Audio Authenticity: Detecting ENF Discontinuity With High Precision Phase Analysis

Daniel Patricio Nicolalde Rodríguez, José Antonio Apolinário, Jr., *Senior Member, IEEE*, and Luiz Wagner Pereira Biscainho, *Member, IEEE*

***Abstract*—This paper addresses a forensic tool used to assess audio authenticity. The proposed method is based on detecting phase discontinuity of the power grid signal; this signal, referred to as electric network frequency (ENF), is sometimes embedded in audio signals when the recording is carried out with the equipment connected to an electrical outlet or when certain microphones are in an ENF magnetic field. After down-sampling and band-filtering the audio around the nominal value of the ENF, the result can be considered a single tone such that a high-precision Fourier analysis can be used to estimate its phase. The estimated phase provides a visual aid to locating editing points (signalled by abrupt phase changes) and inferring the type of audio editing (insertion or removal of audio segments). From the estimated values, a feature is used to quantify the discontinuity of the ENF phase, allowing an automatic decision concerning the authenticity of the audio evidence. The theoretical background is presented along with practical implementation issues related to the proposed technique, whose performance is evaluated on digitally edited audio signals.**

***Index Terms*—Audio authenticity, discrete Fourier transform (DFT), electric network frequency (ENF), forensic analysis, phase estimation.**

I. Introduction

FORENSIC audio authenticity, a branch of audio forensics, has developed remarkably over the last years due to advances in digital signal processing (DSP) and a growing availability of technology [1]. It uses DSP methods to perform signal analysis of recorded audio evidence in legal and law enforcement contexts.

As any other forensic science, authenticity examinations analyze and interpret physical evidence using natural sciences. The goal of this paper is to detail a technique that uses a high precision phase analysis to detect electric network frequency (ENF) discontinuities and thus provide some degree of audio authentication [2], [3]. The proposed technique is, therefore, based on the presence of a small portion of the power grid signal, sometimes embedded in audio recordings.

Manuscript received February 26, 2010; revised April 16, 2010; accepted April 20, 2010. Date of publication June 01, 2010; date of current version August 13, 2010. This work was supported in part by the Brazilian Agencies CAPES, CNPq, and FAPERJ. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Darko Kirovski.

D. P. Nicolalde Rodríguez and J. A. Apolinário, Jr. are with the Department of Electrical Engineering, Military Institute of Engineering (IME), Rio de Janeiro, RJ, Brazil (e-mail: danielnicolalde@hotmail.com; apolin@ime.eb.br).

L. W. P. Biscainho is with the Program of Electrical Engineering, COPPE/Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brazil (e-mail: wagner@lps.ufrj.br).

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

Digital Object Identifier 10.1109/TIFS.2010.2051270

The importance of this topic is enhanced by the advent of personal computers and all sorts of digital technology: we may say that, today, editing digital audio has become a simple task [4]. Moreover, if a good job is carried out, it is hard, even for well-trained ears, to detect this type of fraud, hence, the importance of this subject in the field of audio authenticity.

To tackle the digital audio authenticity problem, this paper resorts to modern DSP techniques which, to some extent, can be quite effective in detecting subtle changes in the phase of the ENF, provided it is present in the recorded material.

The paper is organized as follows. Section II provides some background about the power grid signal: its generation, behavior of the ENF and its phase, and how it is embedded in audio signals. Section III deals with estimating the phase of a sinusoidal signal. We start from a simple concept, the use of the discrete Fourier transform (DFT), and discuss a high-precision Fourier analysis technique for which we propose an efficient phase estimation scheme. Section IV details the proposed method for audio authenticity based on the phase estimate of the power grid signal. The method includes a visual characterization as well as an automatic discrimination. Section V evaluates the proposed method with real audio signals. The signals belong to two public *corpora*. Examples of the two types of editing (insertion and removal of a signal fragment) are also shown in this section. Finally, after a few practical issues discussed in Section VI, conclusions are summarized in Section VII.

II. The Power Grid Signal

The electric power system, as an important element for modern society, constitutes a fundamental factor for the development of countries and can be defined as a group of apparatuses, wires, and machines, that links the power plants to costumers and their needs. Power plants may generate energy by different ways including thermal (coal, oil, nuclear, geothermal), hydroelectric, solar, and wind. The public power grid signal may be viewed as a single sinusoidal waveform with a fixed frequency (the so-called ENF).

Most of the power provided by the power grid comes from turbines that work as generators of alternating current. The rotation velocity of these turbines determines the ENF, whose standard nominal values are 50 and 60 Hz. The first value is adopted in European countries, Asian countries (except Saudi Arabia), African countries (except Liberia), Australia, and in some South American countries like Argentina, Bolivia, Chile, Uruguay, and Paraguay. Meanwhile, 60 Hz is used in Central and North America and in some other South American countries including Ecuador, Venezuela, Peru, Colombia, and Brazil.

Japan is a peculiar case that adopts both 50 and 60 Hz as ENF nominal values.

It is important to mention that, for a correct operation of the power system, frequency and phase of all power generation units should remain synchronous within narrow limits. It is, therefore, of paramount importance that the ENF remains stable. If, for example, a generator drops 2 Hz below the nominal ENF, it will rapidly build up enough heat to destroy itself [5]. Therefore, in the majority of cities, especially those in the most developed regions, a tight control is kept over operator units.

Every type of electric equipment operating connected to the power grid emits an electromagnetic field. This fact causes the power grid signal to be embedded in some recorded signals when a recording device is connected to an electrical outlet or to certain microphones in an ENF magnetic field [6]. Its presence in recorded signals and its expected frequency and phase stability make the ENF useful in some audio authenticity examinations [7], [8].

In [4] and [9], the ENF is used for the task of audio authenticity; the method therein is based on comparing the pattern of the ENF embedded in a recorded signal with the patterns of the power grid signals from a few (suspect) regions, which have been previously stored in a database. It is then possible to obtain, besides audio authentication, information about the place where and the time when the recording was carried out. The Forensic Speech and Audio Analysis Working Group of the European Network of Forensic Science Institutes recently published a document giving guidelines for the use of ENF analysis in forensic authentication of audio recordings [10], attesting to the importance of this subject.

The present work is based on estimating the phase of the power grid signal embedded in the recorded audio signal assuming that a database with ENF information is not available. We use abrupt changes in the estimated phase to infer whether or not the signal has been digitally edited.

III. Estimating Frequency and Phase of a Single Tone

The power grid signal may be viewed as a single tone whose frequency and phase can be estimated. This section starts with the short-time DFT [11] and proceeds to a high-precision Fourier analysis method named DFT^1 (the term DFT^k was coined in [12] denoting the DFT of the kth derivative of a signal, DFT^0 representing its regular DFT).

A. Phase Estimation Using the DFT

Let $s_{\text{tone}}(n)$ be an M-sample single tone sequence, whose frequency and phase are to be estimated. The application of a smoothing window $w(n)$ (e.g., Hann) yields the signal $x(n) = s_{\text{tone}}(n)w(n)$. The N_{DFT}-point DFT of $x(n)$, with $N_{\text{DFT}} \geq M$, will be called $X(k)$.

Let k_{peak} be the integer index associated with the maximum value of $|X(k)|$. Then, the estimated value of the tone frequency is

$$f_{\text{DFT}} = k_{\text{peak}} \frac{f_s}{N_{\text{DFT}}} \tag{1}$$

where f_s is the sampling frequency of $s_{\text{tone}}(n)$. The resolution of f_{DFT}, which can only assume discrete values, is f_s/N_{DFT}. This means that the greater the value of N_{DFT}, the better the accuracy of f_{DFT}, at the expense of increased computational burden. The tone phase is simply the argument (or angle) of $X(k_{\text{peak}})$

$$\phi_{\text{DFT}} = \arg\left[X(k_{\text{peak}})\right]. \tag{2}$$

B. The Novel Phase Estimation Method

The method in [12], named DFT^1, refines the DFT-based frequency estimation of a single tone, and is commonly used to extract spectral modeling parameters from audio signals. It uses the short-time DFT of the first-order signal derivative. Practical experiments show that DFT^1 attains an improved accuracy in finding the peak of the signal spectrum (i.e., the actual value of its frequency) compared to the DFT method, even for small values of N_{DFT}.

The basic steps to estimate the frequency, as presented in [12], are the following:

1) Compute the approximate first derivative of the signal at instant n

$$s'_{\text{tone}}(n) = f_s\left[s_{\text{tone}}(n) - s_{\text{tone}}(n-1)\right].$$

2) Obtain the windowed version of $s_{\text{tone}}(n)$ and $s'_{\text{tone}}(n)$

$$x(n) = s_{\text{tone}}(n)w(n)$$
$$x'(n) = s'_{\text{tone}}(n)w(n).$$

3) Obtain the N_{DFT}-point DFT of $x(n)$ and $x'(n)$. They will be denoted as $X(k)$ and $X'(k)$, respectively.
4) Compute $|X(k)|$ and $|X'(k)|$ as well as k_{peak}, obtained as in Section III-A.
5) Multiply $|X'(k)|$ by the scaling factor $F(k)$

$$F(k) = \frac{\pi k}{N_{\text{DFT}} \sin\left(\frac{\pi k}{N_{\text{DFT}}}\right)}.$$

At this point, we have $\text{DFT}^0[k] = |X(k)|$ and $\text{DFT}^1[k] = F(k)|X'(k)|$.

6) Finally, the value of the estimated frequency

$$f_{\text{DFT}^1} = \frac{1}{2\pi} \frac{\text{DFT}^1[k_{\text{peak}}]}{\text{DFT}^0[k_{\text{peak}}]}.$$

According to [12], k_{peak} is expected to be the closest integer to $f_{\text{DFT}^1} N_{\text{DFT}}/f_s$; then, in order for f_{DFT^1} to be considered a valid solution,

$$\frac{f_s\left(k_{\text{peak}} - \frac{1}{2}\right)}{N_{\text{DFT}}} \leq f_{\text{DFT}^1} < \frac{f_s\left(k_{\text{peak}} + \frac{1}{2}\right)}{N_{\text{DFT}}}$$

must be satisfied, otherwise the method has failed for this frequency. If we define $k_{\text{DFT}^1} = N_{\text{DFT}} f_{\text{DFT}^1}/f_s$, the validation condition can be rewritten as

$$\left(k_{\text{peak}} - \frac{1}{2}\right) \leq k_{\text{DFT}^1} < \left(k_{\text{peak}} + \frac{1}{2}\right).$$

The mechanism introduced in [12] is intended to estimate the value of the frequencies of single tones present in an audio signal, based on the use of the Fourier transform of signal derivatives. The method proposed below extends this result to estimate the phase of a single tone.

Considering a signal model given by $s_{\text{tone}}(n) = a(n)\cos(\omega_0 n + \phi_0)$, the signal phase corresponds to $\phi(n) = \omega_0 n + \phi_0$, where ϕ_0 is the phase at $n = 0$. An estimation of such a value would be restricted to the interval between $-\pi$ and π, and a plot of $\phi(n) \times n$ would be a saw-tooth-like curve (wrapped phase). This model of $s_{\text{tone}}(n)$ is of a narrowband signal, which would be deterministic were $a(n)$ a constant. In practice, $a(n)$ is assumed to evolve slowly over time, and thus can be taken as approximately constant within a small analysis frame or "window." The model does not include any stochastic part (or broadband component), but can be applied to the target problem of this work, since, as will be seen in the next section, all frequency components outside a small bandwidth defined around the ENF nominal value are carefully filtered out.

Therefore, the signal can be expressed as

$$s_{\text{tone}}(n) = a\cos(\omega_0 n + \phi_0) \tag{3}$$

where $\omega_0 = 2\pi f_{\text{tone}}/f_s$, and f_{tone} is the actual value of the tone frequency.

Consequently, $s'_{\text{tone}}(n)$, as computed in the first step of the DFT^1 frequency estimation procedure, can be expressed as

$$\begin{aligned}\frac{s'_{\text{tone}}(n)}{af_s} &= [\cos(\omega_0 n + \phi_0) - \cos(\omega_0 n - \omega_0 + \phi_0)] \\ &= [\cos(\phi_0) - \cos(\phi_0 - \omega_0)]\cos(\omega_0 n) \\ &\quad - [\sin(\phi_0) - \sin(\phi_0 - \omega_0)]\sin(\omega_0 n).\end{aligned} \tag{4}$$

Additionally, since the first difference of a sinusoid (tone) is in fact another sinusoid with the same frequency, (4) can be represented by

$$\begin{aligned}\frac{s'_{\text{tone}}(n)}{af_s} &= C\cos(\omega_0 n + \theta) \\ &= C\cos(\theta)\cos(\omega_0 n) - C\sin(\theta)\sin(\omega_0 n)\end{aligned} \tag{5}$$

where C is a constant and θ is the phase of s'_{tone}.

Comparing (4) to (5), we can write

$$C\cos(\theta) = \cos(\phi_0) - \cos(\phi_0 - \omega_0) \tag{6}$$

and

$$C\sin(\theta) = \sin(\phi_0) - \sin(\phi_0 - \omega_0). \tag{7}$$

Dividing (7) by (6), we obtain

$$\begin{aligned}\tan(\theta) &= \frac{\sin(\theta)}{\cos(\theta)} = \frac{\sin(\phi_0) - \sin(\phi_0 - \omega_0)}{\cos(\phi_0) - \cos(\phi_0 - \omega_0)} \\ &= \frac{\sin(\phi_0)[1 - \cos(\omega_0)] + \cos(\phi_0)\sin(\omega_0)}{\cos(\phi_0)[1 - \cos(\omega_0)] - \sin(\phi_0)\sin(\omega_0)}.\end{aligned} \tag{8}$$

Dividing both numerator and denominator of (8) by $\cos(\phi_0)$ and isolating $\tan(\phi_0)$, the next expression is obtained

$$\tan(\phi_0) = \frac{\tan(\theta)[1 - \cos(\omega_0)] + \sin(\omega_0)}{1 - \cos(\omega_0) - \tan(\theta)\sin(\omega_0)}. \tag{9}$$

The value of ϕ_0 represents the initial phase of $s_{\text{tone}}(n)$; since it is being estimated from the DFT^1, we write it as

$$\phi_{\text{DFT}^1} = \arctan\left\{\frac{\tan(\theta)[1 - \cos(\omega_0)] + \sin(\omega_0)}{1 - \cos(\omega_0) - \tan(\theta)\sin(\omega_0)}\right\} \tag{10}$$

where the value of ω_0 is approximated as $\omega_0 \approx 2\pi f_{\text{DFT}^1}/f_s$.

For the value of θ, we carry out a linear interpolation in the argument of $X'(k)$. Let k_{low} and k_{high} be defined as

$$k_{\text{low}} = \text{floor}[k_{\text{DFT}^1}]$$

and

$$k_{\text{high}} = \text{ceil}[k_{\text{DFT}^1}]$$

where $\text{floor}[\alpha]$ rounds the value of α to the nearest integer less than or equal to α and $\text{ceil}[\beta]$ rounds the value of β to the nearest integer greater than or equal to β.

Recalling that $k_{\text{DFT}^1} = N_{\text{DFT}} f_{\text{DFT}^1}/f_s$, a linear interpolation between points $(k_{\text{low}}, \theta_{\text{low}} = \arg[X'(k_{\text{low}})])$ and $(k_{\text{high}}, \theta_{\text{high}} = \arg[X'(k_{\text{high}})])$ can yield point $(k_{\text{DFT}^1}, \arg[X'(k_{\text{DFT}^1})])$, whose argument corresponds to the value of θ used in (10), i.e.,

$$\theta \approx (k_{\text{DFT}^1} - k_{\text{low}})\frac{\theta_{\text{high}} - \theta_{\text{low}}}{k_{\text{high}} - k_{\text{low}}} + \theta_{\text{low}}. \tag{11}$$

From (10), it is worth mentioning that ϕ_{DFT^1} can have two possible values. If $\arctan(\phi_{\text{DFT}^1})$ has a positive value, ϕ_{DFT^1} could be in the first or in the third quadrant of a two-dimensional Cartesian system; if, on the other hand, $\arctan(\phi_{\text{DFT}^1})$ has a negative value, ϕ_{DFT^1} could be in the second or in the fourth quadrant. A simple decision can be taken by using the value of ϕ_{DFT} as a reference: choose the value of ϕ_{DFT^1} closer to ϕ_{DFT}.

C. Preliminary Experiments

In order to understand better and evaluate the proposed method, we provide the results of a few preliminary computer experiments.

We have initially considered a 60.98-Hz sinusoidal tone sampled at 1200 Hz. In Fig. 1, the true spectrum of this signal, zoomed around the nominal frequency of the tone, is shown together with its associate discrete spectra computed via 200- and 2000-point DFTs.

In this experiment, we obtained the first 100 estimated frequencies and phases for consecutive frames of the test tone delimited by a 200-sample sliding window (i.e., advancing sample by sample). For this particular signal, the DFT procedure provided a constant estimated frequency value of 60 Hz for $N_{\text{DFT}} = 200$, and 61.20 Hz for $N_{\text{DFT}} = 2000$. Meanwhile, when using the DFT^1 method, the values of the estimated frequency had a mean of 60.9719 Hz with a standard deviation of 0.0025 Hz for $N_{\text{DFT}} = 200$, and a mean of 60.9818 Hz with a standard deviation of 0.0032 Hz for $N_{\text{DFT}} = 2000$.

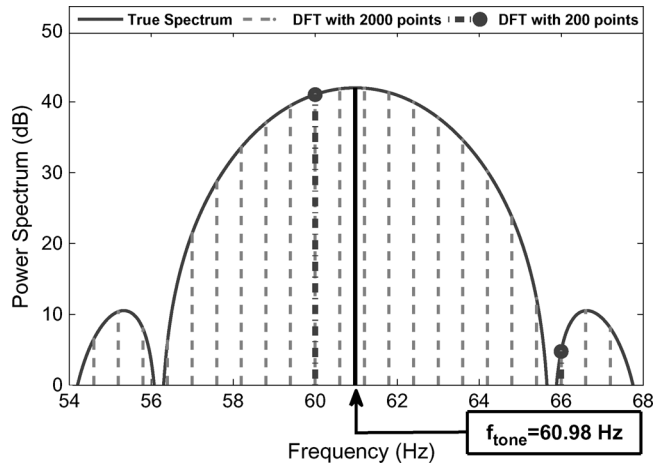

Fig. 1. Spectra of 500 windowed samples of a single 60.98-Hz tone sampled at 1200 Hz: continuous spectrum; 200-point DFT; and 2000-point DFT.

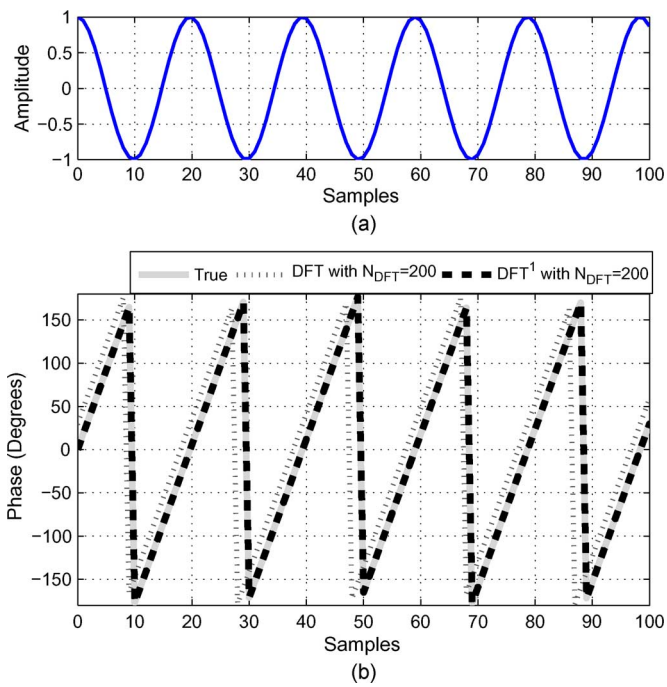

Fig. 2. Phase estimation of an artificial 60.98-Hz tone: (a) signal; (b) phase estimate.

A mean relative error[1] averaged over the frames $\overline{e}_f$ has been computed for both DFT and DFT^1. The values of $\overline{e}_f$ obtained were: 1.61% with $N_{\mathrm{DFT}} = 200$, and 0.36% with $N_{\mathrm{DFT}} = 2000$ for the DFT method; 0.014% with $N_{\mathrm{DFT}} = 200$, and 0.005% with $N_{\mathrm{DFT}} = 2000$ for the DFT^1 method. The errors attained with the DFT^1 method are substantially lower than those with the DFT method.

The respective 100 estimations of phase using DFT and DFT^1 are shown in Fig. 2. For both methods, the same number of points N_{DFT} and window size have been used. Analogously to the case of frequency estimation, a mean error $\overline{e}_\phi$ has been obtained as an average over the phase error.[2] The mean phase errors obtained were: 29.25° with $N_{\mathrm{DFT}} = 200$, and 6.57° with $N_{\mathrm{DFT}} = 2000$ for the DFT method; 0.25° with $N_{\mathrm{DFT}} = 200$, and 0.0912° with $N_{\mathrm{DFT}} = 2000$ for the DFT^1 method. A considerable improvement has been be obtained in phase estimation using the new method.

TABLE I
EVALUATION OF FREQUENCY AND PHASE ESTIMATIONS USING DFT AND DFT^1. THE EXPERIMENT WAS CARRIED OUT WITH 1000 TONES WITH FREQUENCIES VARYING RANDOMLY BETWEEN 59.0 AND 61.0 HZ. $\bar{e}_f$ AND $\bar{e}_\phi$ REPRESENT THE MEAN ERRORS IN FREQUENCY AND PHASE, RESPECTIVELY

Method	M (samples)	N_{DFT} (points)	$\bar{e}_f$ (%)	$\bar{e}_\phi$ (Degrees)
DFT	60	200	0.8306	4.4086
		2,000	0.2696	1.4309
		20,000	0.0679	0.3606
	100	200	0.8189	7.2944
		2,000	0.2688	2.3949
		20,000	0.0261	0.2329
	200	200	0.8180	14.6496
		2,000	0.2588	4.6359
		20,000	0.0246	0.4411
DFT^1	60	200	0.0559	0.2907
		2,000	0.0543	0.2831
		20,000	0.0538	0.2802
	100	200	0.0138	0.1217
		2,000	0.0131	0.1160
		20,000	0.0130	0.1152
	200	200	0.0024	0.1221
		2,000	0.0024	0.0559
		20,000	0.0015	0.0436

A statistical evaluation of frequency and phase estimation for both methods, DFT and DFT^1, was performed. For that, 1000 tones with frequencies randomly varying (with uniform distribution) between 59.0 and 61.0 Hz were synthesized. Subsequently, the errors in the estimates of frequency and phase for different DFT lengths and window sizes (M) were computed. Table I summarizes the results.

It can be seen that when the DFT length increases, given a constant window size, frequency and phase estimates improve in both methods; this is due to the fact that the signal spectrum is sampled with higher resolution. Additionally, the DFT^1 method provides a substantial improvement in frequency and phase estimation when compared to the DFT method, for the same N_{DFT}. This effect can be seen in Table I: the DFT^1 estimates with lowest resolution $(N_{\mathrm{DFT}} = 200)$ are better than the DFT estimations with highest resolution $(N_{\mathrm{DFT}} = 20\,000)$. For stationary signals, as in the present experiment, increasing window size improves frequency estimation in both methods. However, this parameter cannot grow unbounded if one needs to detect abrupt phase changes, thus it should be kept low for the target application.

IV. THE PROPOSED METHOD

As mentioned before, the power grid signal may be embedded in recorded signals. Consequently, considering that audio editing means removal or insertion of a portion of audio, the same action is carried out in the embedded power grid

[1]The relative frequency error at the n_bth frame is defined by the rate between the absolute value of the difference *estimated value minus correct value* and the correct value, i.e., $e_f(n_b) = (|\hat{f}(n_b) - f(n_b)|/f(n_b)) \times 100\%$.

[2]The phase error at the n_bth frame is defined as the absolute value of the difference between the estimated value and the correct phase and is given as $e_\phi(n_b) = |\hat{\phi}(n_b) - \phi(n_b)|$.

signal. Following this reasoning, a method that attempts to detect abrupt changes in the phase of the embedded ENF signal is proposed here.

The method can be divided in two parts. The first part comprises a visual mechanism that allows the observation of the behavior of the estimated phase of the power grid signal. The other part automatically discriminates between original and edited signals by means of a decision ratio. The basic idea of this method, without employing high-precision phase analysis, can be found in [13] and [14].

A. Visual Method

The steps of the visual method are detailed below:

1) Down-sample the audio signal to a frequency $f_{\rm ds}$ which, as a suggestion, could be 1000 or 1200 Hz, depending on the value of the nominal ENF being 50 or 60 Hz, respectively. This synchronous sampling, besides reducing the analysis computational burden, allows working with an exact number of samples per cycle of the nominal ENF (or, in the frequency domain, locating one DFT bin exactly on the nominal ENF).
2) Use a very sharp linear-phase FIR filter to bandpass the down-sampled signal. This filter should be centered in the nominal ENF value, and have a passband width between 0.6 and 1.4 Hz, depending on the ENF tolerance guaranteed by the electrical company. In the experiments carried out in this work, a 10 000-coefficient zero-phase filter has been employed (using Matlab function *filtfilt* to avoid delay).
3) Divide the filtered signal in blocks of N_C cycles of the nominal ENF, each block overlapping the former by (N_C-1) cycles. The signal is then segmented in $N_{\rm Block}$ blocks. In Fig. 3, blocks of $N_C = 3$ cycles of the nominal ENF are shown.
4) Estimate the phase of every segmented block using DFT or DFT[1]. Let $\hat{\phi}(n_b)$ be the corresponding phase estimate for the block index n_b.
5) Plot phase values in degrees versus cycles of nominal ENF for visual inspection.

B. Automatic Method

The automatic discrimination between edited and original signals requires a feature that characterizes the detection of abrupt phase changes in the power grid signal embedded in the recorded audio, related to audio editing. The variation of the estimated ENF phase for the n_bth block under analysis

$$\hat{\phi}'(n_b) = \hat{\phi}(n_b) - \hat{\phi}(n_b - 1) \tag{12}$$

for $2 \le n_b \le N_{\rm Block}$, is chosen for this purpose.

Taking $m_{\hat{\phi}'}$ as the average of $\hat{\phi}(n_b)$ from $n_b = 2$ to $N_{\rm Block}$, the proposed feature F is then defined as

$$F = 100 \log \left\{ \frac{1}{N_{\rm Block} - 1} \sum_{n_b=2}^{N_{\rm Block}} \left[\hat{\phi}'(n_b) - m_{\hat{\phi}'} \right]^2 \right\}. \tag{13}$$

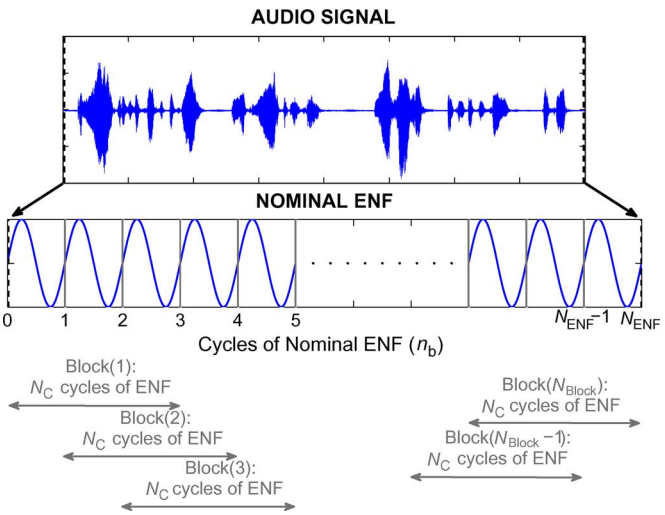

Fig. 3. Block fragmentation of an audio signal. $N_{\rm ENF}$ is the number of ENF cycles in the audio signal; $N_{\rm Block}$ is the number of fragment blocks; N_C is the number of ENF cycles in each block.

For the detection process, the hypothesis group $\{H_O, H_E\}$ is defined; H_O and H_E represent the hypothesis for an audio signal being original and edited, respectively. A decision ratio for the automatic detection can be expressed as

$$F \underset{H_O}{\overset{H_E}{\gtrless}} \gamma \tag{14}$$

where γ is a threshold. For F greater than γ, it is decided that the audio signal has been edited, i.e., hypothesis H_E. Otherwise, hypothesis H_O is favored.

Let P_D be the probability of *detection*, or *hit* (i.e., the audio signal is considered as edited when it indeed has been edited), P_F be the probability of *false alarm* (i.e., the audio signal is considered as edited when it has actually not been edited), and P_M as the probability of a *miss* (i.e., the audio signal is considered as not edited when it indeed has been edited). The expressions for P_D, P_F, and P_M are

$$P_D = P(\hat{H} = H_E | H_E) = P(F > \gamma | H_E)$$
$$P_F = P(\hat{H} = H_E | H_O) = P(F > \gamma | H_O)$$

and

$$P_M = P(\hat{H} = H_O | H_E) = P(F < \gamma | H_E).$$

Additionally, $P_D = 1 - P_M$. For optimal detection, the goal is to obtain a value of γ that maximizes the value of P_D. To establish this threshold, it is necessary to prepare a *corpus* of audio signals including their original and edited (in a controlled way) versions, and evaluate this database with the proposed method for an extended range of γ values; with the corresponding values of P_M and P_F, the so-called detection error tradeoff (DET) curve [15] (P_M as a function of P_F) is constructed. The point in the curve where $P_M = P_F$ is known as the equal error rate (EER). The value of γ that corresponds to the EER point will be taken as the decision threshold in (14). The EER allows the characterization of the detection system error by a single parameter, since the lower this value the better is the system performance.

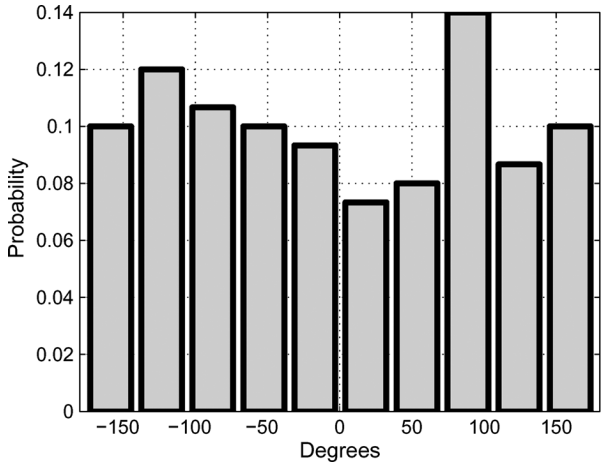

Fig. 4. Histogram of phase change for the signals forming the *corpus* of edited speech.

V. Assessing the Method

In order to evaluate the proposed method, it was employed to check the authenticity of digitally edited recordings. The original recordings [digitized with 16-bit quantization and sampling rates of 8 kHz in the case of telephone and 16 kHz in the case of microphone signals (see additional details in [16])] for the main evaluation were taken from two public databases in Spanish, AHUMADA and GAUDI, obtained from the website http://atvs.ii.uam.es/databases.jsp. The signals chosen for the evaluation were checked and determined to be neither digitally saturated nor having a low signal-to-noise ratio (SNR). Since both databases came from Spain, the nominal ENF associated with all signals was 50 Hz. Overall, 100 signals (speech utterances) were used: 50 by female and 50 by male speakers. They were edited in such a way that half of the speech files had an audio portion deleted, while the other half had a portion of audio inserted. Insertions were carried out with a fragment of audio belonging to the same file in order to avoid strong short-time spectral changes (due to, for example, a difference in sampling rate), which could make the detection easier. In [13], the authors address the problem of detecting audio discontinuities from spectral distances.

It is important to mention that the editing was carried out without regard to the phase changes ocurring in the power grid signal, in an attempt to emulate the way most files are digitally edited. Therefore, the phase changes resulted in a random distribution among the speech files, as depicted in Fig. 4. This histogram shows a relatively uniform distribution between $-180°$ and $+180°$. That means that all graduations of difficulty are covered by the edited databases, as would happen in real life.

In the following, the values of the different variables used in the proposed method to detect audio editing from ENF discontinuity are detailed. Because of the ENF nominal value, 50 Hz for this evaluation, the sampling frequency after decimation was set to $f_{\rm ds} = 1000$ Hz. The value passband width of the tuned filter around the nominal ENF was chosen to be 0.8 Hz. Additionally, window size values N_C were chosen as 3, 5, and 10 cycles of the nominal ENF (1 cycle = 20 ms of signal); and DFT length values $N_{\rm DFT}$ were chosen as 200, 2000, and 20 000 points.

TABLE II
Evaluation of Audio Authenticity for the Test Audio *corpus* (100 Original and 100 Edited Signals). N_C Represents the Analysis Window Size in Cycles of the Nominal ENF

Method	$N_{\rm C}$	$N_{\rm DFT}$	EER (%)
DFT	3 (60 samples)	200	6
		2,000	6
		20,000	7
	5 (100 samples)	200	6
		2,000	6
		20,000	8
	10 (200 samples)	200	6
		2,000	6
		20,000	9
DFT^1	3 (60 samples)	200	6
		2,000	6
		20,000	6
	5 (100 samples)	200	6
		2,000	6
		20,000	6
	10 (200 samples)	200	8
		2,000	6
		20,000	6

Table II summarizes the results obtained by the automatic discrimination method according to the decision rule expressed in (14). It can be seen that the use of DFT^1 yields more stable, almost constant results: around 6% EER in the audio editing decision, regardless of N_C and $N_{\rm DFT}$ values (the exception, EER $= 8\%$, occurs for the only case where the DFT length is not greater than the window size).

After examination of the results in Table II, allowing for some safety margin in both parameters without increasing too much of the computational load, a cautious choice is to use the DFT^1 method with a window size of 5–10 ENF cycles and $N_{\rm DFT} = 2000$ points. Next, a particular case using the DFT^1 method with this choice ($N_C = 10$ cycles and $N_{\rm DFT} = 2000$) is further detailed. Fig. 5 presents the histograms of feature F for both edited and original signals in the test audio *corpus* as well as the localization of γ, the optimum decision threshold for the detection process. It can be seen that the distribution of original signals are reasonably separated from the distribution of edited signals.

The DET curve (P_M versus P_F) as well as the localization of the EER point (6%) for this particular case are shown in Fig. 6.

In an attempt to show results of the visual aid provided by the proposed method, two examples of audio editing of signals from the test audio *corpus* are presented. The preset is the same at that used for the particular case previously detailed.

Fig. 7 presents an example where an audio portion has been deleted from a speech file. The phase estimation for the original ENF signal has rectilinear behavior, whereas the edited signal has an abrupt phase change at the edit point P_1 (in this case, a positive change).

Fig. 8 presents an example where an audio portion has been inserted into a speech file. There are, in this case, two edit points: P_1 and P_2. Consequently, we can observe two phase changes in the edited signal (the first one negative, the second one positive).

Phase estimation using the DFT^1-based method exhibits better resolution than using the DFT-based method (specially in the regions where phase transitions occur). This greater accuracy in phase estimation improves the visual aid.

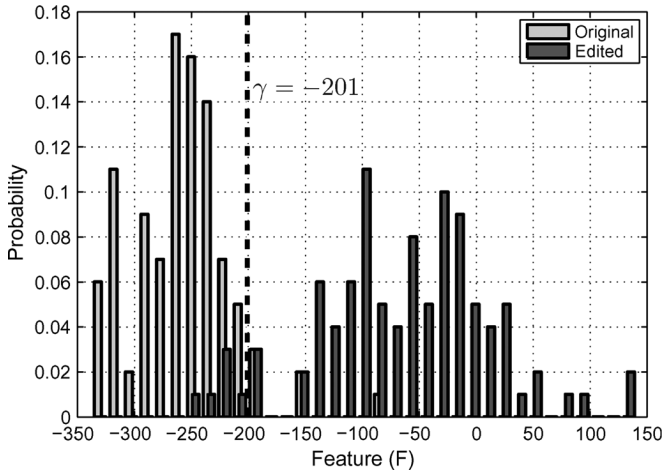

Fig. 5. Histograms of feature F for the test audio *corpus*. The DFT[1]-based phase estimation method was used, with a window size of 10 cycles of the nominal ENF and $N_{\mathrm{DFT}} = 2000$ points.

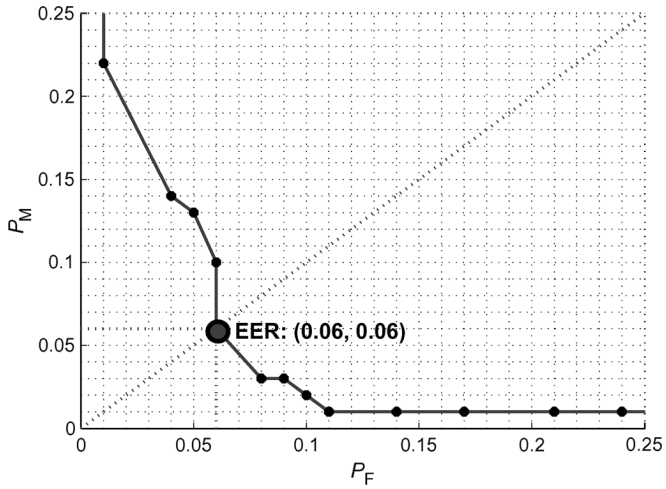

Fig. 6. DET curve: P_M versus P_F for the test audio *corpus*. The DFT[1]-based phase estimation method was used, with a window size of 10 cycles of the nominal ENF and $N_{\mathrm{DFT}} = 2000$ points.

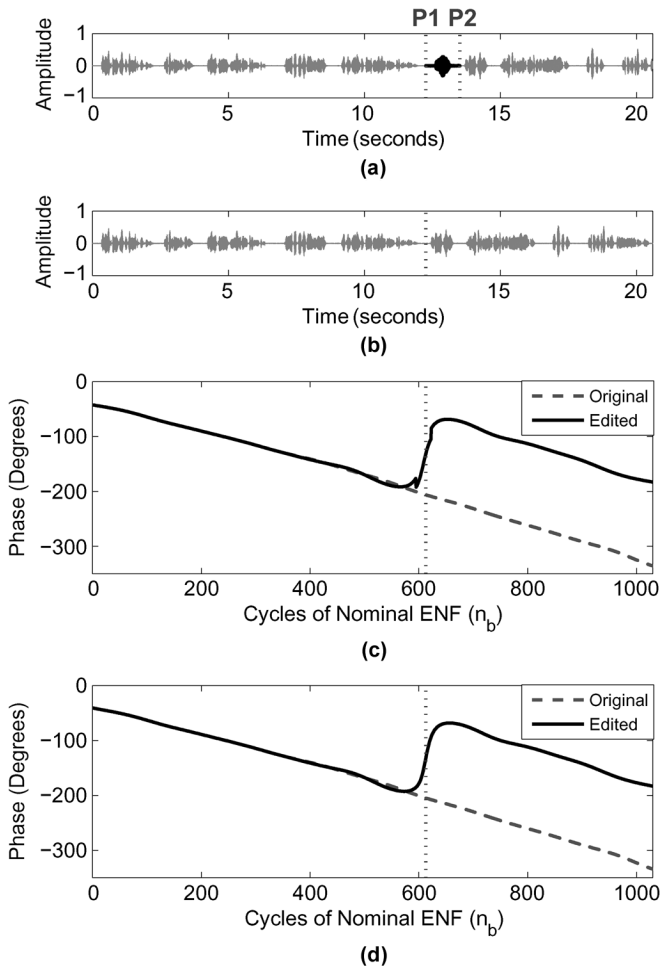

Fig. 7. Visualization of a fragment deletion. Points P_1 and P_2 bound the portion eliminated from the original signal. Consequently, P_1 is the edited point within the signal. The nominal ENF is 50 Hz and the passband width of the bandpass filter is 0.8 Hz. The phase estimation methods used a window size of 10 cycles of the nominal ENF and $N_{\mathrm{DFT}} = 2000$ points. (a) Original signal. (b) Edited signal. (c) Phase estimation using DFT. (d) Phase estimation using DFT[1].

VI. Practical Issues

The well-behaved ENF variation of the Spanish *corpus* (signals from AHUMADA and GAUDI edited) yielded very nice results. But what if the proposed method is to be used in real-life situations where signals are degraded in a number of ways? To answer this question, two additional local *corpora* were prepared, containing recordings in Portuguese as spoken in Rio de Janeiro, Brazil: Carioca 1 (digitized with 16-bit quantization and a sampling rate of 44 100 Hz) and Carioca 2 (16 bits and 11 050 Hz), both with the same structure of the edited Spanish *corpus*: a total of 100 original and 100 edited signals. As Brazilian speech databases, their nominal ENF is 60 Hz.

Carioca 1 speech signals were recorded with low background noise and without saturation. The EER obtained for this *corpus* was 7%. This result, only slightly worse than the one obtained for the Spanish *corpus*, could be due to the slightly faster variation of the ENF contained in Carioca 1 recordings. Although in both cases the ENF has a similar deviation around their nominal values, the ENF in Spain seems to vary more slowly than the ENF recorded in the city of Rio de Janeiro. Since the difference in EER was very small (1%), no further investigation was carried out; nevertheless, this result reinforces the expectation that the performance of the proposed method would degrade in a region without a tight control over the ENF.

The Carioca 2 *corpus* was prepared under unfavorable conditions: among its 100 signals, 21 exhibited a moderate degree of saturation; and the *corpus* average SNR was around 30 dB (the average SNR of the Spanish *corpus* was estimated in 35 dB). The resulting EER for this *corpus* was 15%. In the following subsections, both effects will be addressed individually.

A. Effect of Background Noise

The Spanish *corpus* was used to carry out this study. Defining s as the clean speech and $n_{\mathrm{background}}$ as the background noise, both mutually uncorrelated by assumption, the original SNR is given as

$$\mathrm{SNR}_{\mathrm{original}} = \frac{E[s^2]}{E\left[(n_{\mathrm{background}})^2\right]}. \tag{15}$$

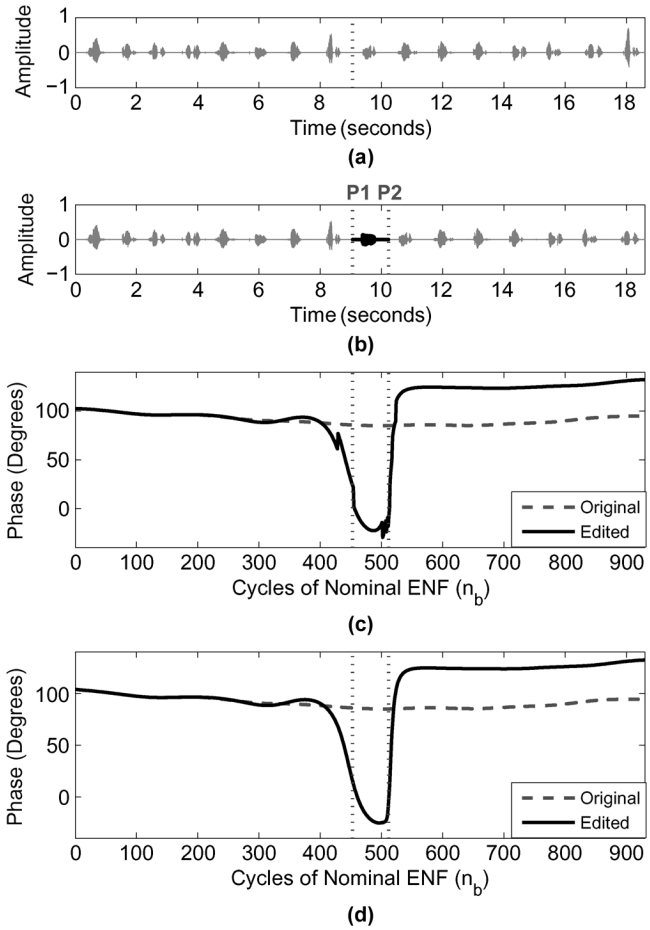

Fig. 8. Visualization of a fragment insertion of length $P_2 - P_1$. P_1 is the insertion point in the original signal. Consequently, P_1 and P_2 are the edited points within the signal. The nominal ENF is 50 Hz and the passband width of the bandpass filter is 0.8 Hz. The phase estimation methods used a window size of 10 cycles of the nominal ENF and $N_{\mathrm{DFT}} = 2000$ points. (a) Original signal. (b) Edited signal. (c) Phase estimation using DFT. (d) Phase estimation using DFT[1].

In order to estimate the original SNR, a *voice activity detector* (VAD) algorithm [17] was employed to separate active speech regions from background noise.

In order to alter the SNR, zero mean uncorrelated noise n_{add} has been added to the signals, such that

$$\mathrm{SNR} = \frac{E[s^2]}{E\left[(n_{\mathrm{background}})^2\right] + E\left[(n_{\mathrm{add}})^2\right]}. \tag{16}$$

The next step is to obtain an error rate as a function of the SNR. For that, the value of $E[(n_{\mathrm{add}})^2]$ was varied and the proposed method for audio authenticity (using the DFT[1]-based phase estimation technique) was applied to the Spanish *corpus*.

Three types of noise were employed:

- NOISE 1: White Gaussian noise.
- NOISE 2: Low-frequency colored noise obtained from white noise filtered through $H_{\mathrm{LP}}(z) = (1/(1 - 0.9z^{-1}))$.
- NOISE 3: High-frequency colored noise obtained from white noise filtered through $H_{\mathrm{HP}}(z) = (1/(1 + 0.9z^{-1}))$. Assuming a frequency sampling of 8000 Hz, the frequency responses of both filters $H_{\mathrm{LP}}(z)$ and $H_{\mathrm{HP}}(z)$, are shown in Fig. 9.

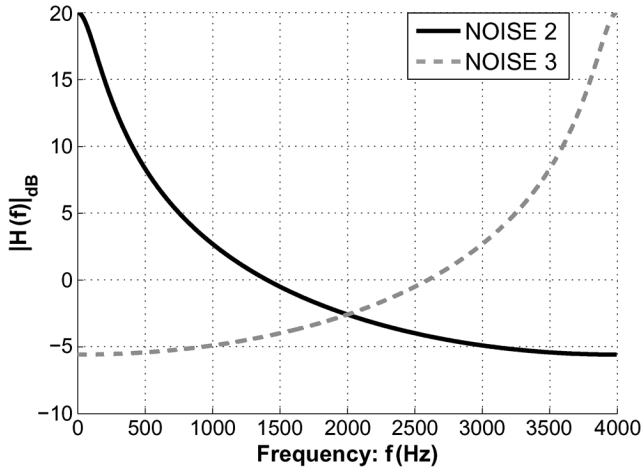

Fig. 9. Frequency response curves corresponding to $H_{\mathrm{LP}}(z)$ and $H_{\mathrm{HP}}(z)$ for NOISE 2 and NOISE 3, respectively. The curves were plotted as a function of the frequency f, in hertz, assuming a sampling rate of 8000 Hz.

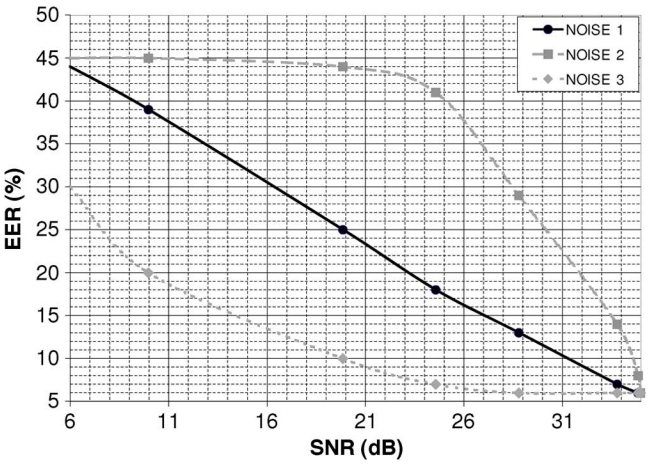

Fig. 10. Effect of background noise: EER in the task of audio authentication as a function of the SNR. The system error without adding extra noise is 6%, and the original SNR is 35 dB.

Fig. 10 presents the (equal) error rates as a function of the SNR when the proposed method is evaluated with additive NOISES 1, 2, and 3. Note that, when the noise has high energy in low frequencies (those components that affect directly the ENF, 50 Hz in this case), the background noise effect is much stronger. Therefore, NOISE 2 (whose curve has a logarithmic shape) is the most harmful to the authentication method, followed by NOISE 1 (linear shape) and NOISE 3 (exponential shape).

B. Effect of Saturation

In order to analyze the effect of the nonlinearity caused by saturation on the audio authenticity method proposed here, the Spanish *corpus* was used once more. Initially, a VAD algorithm [17] was applied to all signals. Then, *a percentage of the active voice samples* (referred to here as saturation level) are clipped to a suitably chosen maximum value. Fig. 11 presents an example of a signal with 3% saturation level.

By varying the saturation level and applying the audio authentication method to the Spanish *corpus*, the curve EER versus sat-

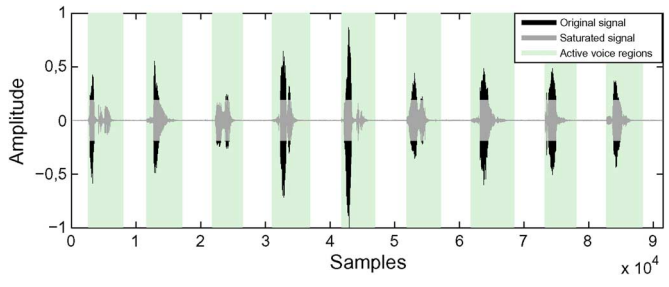

Fig. 11. Example of a signal with 3% saturation level, i.e., with 3% of the samples in active (shaded) regions clipped to a maximum level.

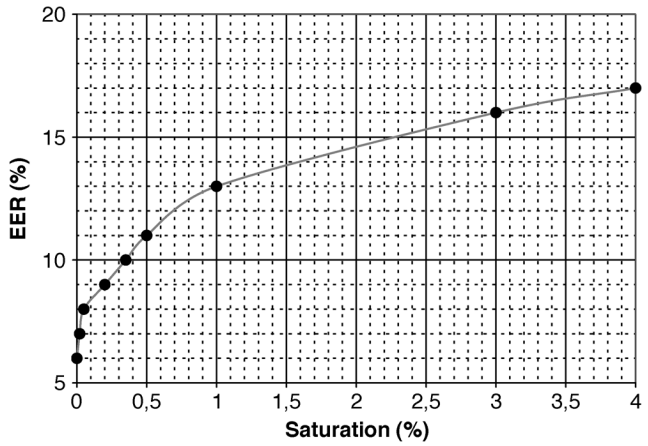

Fig. 12. Effect of saturation: EER in the task of audio authentication as a function of the saturation level. The system error without artificial nonlinearity is 6%.

uration shown in Fig. 12 was obtained. It can be observed that, considering the active voice intervals defined by the shaded regions, saturation levels above 0.5% considerably affected the performance of the authentication method. Also, perceptually speaking, this value corresponds to a high level of saturation.

From what have been studied in this section, assuming that the effect of the unfavorable conditions were linear, one could expect to have, for the Carioca 2 *corpus*, an overall EER given by the sum of the EER for the Spanish *corpus* (6%) with 3 extra terms:

- 1% due to ENF variation—recalling that the experiment with Carioca 1 database, which shares its ENF characteristics with Carioca 2 database, resulted in an additional error of 1% over the Spanish database results;
- 5% due to background error—computing the difference between EERs for 30- and 35-dB SNR in Noise 1 plot of Fig. 10;
- the remaining 3% due to saturation in some signals—a reasonable speculation, not denied by inspection of Fig. 12.

VII. Conclusion

The proposed technique to detect audio editing has yielded favorable results. The idea of finding abrupt phase changes in the power grid signal provides an accurate visual characterization. This visual aid helps in determining the editing points and inferring the type of editing (whether insertion or deletion of audio segments). Additionally, the use of a decision feature allows an automatic discrimination between original and edited signals. In the computer experiments, the error attained by the detection process over clean audio signals was 6%. This small value of EER was probably due to those cases where the editing process caused insignificant ENF phase changes (around 0° in the histogram of Fig. 4).

The use of the DFT[1] method, here adapted to estimate phase with high accuracy, yielded improved resolution in the visual characterization (especially in the regions where the phase transitions are located) as well as good regularity in the automatic discrimination. The use of the DFT[1]-based technique instead of the traditional DFT-based technique is justifiable due to its higher accuracy results with a smaller number of points, thus not increasing computational overhead.

Practical issues related mainly to the effects of nonlinearity and low SNR were independently analyzed.

Considering the presence of power grid signals in some recorded signals, the proposed technique for evaluating audio authenticity can be a useful tool in the field of forensic phonetics. The method described herein becomes even more important in those cases when there is no ENF database available.

It is worth mentioning that several extraneous phenomena either acoustically generated, or inherent to ac transmission or the recording system itself (e.g., transients, power-line spikes and surges, coding artifacts, etc.) may affect the recording under analysis, thus impacting the detection performance. A careful evaluation of those effects should be object of future research.

References

[1] R. C. Maher, "Audio forensic examination: Authenticity, enhancement, and interpratation," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 84–94, Mar. 2009.

[2] B. E. Koenig, "Authentication of forensic audio recordings," *J. Audio Eng. Soc.*, vol. 38, no. 1/2, pp. 3–33, Jan./Feb. 1990.

[3] B. E. Koenig and D. S. Lacey, "Forensic authentication of digital audio recordings," *J. Audio Eng. Soc.*, vol. 57, no. 9, pp. 662–695, Sep. 2009.

[4] R. W. Sanders, "Digital authenticity using the electric network frequency," in *Proc. AES 33rd Int. Conf. Audio Forensics, Theory and Practice*, Denver, CO, Jun. 2008.

[5] M. Amin and J. Stringer, "The electric power grid: Today and tomorrow," *MRS Bulletin*, vol. 33, pp. 399–407, Apr. 2008.

[6] E. Brixen, "ENF quantification of the magnetic field," in *Proc. AES 33rd Int. Conf. Audio Forensic, Theory and Practice*, Denver, CO, Jun. 2008.

[7] C. Grigoras, "Applications of ENF criterion in forensic audio, video, computer, and telecommunication analysis," *Forensic Sci. Int.*, vol. 167, no. 2, pp. 136–145, Apr. 2007.

[8] C. Grigoras, "Applications of ENF analysis in forensic authentication of digital audio and video recordings," *J. Audio Eng. Soc.*, vol. 57, no. 9, pp. 643–661, Sep. 2009.

[9] 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*, Denver, CO, Jun. 2008.

[10] C. Grigoras, A. J. Cooper, and M. Michalek, Best Practice Guidelines for ENF Analysis in Forensic Authentication of Digital Evidence European Network of Forensic Science Institutes, Forensic Speech and Audio Analysis Working Group, 2009, Ref. ENFSI-FSAAWG-BPM-ENF-001.

[11] P. A. Esquef and L. W. Biscainho, "Spectral-based analysis and synthesis of audio signals," in *Advances in Audio and Speech Signal Processing: Technologies and Applications*, H. P. Meana, Ed. Hershey: Idea Group, 2007, pp. 56–92.

[12] M. Desainte-Catherine and S. Marchand, "High-precision fourier analysis of sounds using signal derivatives," *J. Audio Eng. Soc.*, vol. 48, no. 7/8, pp. 654–667, Jul./Aug. 2000.

[13] D. P. Nicolalde R. and J. A. Apolinário, Jr., "Evaluating digital audio authenticity with spectral distances and ENF phase change," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'09)*, Taipei, Taiwan, Apr. 2009.

[14] D. P. Nicolalde R., J. A. Apolinário, Jr., and L. W. Biscainho, "Autenticação de áudio digital com base na mudança de fase da frequência da rede elétrica," in *Proc. XXVII Brazilian Telecommunications Symp. (SBrT'09)* (in Portuguese), Blumenau, Brazil, Sep. 2009.

[15] 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.

[16] J. Ortega-García, J. González-Rodríguez, and V. Marrero-Aguiar, "AHUMADA, a large speech corpus in Spanish for speaker characterization and identification," *Elsevier Speech Commun.*, vol. 31, pp. 255–264, Jun. 2000.

[17] A. Benyassine *et al.*, "ITU-T recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Commun. Mag.*, vol. 35, no. 9, pp. 64–73, Sep. 1997.

Daniel Patricio Nicolalde Rodríguez was born in Quito, Ecuador, in 1982. He completed his undergraduate education in electronics and telecommunications engineering from the Escuela Politécnica del Ejército (ESPE), Quito, Ecuador, in 2007. He is a graduating student at the Military Institute of Engineering (IME), Rio de Janeiro, Brazil, since 2008.

His professional interests include digital signal processing, speech and audio processing, as well as dynamic programming.

José Antonio Apolinário, Jr. (S'95–M'99–SM'04) was born in Taubaté, Brazil, in 1960. He graduated from the Military Academy of Agulhas Negras (AMAN), Resende, Brazil, in 1981 and received the B.Sc. degree from the Military Institute of Engineering (IME), Rio de Janeiro, Brazil, in 1988, the M.Sc. degree from the University of Brasília (UnB), Brasília, Brazil, in 1993, and the D.Sc. degree from the Federal University of Rio de Janeiro (COPPE/UFRJ), Rio de Janeiro, Brazil, in 1998, all in electrical engineering.

He is currently an Adjoint Professor with the Department of Electrical Engineering, IME, where he has already served as the Head of Department and as the Vice-Rector for Study and Research. He was a Visiting Professor at the Escuela Politécnica del Ejército (ESPE), Quito, Ecuador, from 1999 to 2000 and a Visiting Researcher and twice a Visiting Professor at Helsinki University of Technology (HUT), Finland, in 1997, 2004, and 2006, respectively. His research interests comprise many aspects of linear and nonlinear digital signal processing, including adaptive filtering, speech, and array processing. He has recently edited the book *QRD-RLS Adaptive Filtering* (New York: Springer, 2009).

Dr. Apolinário has organized and been the first Chair of the Rio de Janeiro Chapter of the IEEE Communications Society.

Luiz Wagner Pereira Biscainho (S'95–M'03) was born in 1962, in Rio de Janeiro, Brazil. He received the B.Sc. degree (*magna cum laude*) in electronics engineering, the M.Sc. degree and the D.Sc. degrees, both in electrical engineering, all from the Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil, in 1985, 1990, and 2000, respectively.

He is with the Department of Electronics Engineering (DEL/UFRJ) and the Program of Electrical Engineering (COPPE/UFRJ) as an Associate Professor. His research area is digital signal processing, particularly audio processing and adaptive systems.

Besides the IEEE, Dr. Biscainho is an active member of the Audio Engineering Society (AES) and of the Brazilian Telecommunications Society (SBrT).