Circuits Syst Signal Process
https://doi.org/10.1007/s00034-018-0793-9

An ENF-Based Audio Authenticity Method Robust to MP3 Compression

Pablo Zinemanas[1] · **Magdalena Fuentes**[2,3] ·
Pablo Cancela[1] · **José Antonio Apolinário Jr.**[4]

Received: 3 April 2017 / Revised: 21 February 2018 / Accepted: 24 February 2018

Abstract This work presents a novel method for assessing audio authenticity. Assuming that the electric network frequency is embedded in audio signals, the evaluation of audio integrity is carried out by detecting phase discontinuities. This is conducted by using causal and anti-causal filters, in order to avoid the mix of past and future phase information related to the time of analysis. This local phase change is then post-processed and thresholded to obtain the editing times. One remarkable property of the proposed method is its ability to withstand MP3 compression, an audio format widely used in practice. A more accurate evaluation metric is also introduced in this work. For that purpose, the databases used for evaluating the algorithm were automatically labeled indicating the editing times. The procedure to generate the ground truth is

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00034-018-0793-9) contains supplementary material, which is available to authorized users.

✉ Pablo Zinemanas
pzinemanas@fing.edu.uy

Magdalena Fuentes
magdalena.fuentes@l2s.centralesupelec.fr

Pablo Cancela
pcancela@fing.edu.uy

José Antonio Apolinário Jr.
apolin@ime.eb.br

[1] Instituto de Ingeniería Eléctrica, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay

[2] L2S, Univ. Paris-Sud-CNRS-CentraleSupélec, Paris, France

[3] LTCI, Télécom ParisTech, Paris, France

[4] Programs of Electrical and Defense Engineering, Military Institute of Engineering (IME), Rio de Janeiro, Brazil

Published online: 08 March 2018

presented, as well as a discussion on the proposed metric. The performance of the technique presented promising results when evaluated on digitally edited and original audio signals.

Keywords Audio authenticity · PLL · ENF · MP3 · Phase discontinuity

1 Introduction

Recent developments and availability of technology made audio tempering a very simple task [19]. Identifying possible manipulation in speech signals is of great importance to police investigation as well as for forensic analysis. Yet, if the editing (a mere cut or an audio insertion) is carried out carefully, it is hard to determine if the audio has been modified or not, even for trained ears.

Nevertheless, in some cases (e.g., when the recording device is connected to an electrical outlet), a small portion of the power grid signal or electric network frequency (ENF) might be embedded in the recorded signals. The ENF, with typical nominal values equal to 50 or 60 Hz, presents non-periodical temporal variations which depend on the region of recording, i.e., on the electric network where the signal was acquired. Therefore, the ENF embedded in an audio recording can be regarded as a time stamp [11]. At the same time, as the ENF varies smoothly around its nominal value, it is expected to be continuous along the whole audio signal. Editing digital audio without proper care and skills has a large probability of causing a phase discontinuity in the ENF signal.

Since the ENF criterion applied to the authentication of forensic audio recordings was first introduced [11–13], an intensive research effort focused on information forensics, including detecting ENF discontinuities for audio authentication, has been carried out [29]. Most existing research involve the discrete Fourier transform (DFT) or spectral distances for extracting the phase changes of the ENF [20,21,28]. Later on, the use of variations of the ENF to detect audio edit was employed [5] and further improved [6].

More recently, the use of phase-locked loop (PLL) control signals for the task of audio authentication has been proposed in [9]. Besides that, in [10], authors claim that the light intensity from fluorescent lamps and incandescent bulbs varies in accordance to the fluctuation of the power line frequency. They used video signals to detect and estimate the ENF. Different applications of this ENF analysis were reported in this work, including tampering detection and synchronization of audio and video tracks. Other typical applications for ENF analysis are time stamping [1,16,29] and identification of the region of recording [15]. Also, there are commercial products performing ENF analysis in the market. As an example, software EdiTracker (now a plug-in of a forensic audio suite),[1] which scans audio files for ENF phase continuity and has been available for more than 10 years.

In this work, we propose a novel method for ENF-based audio authenticity in which phase discontinuity is assessed by means of causal and anti-causal filters checking the

[1] http://speechpro-usa.com/product/forensic_analysis/editracker.

behavior of the electric network signal in both directions, direct and time-reversed. The proposed scheme, due to its robustness, was evaluated using MP3 compressed audio files. This audio format is widely employed in practice and not well comprehensively addressed in the audio authenticity literature so far. The system performance is compared with two state-of-the-art algorithms achieving promising results. For evaluation purposes, a novel metric is also introduced. This metric takes into account not only if the algorithm is capable of detecting the manipulation of the audio, but also if the algorithm correctly detects the editing times.

The remaining of this paper is organized as follows. The next section details the data sets employed in our experiments. The new method is introduced in Sect. 2, while Sect. 3, devoted to the experiments carried out in this work, shows the promising results of the proposed method. Finally, Section 4 presents a few concluding remarks and outlines a discussion about future work.

2 The Proposed Method

We detail in the following all steps of the proposed method, named local phase change (LPHC).

2.1 Introduction

In order to analyze the ENF continuity in audio signals, it is usual to preprocess the signal by narrowband filtering it in order to extract the embedded ENF. If a phase discontinuity is present in the audio, it is smoothed due to this procedure and the phase change may be less acute. In order to avoid this effect, we propose to take measurements that, at each instant, do not mix past and future phase values of the signal under analysis. This can be achieved by making the measurement based on the use of causal and anti-causal filters. This local phase change measurement is post-processed and thresholded to obtain instants that are likely to correspond to the audio editing points. It is worth noting that any possible phase discontinuity is assumed present only during no voice activity periods of the audio under analysis. Any editing during voice activity would be easily recognized by a trained listener.

2.2 Local Phase Change Measurements (LPHC)

The causal and anti-causal filters are defined by means of complex exponentials as explained in the following.

Let us consider a complex exponential $E(t)$ in a limited interval of analysis:

$$E(t) = \mathrm{e}^{j2\pi f_{\mathrm{E}} t}, \quad -\frac{N}{f_{\mathrm{E}}} < t < \frac{N}{f_{\mathrm{E}}}, \tag{1}$$

where N is the number of cycles considered to estimate the phase of the ENF in the audio signal and f_{E} is the frequency of the component introduced by the electrical network. In order to measure past and future phase values locally, we define left and right complex exponentials, $E_L(t)$ and $E_R(t)$, as $E_L(t) = u(-t)E(t)$ and $E_R(t) = u(t)E(t)$, where $u(t)$ is the Heaviside function. Let $x(t)$ be the signal under analysis, initially assumed a perfect cosine with an arbitrary phase $x(t) = \cos(2\pi f_{\mathrm{E}}t + \phi)$. If we project signal $x(t)$ into $E(t)$ by applying the dot product

$$\langle x(t), E(t)\rangle = \int_{-\infty}^{\infty} E(t)x(t)\mathrm{d}t, \tag{2}$$

we obtain a value whose phase is the relative phase of $x(t)$ with respect to $E(t)$. If we express $x(t)$ as $\frac{\mathrm{e}^{j(2\pi f_{\mathrm{E}}t+\phi)}+\mathrm{e}^{-j(2\pi f_{\mathrm{E}}t+\phi)}}{2}$, and take into account that the first term of the product is null, we obtain

$$\langle x(t), E(t)\rangle = \int_{-\frac{N}{f_{\mathrm{E}}}}^{\frac{N}{f_{\mathrm{E}}}} \mathrm{e}^{j2\pi f_{\mathrm{E}}t}\,\frac{\mathrm{e}^{-j(2\pi f_{\mathrm{E}}t+\phi)}}{2}\mathrm{d}t = \frac{N\mathrm{e}^{-j\phi}}{f_{\mathrm{E}}}. \tag{3}$$

Then, the left and right relative phases of $x(t)$, with respect to the phase of $E(t)$, are

$$\begin{aligned}\phi_L &= \arg[\langle x(t), E_L(t)\rangle] = \arg\left[\frac{N\mathrm{e}^{-j\phi}}{2f_{\mathrm{E}}}\right] = -\phi \text{ and}\\ \phi_R &= \arg[\langle x(t), E_R(t)\rangle] = \arg\left[\frac{N\mathrm{e}^{-j\phi}}{2f_{\mathrm{E}}}\right] = -\phi.\end{aligned} \tag{4}$$

We can easily see that there is no phase change for the case of an ideal cosine, that is

$$\phi_{\text{change}} = \phi_R - \phi_L = 0. \tag{5}$$

If there is an instantaneous phase change at $t = 0$, we can write $x(t)$ as

$$x(t) = u(-t)\cos(2\pi f_{\mathrm{E}}t + \phi_1) + u(t)\cos(2\pi f_{\mathrm{E}}t + \phi_2). \tag{6}$$

Now, at the time of analysis, we obtain:

$$\begin{aligned}\phi_L &= \arg[\langle x(t), E_L(t)\rangle] = \arg\left[\frac{N\mathrm{e}^{-j\phi_1}}{2}\right] = -\phi_1 \text{ and}\\ \phi_R &= \arg[\langle x(t), E_R(t)\rangle] = \arg\left[\frac{N\mathrm{e}^{-j\phi_2}}{2}\right] = -\phi_2,\end{aligned} \tag{7}$$

such that the local phase change would be:

$$\phi_{\text{change}} = \phi_R - \phi_L = \phi_1 - \phi_2, \tag{8}$$

which gives the value of the phase change at the discontinuity.

In order to study the phase change along the whole signal, we can compute the relative phase in terms of convolutions:

$$\begin{aligned}\phi_L(t) &= \arg[x(t) * E_L(t)] \text{ and}\\ \phi_R(t) &= \arg[x(t) * E_R(t)];\end{aligned} \tag{9}$$

the local phase change being computed at every time instant t as

$$\phi_{\text{change}}(t) = \phi_R(t) - \phi_L(t). \tag{10}$$

The measurement of the phase change performed with left and right sided reference functions ensures that the information from one side does not contaminate the other one which otherwise tends to degrade the information about the actual edit point. At the same time, local phase discontinuities are estimated by computing the difference between the phase of the signal with respect to the same exponential. Any delay introduced by the filters would be present in both values (past and future measurements of the phase) and thus would not affect the difference.

2.3 Windowing

If the number of cycles used to estimate the phase of the ENF (N) is very small, the measurement of the phase may be affected by the interference that may be present around the components of the ENF. On the other hand, if N is too large, the phase is not measured very close to the value corresponding to the instant of analysis but at the center of the N cycles. A good tradeoff can be obtained if, instead of using an exponential with constant amplitude, we use a window that weights the measurement with a higher value closer to the time of analysis. Then, the analysis functions become

$$\begin{aligned}E_L(t) &= u(-t)E(t)H(t) \text{ and}\\ E_R(t) &= u(t)E(t)H(t),\end{aligned} \tag{11}$$

where $H(t)$ is the weighting window, for example a Hann window. While this in general produces results with a better behavior, the phase estimation is approximate, but not exact, in the case of an ideal cosine.

2.4 Post-processing

Due to fact that the ENF is not exactly, in mean, equal to the nominal ENF value, there is an offset that should be corrected. The relative phase for the right side ϕ_R in the case where the actual ENF, f_E', is different from the nominal one, f_{E}, can be calculated from the following expressions:

$$\begin{aligned}
\phi_R &= \arg\left[\int_0^{N/f_{\mathrm{E}}} \mathrm{e}^{j2\pi f_{\mathrm{E}} t}\, \frac{\mathrm{e}^{j(2\pi f'_E t+\phi_2)} + \mathrm{e}^{-j(2\pi f'_E t+\phi_2)}}{2} \mathrm{d}t\right] \text{ and} \\
\phi_R &= \arg\left[\frac{\mathrm{e}^{j\left(2\pi \frac{f'_E+f_{\mathrm{E}}}{f_{\mathrm{E}}} N+\phi_2\right)} - \mathrm{e}^{j\phi_2}}{j4\pi(f'_E + f_{\mathrm{E}})} - \frac{\mathrm{e}^{-j\left(2\pi \frac{f'_E-f_{\mathrm{E}}}{f_{\mathrm{E}}} N+\phi_2\right)} - \mathrm{e}^{-j\phi_2}}{j4\pi(f'_E - f_{\mathrm{E}})}\right].
\end{aligned} \tag{12}$$

Similarly, the left side phase is

$$\phi_L = \arg\left[\frac{\mathrm{e}^{j\phi_1} - \mathrm{e}^{j\left(-2\pi \frac{f'_E+f_{\mathrm{E}}}{f_{\mathrm{E}}} N+\phi_1\right)}}{j4\pi(f'_E + f_{\mathrm{E}})} - \frac{\mathrm{e}^{-j\phi_1} - \mathrm{e}^{-j\left(-2\pi \frac{f'_E-f_{\mathrm{E}}}{f_{\mathrm{E}}} N+\phi_1\right)}}{j4\pi(f'_E - f_{\mathrm{E}})}\right]. \tag{13}$$

We can see that $\phi_{\text{change}}(t) = \phi_R - \phi_L$ is independent of t. This introduces an offset which should be removed to obtain a signal with zero-mean, see Fig. 1. Although this can be achieved by subtracting the mean value of the phase change in the whole audio excerpt, removing the median is a more robust choice since the editing of the audio introduces a peak in $\phi_{\text{change}}(t)$:

$$\begin{aligned}
m_{\text{ch}} &= \text{median}(\phi_{\text{change}}(t)), \text{ and} \\
\phi'_{\text{change}}(t) &= \phi_{\text{change}}(t) - m_{\text{ch}}.
\end{aligned} \tag{14}$$

The normal variations of the ENF and the interference in low frequencies around the ENF value produce variations in the local phase change. These variations deviate the value of the measured phase with positive and negative drifts; they, when averaged in a time window long enough, bring the balance to a value that is close to zero. As the dynamics of these variations range in very low frequencies, approximately 1–4 Hz, and is a priori unknown, we use a post-processing procedure to take the minimum of the signal averaged in different time lengths. Let us define the averaging window of time T_i as

$$W_{T_i}(t) = W(t/T_i), \tag{15}$$

where W is an arbitrary window, i.e., a Hann window:

$$W(t) = \frac{1}{2}\left[u(t+0.5) - u(t-0.5)\right](1 + \cos(\pi t)). \tag{16}$$

Then, the undesired variations are diminished by taking the minimum of a set of filters with different averaging times, weighted with their time length in order to make them comparable

$$\phi''_{\text{change}}(t) = \min_i \left\{T_i W_{T_i}(t) * \phi'_{\text{change}}(t)\right\}, \tag{17}$$

where $T_i \in \{0.5, 0.6, \ldots, 1.7, 1.8\}$ (in seconds) was chosen to cover different time scales. This allows to attenuate large local oscillations due to audio interference in the ENF while not affecting strongly the isolated peaks that appear as a consequence

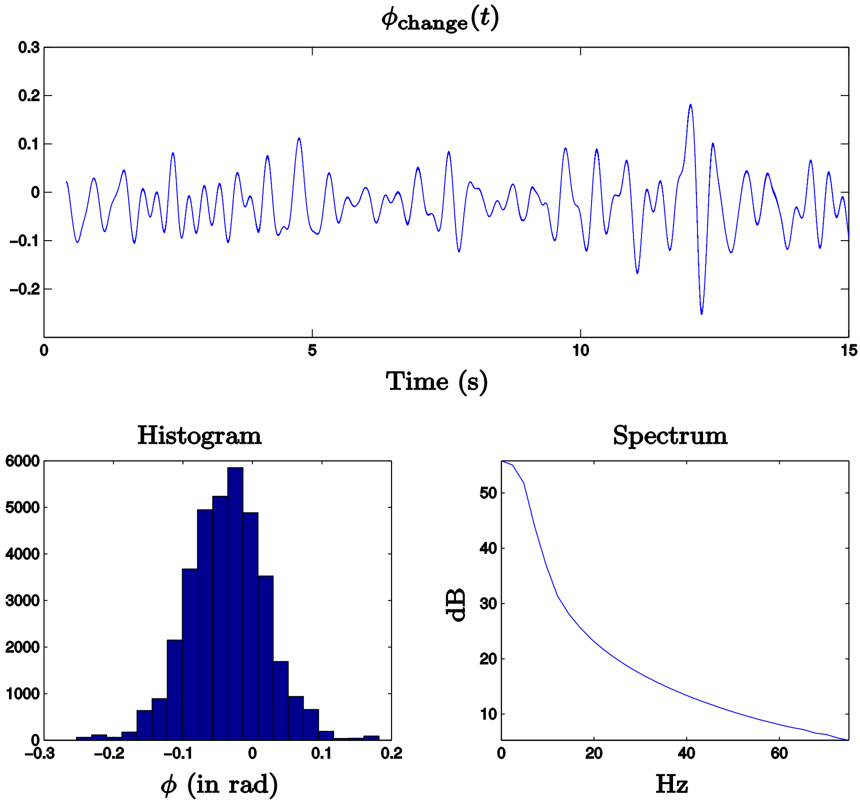

Fig. 1 Example of a typical variation of the phase change for one excerpt. On top, the value of ϕ_{change} in some seconds of the excerpt; on bottom, the histogram and spectrum of ϕ_{change}

of forgery; see Fig. 2 for two examples. It can be observed that isolated peaks that correspond to an edited point are kept while quick oscillations are dampened. Finally, $\phi''_{\text{change}}(t)$ is compared with a threshold γ to detect the existence of an editing point.

2.5 Voice Activity Detection

The presence of voice in the audio could severely interfere the ENF signal, making extremely difficult to detect phase changes. Also, it is a reasonable assumption that when an audio excerpt is forged, in order to avoid making the editing evident, it will be edited at times where there is no voice activity. As a way to avoid detecting edit points during voice activity, we use a *voice activity detection* (VAD) method to determine if the voice is present.

In this work, algorithms LPHC and PLL were applied using the same criteria as in [6]. This allows to focus the comparison on the detection of the phase change independently of the voice interference.

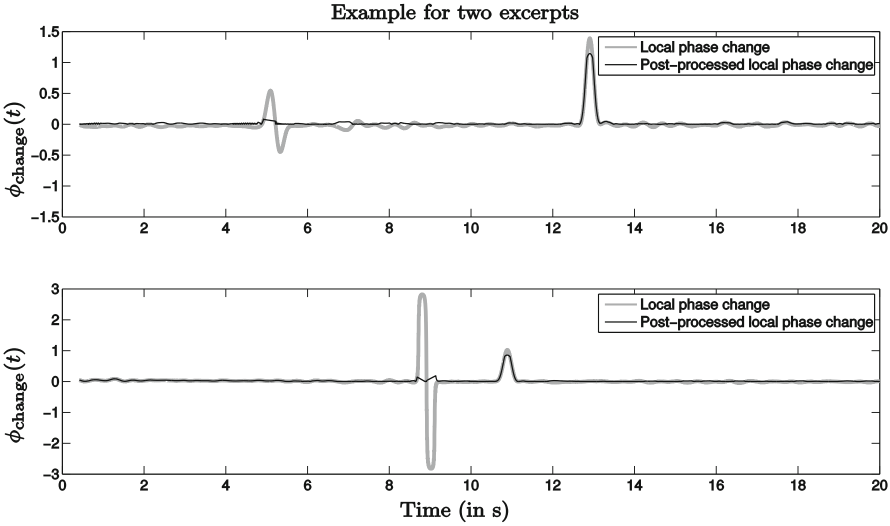

Fig. 2 Example of two *difficult* audio files in which there is one time which corresponds to a forgery (second 13 in the upper and 11 in the lower) and another that is a strong local oscillation due to interference in the ENF (at second 5 in the upper and second 9 in the lower). The black plot is the local phase measurement (ϕ_{change}) and the gray one is the same after the post-processing (ϕ''_{change})

3 Experiments and Results

3.1 Datasets

The databases used for evaluating the algorithms are CARIOCA [21] and AHUMADA [23]. Each database has a total of 100 original files and 100 corresponding edited files. Half of the edited files have one insertion and the remaining have one deletion. One of the main differences between Databases AHUMADA and CARIOCA is that the ENF components are 50 and 60 Hz, respectively.

3.1.1 Ground Truth Extraction

Edited signals of both AHUMADA and CARIOCA databases do not have temporal editing marks. This ground truth information is, to our understanding, important to evaluate correctly the performance of the algorithms. To accomplish this task, a dynamic time warping algorithm (DTW) [27] was implemented. The algorithm calculates the spectrogram of the original and edited audio files and measures the distance between the spectrum S_o at the frame i of the original file and the spectrum S_e at the frame j of the edited file as:

$$c(i,j) = \min\left\{\frac{\sum_{k=0}^{N/2-1} ||S_o(i,k)| - |S_e(j,k)||}{\sum_{k=0}^{N/2-1} ||S_o(i,k)|||S_e(j,k)||}, 0.5\right\}, \tag{18}$$

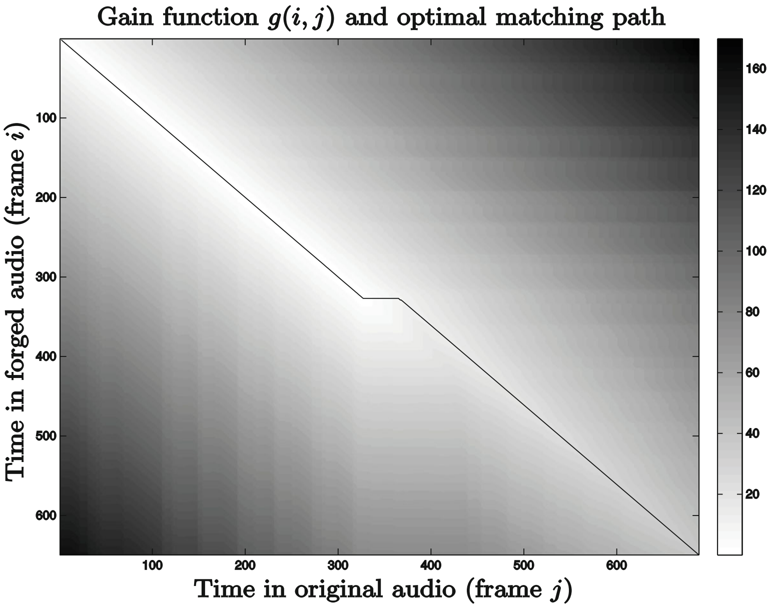

Fig. 3 Example of the distance map, showing the gain function $g(i, j)$, and the optimal path matching original and forged audio files. The horizontal section in this path corresponds to the deletion of a part of the audio as there is a jump in the corresponding time

where N is the spectrogram window size and $S(i, k)$ is the k-th frequency bin of the FFT of frame i. This normalization is employed to allow similarity measurements, independently of the intensity of the audio signals. The difference is saturated to a value of 0.5 in order not to overweight large differences in the spectrum. The gain function in the dynamic programming setup is defined as

$$g(i, j) = \min\{g(i-1, j-1), g(i-1, j), g(i, j-1)\} + c(i, j). \tag{19}$$

After the best alignment is computed for the minimum value of g of the complete original and edited sequence of frames, the edit points can be easily detected looking for time jumps in the alignment, as the example depicted in Fig. 3: A horizontal section in the match corresponds to the deletion of that section of the audio as there is no corresponding match of the original audio in the forged one. A vertical jump is due to the insertion of audio, in the case of a forged excerpt.

3.1.2 Compression

In recent years, the problem of determining authenticity in MP3 compressed audio recordings has received considerable attention. This is because MP3 is a common audio format widely used for consumer audio storage as well as a standard for music playback on digital audio players and for transfers over the internet. Some recent works have dedicated significant effort to detect fake quality audio tracks (e.g., dou-

Table 1 Databases bitrates

Database	f_s (Hz)	Bitrates (kbps)
CARIOCA	44,100	128–64–32–16
AHUMADA	8000	64–32

ble compression) [4,24,25], while others focused on localizing possible tampered portions in audio files [26]. For instance, a method using a machine learning algorithm to decide if an audio recording has been transcoded (from a lower to a higher bitrate) was introduced in [4]; it is based on a support vector machine (SVM) classifier where five different bitrates were chosen as classes. Another recent work based on SVM presents an estimation of signal parameters via rotational invariance techniques (ESPRIT) scheme for tampering detection [26]. Also, other methods were proposed to determine both double compression and tampering based on modified discrete cosine transform (MDCT) coefficients information [2,17,18]. Furthermore, a scheme for detecting the more common forgeries as deletion or insertion using frame offsets was presented, reporting interesting results [30]. To evaluate the robustness to compression of the algorithms under investigation, we have created six new databases with different bitrates. Table 1 shows these values for each database. Due to the low sampling rate of AHUMADA database, the maximum bitrate in that case is 64 kbps. The new databases were coded and decoded using an MP3 codec (libmp3lame) included in *FFmpeg* tool [8]. Because we have uncompressed original and edited files, the process to create this database differs from what could be expected a normal process of manipulating an MP3 file, which would be: (1) decode a compressed file, (2) edit it, and (3) code it again. In this work, the first step is missing as the databases are composed of uncompressed audio files already edited. We present a causal and an anti-causal ENF estimation for determining editing times in audio signals. We report the results using MP3 signals with different bitrates and a novel and more rigorous assessment. The proposed method is described in the following section.

3.2 Experiments

The performance of the proposed method was compared with two recently proposed methods, referred to as PLL [9] and TMVAD (template matching with voice activity detection) [6].Those approaches, having presented better results than previous works for the databases under investigation, were assumed the state-of-the-art in ENF-based methods for audio authentication. For that reason, we use them as the baseline for the present work. A brief description of those algorithms is presented in the following. Afterward, a study on the robustness of the three algorithms to compression artifacts is introduced.

3.2.1 PLL and TMVAD Algorithms

Both PLL and TMVAD algorithms are composed of three main stages: preprocessing, ENF estimation and post-processing or/and decision step.

The PLL is composed of a preprocessing step in which the signal is down-sampled and filtered with a sharp bandpass filter centered in the ENF nominal value. The obtained signal is then normalized using an estimation of its envelope. The second stage performs the estimation of the ENF phase variations using a phase-locked loop that compares the ENF with a synthetic reference. Finally, the variations are smoothed with a moving average filter and compared with a fixed threshold which was empirically determined by studying the databases CARIOCA and AHUMADA.

The TMVAD algorithm preprocessing stage consists of a sample rate reduction, followed by a voice activity detector and a narrow bandpass filter centered at the ENF nominal value. The ENF estimation is then carried out via Hilbert's method, which detects subtle variations in the ENF frequency and is considered sensitive to background noise. The ENF variations are computed by taking the absolute value of the median-compensated of the estimated ENF. Finally, for those points in which the variations are above a variable threshold, the correlation with prototypical templates generated from synthetic ENF is computed. An edit point is detected when the maximum correlation for the set of templates is above a threshold.

3.2.2 Performance Assessments

In previous works, the performance of the algorithms was measured in terms of the equal error rate (EER) operating point [5,6,9]. The EER value is calculated for a threshold γ such that

$$\text{EER} = P_{\text{FP}}(\gamma) = P_{\text{FN}}(\gamma), \tag{20}$$

where $P_{\text{FP}}(\gamma)$ and $P_{\text{FN}}(\gamma)$ are the false positive and the false negative probabilities. In this context, $P_{\text{FP}}(\gamma)$ is usually calculated as the probability of recognizing at least one editing in original files. Similarly, $P_{\text{FN}}(\gamma)$ is generally computed as the probability of not recognizing an edition in manipulated files. This evaluation method does not make a comparison between the editing times detected for the algorithms and the real editing times.

In this work, we propose a novel method to compute $P_{\text{FP}}(\gamma)$ and $P_{\text{FN}}(\gamma)$ using the information of temporal editing marks. This method obtains more trustworthy results. We start by defining:

- m_i, $i = 1, \ldots, N$: temporal marks of editing (ground truth) extracted as explained in Sect. 3.1.1.
- t_j, $j = 1, \ldots, M$: temporal points that the algorithm finds as an editing (points of insertion or deletion).
- $\tau = 250$ ms: temporal tolerance.[2]

Therefore, for audio file k, we check for the existence of a correct classification (true positive or true negative), OK[k], of false positive, FP[k], or of a false negative, FN[k], as follows.

[2] The tolerance parameter τ was chosen to avoid losing editing points. Our assumption is that the duration of a short word is at least 500 ms, so in the case of the insertion of short words, the value $\tau = 250$ ms permits an evaluation of the system's performance without losing any editing point.

Table 2 EER (in %) results of usual approach

	AHUMADA	CARIOCA
LPHC	3.0	**1.5**
TMVAD	**1.0**	2.0
PLL	4.5	2.0

The bold values indicate that the best result for each experiment

$$\mathrm{OK}[k] = \begin{cases} 1 & \text{if } \forall i, \exists j \ / \ |m_i - t_j| < \tau \ \text{ and } N = M \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{FP}[k] = \begin{cases} 1 & \text{if } \exists j \ / \ |m_i - t_j| > \tau \ \forall i = 1, \ldots, N \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{FN}[k] = \begin{cases} 1 & \text{if } \exists i \ / \ |m_i - t_j| > \tau \ \forall j = 1, \ldots, M \\ 0 & \text{otherwise} \end{cases}$$

Keeping in mind that a given piece of audio can fulfill both FP and FN conditions simultaneously, we calculate, for a set of signals employed in the evaluation of each audio authenticity method, the total numbers of OK, FP, and FN as follows:

1. $N_{\mathrm{OK}} = \sum_k \mathrm{OK}[k]$;
2. $N_{\mathrm{FP}} = \sum_k \mathrm{FP}[k]$; and
3. $N_{\mathrm{FN}} = \sum_k \mathrm{FN}[k]$.

Finally, we estimate the probability of false positives as:

$$P_{\mathrm{FP}} = \frac{N_{\mathrm{FP}}}{N_{\mathrm{OK}} + N_{\mathrm{FP}} + N_{\mathrm{FN}}} \tag{21}$$

and the probability of false negatives as

$$P_{\mathrm{FN}} = \frac{N_{\mathrm{FN}}}{N_{\mathrm{OK}} + N_{\mathrm{FP}} + N_{\mathrm{FN}}}. \tag{22}$$

3.2.3 Performance Evaluation with Original PCM Signals

Table 2 shows the results of comparing the novel algorithm (LPHC) with previous works using the same evaluation assessment than in [6].

To compare the performance of the algorithms in terms of this new and more rigorous assessment, we compute the detection error tradeoff (DET) curve that plots $P_{\mathrm{FP}}(\gamma)$ versus $P_{\mathrm{FN}}(\gamma)$ for various γ values. Each database (AHUMADA and CARIOCA) is tested separately. Figure 4 shows the comparative DET curves of the three algorithms testing with AHUMADA and CARIOCA databases, respectively.

As shown in Fig. 4, it is easy to notice that testing with CARIOCA database, LPHC and TMVAD have similar performance, and PLL is the algorithm that performs worse in most of the cases for both databases. In turn, when testing with AHUMADA, LPHC obtains better results than the other algorithms. If we choose the EER as an operating point, it is possible to compare the performances numerically, as shown in Table 3. In brief, it can be concluded that the algorithms investigated here have better results for the CARIOCA *corpus*. It is also clear that the LPHC algorithm has better performance

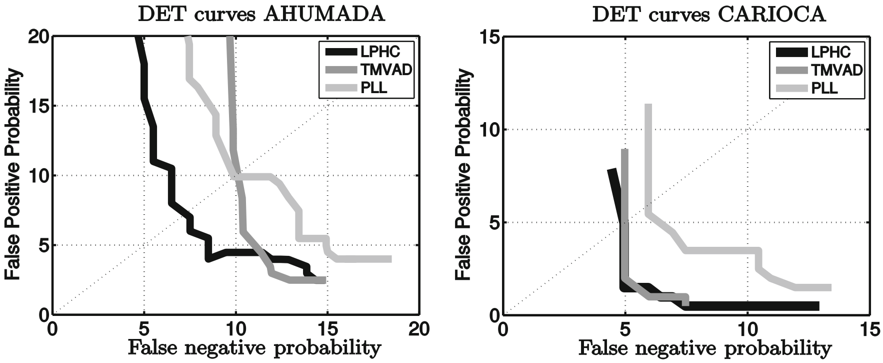

Fig. 4 Comparative DET curves for the different algorithms on AHUMADA (left) and CARIOCA (right) databases (PCM files) with the new evaluation method

Table 3 Results of EER (in %) with the new temporal assessment method

	AHUMADA	CARIOCA
LPHC	**7.4**	**4.9**
TMVAD	10.1	5.0
PLL	9.9	5.9

The bold values indicate that the best result for each experiment

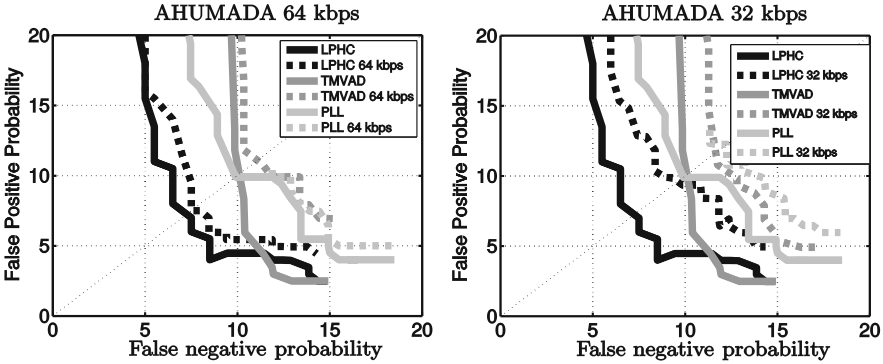

Fig. 5 Comparative DET curves for the different algorithms on AHUMADA 64 kbps (left) 32 kbps (right) databases (MP3 files) with the new evaluation method

for the EER operating point computed with the assessment method described herein than the others in both databases, AHUMADA and CARIOCA.

3.2.4 Performance Comparison with MP3-Coded Signals

In this experiment, we compare the result of testing the algorithms in coded databases to analyze their robustness to compression artifacts. Figure 5 shows DET curves comparing algorithms results in original and compressed AHUMADA database. Table 4 compares the EER for this case.

Table 4 Comparing EER (in %, with the temporal assessment method) for original and compressed AHUMADA database

	Original	64 kbps	32 kbps
LPHC	**7.4**	**8.4**	**9.8**
TMVAD	10.1	11.1	11.7
PLL	9.9	11.4	12.3

The bold values indicate that the best result for each experiment

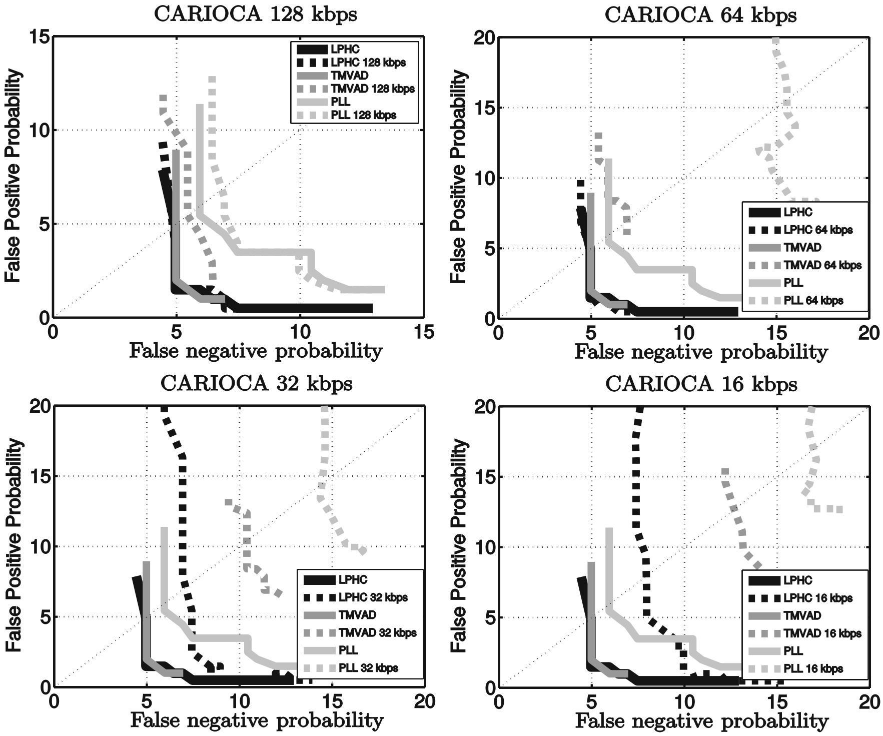

Fig. 6 Comparative DET curves for the different algorithms on CARIOCA database (MP3 files) with the new evaluation method

In a similar way, Figure 6 and Table 5 show the results for original and compressed versions of the CARIOCA database. As seen in the results, among those algorithms under investigation, the LPHC is the most robust one for MP3-coded signals.

3.2.5 Performance with Clipping and Additive Noise

In order to show that LPHC works better not only under the effects of MP3 compression, but also other kinds of distortions, two experiments were carried out as described in the following. The first one consists in adding white noise with a SNR of 10 dB and the second one in clipping the signal at 0.5% of the samples. The experiments

Table 5 Comparing EER (in %, with the temporal assessment method) for original and compressed CARIOCA Database

Algorithm	Original	128 kbps	64 kbps	32 kbps	16 kbps
LPHC	**4.9**	**5.0**	**5.0**	**7.1**	**8.0**
TMVAD	5.0	5.5	6.9	10.4	12.7
PLL	5.9	6.8	15.6	14.4	17.0

The bold values indicate that the best result for each experiment

Table 6 EER (method with temporal edition marks) for CARIOCA and AHUMADA Databases with noise (10 dB SNR) and clipping (0.5%)

	AHUMADA			CARIOCA		
	Original	Noise	Clipping	Original	Noise	Clipping
LPHC	**7.4**	**33.8**	**14.8**	**4.9**	**33.0**	**8.70**
TMVAD	10.1	39.6	18.9	5.0	34.9	14.4

The bold values indicate that the best result for each experiment

were evaluated with methods LPHC and TMVAD, but not with the PLL method as its results are sensitively worse. In all the experiments the LPHC outperforms the TMVAD method. This can be seen in Table 6 where we observe that the LPHC method has lower EER values in all cases.

3.3 Analysis of the Results

Figure 7 shows estimates of the probability density functions (PDF) of the phase change between two consecutive samples measured before post-processing at all instants of the non manipulated audio files and at the ground truth edition points of each file of the manipulated database. If the ENF was exactly at the nominal value at all times, without any additive noise nor any other interference, the phase variation should have a null value at all times. The AHUMADA database presents more variability around the nominal ENF than the CARIOCA database. This variability has a direct impact on the measurement of the phase change for its larger deviation from zero.

The phase change of the ENF at an editing (cut or insertion) time is expected to be distributed uniformly in the range $(-\pi, \pi]$ [21], which can be noted in Fig. 7. Values that are close to 0 may be undetectable as they are masked by the natural variations of the ENF, making virtually impossible the use of the ENF as an indication that the file has been edited.

Relatively large phase changes due to noise or interference of the audio content may produce false positives in a tradeoff of the amount of false negatives that are allowed.

As we have the editing times ground truth, if we make the reasonable assumption that there is no voice activity at the times where an edit was made, we can measure the actual phase change at those times. The LPHC measurement, without applying the post-processing described in Sect. 2.4, is a good estimate of that value, and it

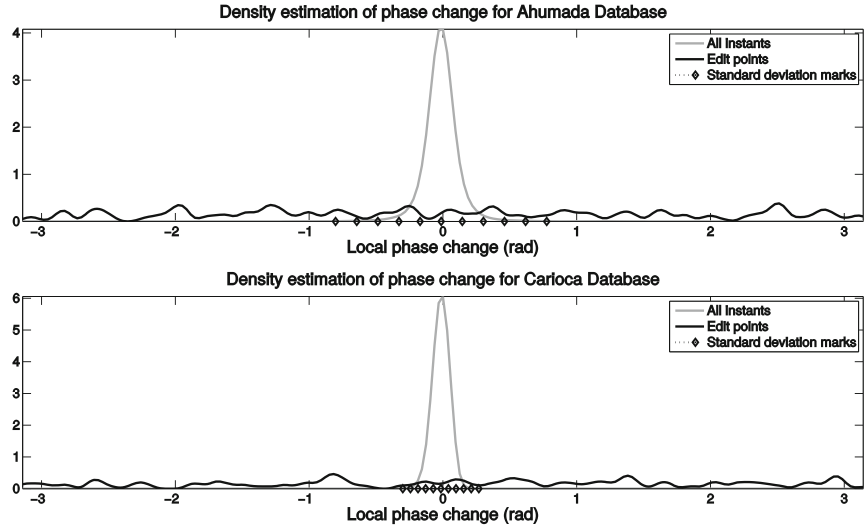

Fig. 7 Estimated PDFs of phase change obtained from *corpora* AHUMADA and CARIOCA. The standard deviation (multiples marked with $\diamond$ in the horizontal axis) of the phase change at all instants without any manipulation are 0.1578 and 0.0568 radians, respectively. The corresponding mean values are − 0.0129 and − 0.0155 radians

becomes useful to study the distribution of phase changes for false negative values in each algorithm.

Figure 8 shows the false negatives phase change measurements which are concentrated at low angular values, as expected. The LPHC technique appears to be the one for which those values are smaller, which suggests that those false negatives are harder to detect due to a small phase change in the fundamental component of the ENF. A small or even null phase change can occur by chance or when the audio manipulation is carried out by someone who knows the problem and has the technical skills to avoid it. In any case, it is worth mentioning that, when using the proposed method, not detecting evidence does not imply that the signal is free of tampering.

Any harmonic of the ENF present in the audio may also be used to detect phase changes [14,22]. As the phase change is amplified proportionally to the number of harmonics, small values in the fundamental frequency may reach values that become easily observable in the second or third harmonic. The presence, as observed in our experiments, of the second harmonic in the AHUMADA database and of the third harmonic in the CARIOCA database, suggests that there is more potential in taking them into account rather than improving the detection based exclusively on the fundamental component.

It seems surprising that with some extent of MP3 compression, the algorithms still are able to detect local phase changes in MP3 as it is usually stated that MP3 does not keep the phase of the signals. Our understanding based on [3] is that MP3 codifies differently the components that are tonal and the ones that are not. If the phases of the same tonal component in two consecutive time frames were distorted, it

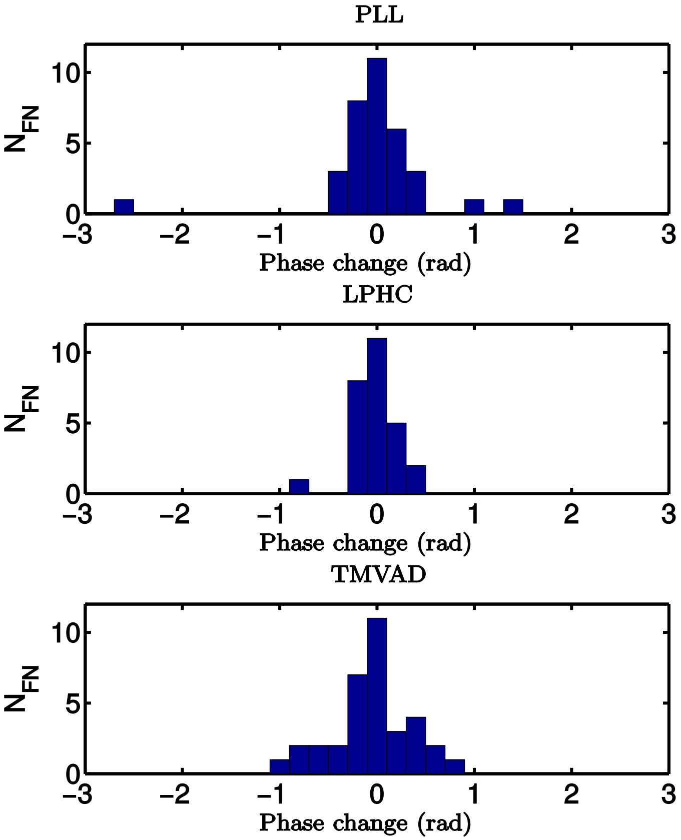

Fig. 8 Histograms of false negatives according to phase change

would clearly affect the audio quality as the tonal nature would disappear if they are not coherent anymore. Since the ENF component has a tonal nature, this holds. We checked this property by comparing the spectrograms of a signal with a slowly varying tonal component and noise with itself after the coding and decoding processes. In the residual, the tonal component almost disappeared and the noise was amplified as its phase was destroyed and a different noise was the result of MP3 compression.

4 Discussion and Future Work

A new method for evaluating audio authenticity, the LPHC, is presented in this work. The new technique measures the instant phase change of the ENF avoiding the smoothing of phase discontinuities.

Furthermore, a novel evaluation metric is introduced in order to assess the accuracy of the algorithm for computing the times of editing of the audio signal.

The proposed method behaves similarly as state-of-the-art algorithms for the detection of audio editing points in uncompressed audio. Moreover, the technique was tested

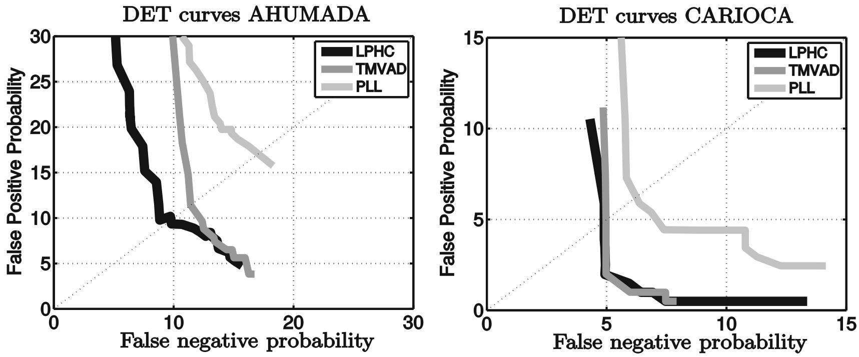

Fig. 9 DET curves of point-based method trained on AHUMADA (left) and CARIOCA (right) database

in compressed audio files and compared with previous works. In this scenario, the LPHC outperforms the other algorithms presenting more robustness to MP3 compression at different ratios. This result is of great interest as nowadays many recording devices compress the audio without giving the possibility to access the raw original data, i.e., dealing with compressed audio seems to be the most likely scenario in modern forensic analysis.

Regarding the evaluation metric introduced in this work, it takes into account not only if the algorithm is capable of detecting the manipulation of the audio, but also if the algorithm detects correctly the editing times. This evaluation is a more realistic measurement of the performance of the algorithm than comparing the output of the algorithm to an edited/original label, false positives and negatives being computed in a more accurate manner. The case of the algorithm detecting a false alarm (e.g., a phase abrupt change produced without audio editing) is not detected if only the information of edited/original is available. For a more trustworthy evaluation, the editing times and the detecting times of the algorithms were compared.

However, it may be reasonable to study the performance of each method in a time-of-editing based way and not in a file oriented way, following pixel-based evaluation criteria commonly used in image forensics [7]. In this case, the evaluation should take into consideration FP and FN of each recognized editing time, which is a more strict measurement than those presented in previous works and even than the one introduced here. Figure 9 shows the DET curves of this new evaluation method for the CARIOCA and AHUMADA databases.

As a suggestion for future work, an indication of the level of confidence could be estimated from statistics of the audio such as noise level and duration of excerpt under analysis, besides the value of the phase change itself. It would be desirable to use a larger and more diverse database than CARIOCA and AHUMADA in order to have more reliable results in this new approach. Also, a natural option worth investigating is using machine learning techniques to make a decision on the audio authenticity using as features the output of the different algorithms (LPHC, TMVAD [6], PLL [9], and SPHINS [26], for instance). Including the harmonics for a better estimation of the ENF should also be object of future research as well as linear modeling or even nonlinear fitting of the ENF.

References

1. I.F. Apolinário , C.A. Rossi, Time-stamp of digital audio recording based on the ENF estimated from another audio signal, in *I IEEE Latin American Symposium on Circuits and Systems (LASCAS)* (2010)
2. T. Bianchi, A.D. Rosa, M. Fontani, G. Rocciolo, A. Piva, Detection and localization of double compression in MP3 audio tracks. EURASIP J. Inf. Secur. **2**, 1–14 (2014)
3. M. Bosi, R.E. Goldberg, *Introduction to Digital Audio Coding and Standars* (Kluwer Academic Publishers, Alphen aan den Rijn, 2003)
4. B. D'Alessandro, Y.Q. Shi, MP3 bit rate quality detection through frequency spectrum analysis, in *Proceedings of the 11th ACM Workshop on Multimedia and Security*, pp. 57–62. ACM, New York, (2009). https://doi.org/10.1145/1597817.1597828
5. P.A.A. Esquef, J.A. Apolinário Jr., L.W.P. Biscainho, Edit detection in speech recordings via instantaneous electric network frequency variations. IEEE Trans. Inf. Forensics Secur. **9**(12), 2314–2326 (2014). https://doi.org/10.1109/TIFS.2014.2363524
6. P.A.A. Esquef, J.A. Apolinário Jr., L.W.P. Biscainho, Improved edit detection in speech via ENF patterns, in *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6 (2015). https://doi.org/10.1109/WIFS.2015.7368585
7. H. Farid, Image forgery detection. IEEE Signal Process. Mag. **26**(2), 16–25 (2009)
8. FFmpeg: A complete, cross-platform solution to record, convert and stream audio and video. http://www.ffmpeg.org. Accessed 10 June 2016
9. M. Fuentes, P. Zinemanas, P. Cancela, J.A. Apolinário Jr., Detection of ENF discontinuities using PLL for audio authenticity, in *VII IEEE Latin American Symposium on Circuits and Systems (LASCAS)*, pp. 79–82 (2016). https://doi.org/10.1109/LASCAS.2016.7451014
10. R. Garg, A.L. Varna, A. Hajj-Ahmad, M. Wu, Seeing ENF: power-signature-based timestamp for digital multimedia via optical sensing and signal processing. IEEE Trans. Inf. Forensics Secur. **8**(9), 1417–1432 (2013)
11. C. Grigoras, Forensic analysis of digital recordings—the electric network frequency criterion, in *Forensic Science International* (2003)
12. C. Grigoras, Digital audio recording analysis, the electric network frequency (ENF) criterion. Int. J. Speech Lang. Law **12**, 43–49 (2005). https://doi.org/10.1558/sll.2005.12.1.63
13. C. Grigoras , Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis (Selected Articles of the 4th European Academy of Forensic Science Conference (EAFS2006) June 13–16, 2006 Helsinki, Finland), pp. 136–145 (2007). https://doi.org/10.1016/j.forsciint.2006.06.033. http://www.sciencedirect.com/science/article/pii/S0379073806004312
14. A. Hajj-Ahmad, R. Garg, M. Wu, Spectrum combining for ENF signal estimation. IEEE Signal Process. Lett. **20**(9), 885–888 (2013)
15. A. Hajj-Ahmad, R. Garg, M. Wu, ENF-based region-of-recording identification for media signals. IEEE Trans. Inf. Forensics Secur. **10**(6), 1125–1136 (2015). https://doi.org/10.1109/TIFS.2015.2398367
16. M. Huijbregtse, Z. Geradts, Using the ENF criterion for determining the time of recording of short digital audio recordings, in *Springer Series Lecture Notes in Computer Science* vol. 5718, pp. 116–124 (2009)
17. R. Korycki, Authenticity investigation of digital audio recorded as MP3 files. Probl. Kryminal. **283**(1), 54–67 (2014)
18. Q. Liu, A.H. Sung, M. Qiao, Detection of double MP3 compression. Cogn. Comput. **2**(4), 291–296 (2010). https://doi.org/10.1007/s12559-010-9045-4
19. R.C. Maher, Audio forensic examination: authenticity, enhancement, and interpretation. IEEE Signal Process. Mag. **26**(2), 84–94 (2009). https://doi.org/10.1109/MSP.2008.931080
20. D.P. Nicolalde-Rodríguez, J.A. Apolinário Jr., Evaluating digital audio authenticity with spectral distances and ENF phase change, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1417–1420 (2009)
21. D.P. Nicolalde-Rodríguez, J.A. Apolinário Jr., L.W.P. Biscainho, Audio authenticity: detecting ENF discontinuity with high precision phase analysis. IEEE Trans. Inf. Forensics Secur. **5**(3), 534–543 (2010). https://doi.org/10.1109/TIFS.2010.2051270

22. D.P. Nicolalde-Rodríguez, J.A. Apolinario Jr., L.W.P. Biscainho, Audio authenticity based on the discontinuity of ENF higher harmonics, in *Proceedings of the 21st European Signal Processing Conference (EUSIPCO), IEEE*, pp. 1–5 (2013)
23. J. Ortega-Garcia, J. Gonzalez-Rodriguez, V. Marrero-Aguiar, AHUMADA: A large speech corpus in Spanish for speaker characterization and identification. Speech Commun. **31**(2), 255–264 (2000). https://doi.org/10.1016/S0167-6393(99)00081-3. http://www.sciencedirect.com/science/article/pii/S0167639399000813. Accessed 20 July 2017
24. M. Qiao, A.H. Sung, Q. Liu, Revealing real quality of double compressed MP3 audio, in *Proceedings of the 18th ACM International Conference on Multimedia, MM '10*, pp. 1011–1014. ACM, New York (2010). https://doi.org/10.1145/1873951.1874137
25. M. Qiao, A.H. Sung, Q. Liu, Improved detection of MP3 double compression using content-independent features, in *2013 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2013)*, pp. 1–4 (2013). https://doi.org/10.1109/ICSPCC.2013.6664121
26. P.M.G.I. Reis, J.P.C.L. da Costa, R.K. Miranda, ESPRIT-Hilbert-based audio tampering detection with SVM classifier for forensic analysis via electrical network frequency. IEEE Trans. Inf. Forensics Secur. **12**(4), 853–864 (2017). https://doi.org/10.1109/TIFS.2016.2636095
27. H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. **26**, 43–49 (1978)
28. R.W. Sanders, Digital audio authenticity using the electric network frequency, in *Audio Engineering Society Conference: 33rd International Conference: Audio Forensics-Theory and Practice* (2008). http://www.aes.org/e-lib/browse.cfm?elib=14403. Accessed 20 July 2017
29. M.C. Stamm, M. Wu, K.J.R. Liu, Information forensics: an overview of the first decade. IEEE Access **1**, 167–200 (2013)
30. R. Yang, Z. Qu, J. Huang, Exposing MP3 audio forgeries using frame offsets. ACM Trans. Multimed. Comput. Commun. Appl. **8**(2S), 35:1–35:20 (2012). https://doi.org/10.1145/2344436.2344441