Edit Detection in Speech Recordings via Instantaneous Electric Network Frequency Variations

Paulo Antonio Andrade Esquef, *Member, IEEE*, José Antonio Apolinário, Jr., *Senior Member, IEEE*, and Luiz W. P. Biscainho, *Member, IEEE*

***Abstract*— In this paper, an edit detection method for forensic audio analysis is proposed. It develops and improves a previous method through changes in the signal processing chain and a novel detection criterion. As with the original method, electrical network frequency (ENF) analysis is central to the novel edit detector, for it allows monitoring anomalous variations of the ENF related to audio edit events. Working in unsupervised manner, the edit detector compares the extent of ENF variations, centered at its nominal frequency, with a variable threshold that defines the upper limit for normal variations observed in unedited signals. The ENF variations caused by edits in the signal are likely to exceed the threshold providing a mechanism for their detection. The proposed method is evaluated in both qualitative and quantitative terms via two distinct annotated databases. Results are reported for originally noisy database signals as well as versions of them further degraded under controlled conditions. A comparative performance evaluation, in terms of equal error rate (EER) detection, reveals that, for one of the tested databases, an improvement from 7% to 4% EER is achieved, respectively, from the original to the new edit detection method. When the signals are amplitude clipped or corrupted by broadband background noise, the performance figures of the novel method follow the same profile of those of the original method.**

***Index Terms*— Acoustical signal processing, edit detection, spectral analysis, instantaneous frequency, voice activity detection.**

I. Introduction

ELECTRIC Network Frequency (ENF) analysis has been demonstrated in the literature to be a powerful tool to audio forensic tasks such as audio authentication, time-of-recording estimation, audio edit detection [1]–[7]. The first two applications involve the so called ENF Criterion which, in brief terms, consists in extracting the ENF component (ENFC) from the audio signal of interest and comparing its frequency pattern with those of referential ENF records stored in properly designed databases. Since frequency variations of the ENF are unique over time, one can determine the time-of-recording for authentication purposes, if a matching is found [1].

The success of the ENF Criterion depends on the quality of both the database and the ENF analysis. The importance of a reliable referential ENF database for forensic examination purposes is discussed in [8], where best practices for database production are suggested.

As regards ENF estimation and analysis, a comparative study of several high-resolution frequency estimators has been reported in [9], together with results of an application on recording location. In [10], the authors evaluate the use of quadratic interpolation over DFT samples as a means to improve the accuracy of ENF estimation. Nonparametric, adaptive, and high-resolution techniques are reported in [11] as ways to improve ENF estimation in the presence of interfering sources. In [12] the authors first run statistical analysis on ENF signals from a wide-area frequency monitoring network in different time scales. Then, ENF estimation with refined short-time Fourier analysis and oscillation error correction [13] are carried out for forensic audio authentication.

Modeling the ENF signal via an autoregressive process is proposed in [14]; the authors use a model-based decorrelation method in the process of ENF matching to improve the performance of timestamp verification tasks. In [15] a multitone harmonic model is used to build a Maximum-Likelihood estimator for the ENF. As with [15], ENF estimation using multiple harmonics is also investigated in [16]. In line with the rationale behind the ENF Criterion [1]–[3], the maximum cross correlation between the extracted ENF signal and the reference signal is reported in [17] to be closely related to the phase of the ENF signal. Therefore, the measure can be used for the purposes of forensic audio authentication, edit detection, and spatial location.

Difficulties in the application of the ENF Criterion for authentication of forensic audio are investigated in [2]. Among them, the authors cite signals with low level ENFC as well as its degradation due to noise contamination and low bit-rate coding. Other source of ENF interference is the

Manuscript received April 15, 2014; revised July 23, 2014 and October 7, 2014; accepted October 8, 2014. Date of publication October 16, 2014; date of current version November 12, 2014. This work was supported in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil, under Grant 475566/2012-2 and Grant 304800/2013-9 and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Pró-Defesa under Grant 23038.009094/2013-83. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. C.-C. Jay Kuo.

P. A. A. Esquef is with the National Laboratory of Scientific Computing, Petrópolis 25651-075, Brazil (e-mail: pesquef@lncc.br).

J. A. Apolinário Jr. is with the Military Institute of Engineering, Rio de Janeiro 22290-270, Brazil (e-mail: apolin@ieee.org).

L. W. P. Biscainho is with the Federal University of Rio de Janeiro, Rio de Janeiro 21941-972, Brazil (e-mail: wagner@smt.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.2014.2363524

superposition of two or more ENFCs caused by recapturing or medium transfers of audio recordings [18]. Limitations of the ENF Criterion for short-duration recordings are investigated in [19]. The existence of ENF interference in battery-powered digital recorders is investigated and demonstrated in [20]. Anti-forensic measures that can prevent audio authentication and forgery detection via ENF analysis are discussed in [21].

ENF analysis finds use in the detection of intentional signal modifications such as cuts and inserts in audio. In [22] the authors propose measuring spectral distances and evaluating phase changes of the ENF for edit detection. Later in [23], they evaluate an edit detection scheme based on high-resolution ENF phase analysis, under realistic scenarios including signal contamination with noise and degradation by amplitude clipping. Edit detection in audio through analysis of ENF amplitude and phase changes is also addressed in [24], when cuts and inserts are made at the zero-crossing points and cross-fading is used. Analysis of higher ENF harmonics is employed in [25] for the purpose of edit detection in forensic audio.

ENF analysis for forensic purposes goes beyond audio examination. In [26] and [27] the authors describe ways to capture and analyze the ENF from signals related to indoor lighting and video cameras for the purpose of forgery detection.

In this paper, an edit detection method for forensic audio is proposed. Besides a series of modifications on the signal processing chain of the method introduced in [23], a novel data-driven threshold-based detection strategy is designed. As with [23], the rationale behind the method is that edits, such as cuts and inserts, are likely to break the quasi-periodical nature of the ENFC, thus producing jumps in its instantaneous phase and, consequently, anomalous and sudden variations in the ENF. Hence, unusual ENF variations can be effective indicators of edits in the audio signal under test (SUT).

As will be seen later in the paper, the process of isolating the ENFC from the SUT involves passing it through narrow passband filters that will strongly smooth any phase jumps of the ENF, estimated as the instantaneous frequency (IF) of the ENFC. Despite the temporal spread, searching for unusual variations of the ENF constitutes a practical and powerful solution for edit detection. For that a variable threshold is devised to serve as an upper limit for the ENF variations observed in unedited SUTs. Any local variation of the ENF exceeding the threshold is considered an edit occurrence.

The proposed method is evaluated in qualitative terms, through properly designed case studies. Furthermore, a comprehensive quantitative performance evaluation is conducted using two distinct annotated databases, with both unedited and edited signals. Results are reported for both databases in their original state (with noisy signals) as well as with the signals further degraded under controlled conditions. Comparative performance analyses are carried out between the proposed edit detection method and that published in [23].

The paper is organized as follows. Section II starts with an overview of the proposed edit detection method and then gives detailed descriptions of its processing steps. In Section III, a qualitative evaluation of the detection capabilities of the method w.r.t. edit duration and location, as well as to signal contamination with noise and amplitude clipping, is reported. A quantitative performance evaluation of the method by means of two distinct annotated databases, whose signals are degraded under controlled conditions, is presented in Section IV. Finally, conclusions are drawn in Section V.

II. Proposed Method

The proposed edit detection method for forensic audio is built on and develops from ideas presented in [23]. Any SUT is assumed to be available in digital PCM format, sampled at f_s Hz, and represented at a suitable wordlength, e.g., 16 bits. Moreover, its duration should lie within 10 s and 50 s. If necessary, audio extracts of longer duration can be conveniently processed in blocks.

A. Initial Assumptions

Besides having an ENFC, the following assumptions are made on the SUT:

A1: The ENFC should be the energy-dominant signal around its nominal frequency;
A2: Contamination with background noise should be low enough to assure A1;
A3: Automatic gain control has been turned off or set to low levels during recording;
A4: Other signal disruptions such as the presence of impulsive noise and nonlinear distortions, e.g., hard amplitude clipping, should be absent or occur at low levels.
A5: Edits (cuts or inserts) made in the signal should be inaudible, with their initial and end points located at voice-inactive passages of the signal.

As regards A1, narrowband speech signals recorded from PSTN phone calls meet the requirement, since the communication channel takes care of filtering out the frequency content below 100 Hz [28], thus easing the prominence of the interfering ENFC that lies in the low-frequency range.

In general, the SUT should be free of any signal distortion that causes significant variations in the ENF. The presence of background noise may decrease the local SNR around the ENFC, thus increasing the variance of the ENF estimate. Assumptions A3 and A4 relate to time-localized interferences that may cause sudden energy increases around the ENFC and, consequently, variations in the estimated IF of the ENFC [29]. Even being a restriction, A5 is reasonable since edits at voice-active parts of the signal would be more difficult to implement without leaving audible clues, and could be detected by spectral changes [22].

The assumptions on the SUT listed above are those for which the proposed edit detection method yields more reliable results. By no means they imply a limited applicability of the method only to unrealistic signals, produced under laboratory conditions. On the contrary, all database signals employed in this work are real-world speech recordings, made in a typical workplace (office) environment. As such, they originally contain background noise and interferences coming from various everyday sound sources, as can be verified by listening to

TABLE I
PROCESSING STEPS OF THE PROPOSED EDIT DETECTION METHOD

Step	Description
1	Sampling rate reduction of the SUT $x[n]$ by a rational factor such that, in the resulting signal $x_{\rm d}[n]$, the nominal ENF be at $\omega_0 = \pi/10$ rad/sample. This implies setting the new sampling frequency in Hz to 20 times the nominal ENF.
2	Determination of a binary vector $v[n]$ that indicates the voice activity regions in $x_{\rm d}[n]$.
3	Isolation of the ENFC $x_{\rm enfc}[n]$ by bandpass filtering $x_{\rm d}[n]$, with a very narrow bandwidth centered at $\omega_0 = \pi/10$ rad/sample.
4	Estimation of the ENF as the instantaneous frequency $f[n]$ of $x_{\rm enfc}[n]$, via Hilbert's method [29].
5	Definition of a detection signal $d[n]$ as the absolute value of the median-compensated $f[n]$, i.e., a signal that represents the magnitude of ENF variations.
6	Computation of a variable magnitude threshold $t[n]$ for $d[n]$ that represents the maximum allowed limits for ENF variations in edit free conditions.
7	Evaluation of the following detection criterion: an edit is considered detected if there is at least one local maxima of $d[n]$ above $t[n]$ and within a voice-inactive region of $x_{\rm d}[n]$. Otherwise, the signal is considered unedited.

selected database signals available from [30]. Using these real-world noisy signals, performance evaluations of the proposed method, reported in Sections III and IV, demonstrate it is effective for edit detection. Robustness to noise is further investigated through the addition of artificially generated noise on top of the originally present background noise.

B. Overall Description of the Proposed Method

The main processing steps of the edit detection method are given in Table I. In comparison with the method presented in [23], the following differences can be highlighted:

- Instead of a high-order FIR bandpass filter, a nonlinear-phase elliptic filter has been used to isolate the ENFC, yielding significant reduction in computational cost;
- No block-based processing is used to compute the phase or frequency of the ENFC;
- The detection criterion is simpler, being the overall height of the variable magnitude threshold $t[n]$ controlled by a single parameter.
- Information on voice activity is incorporated into the detection criterion.

C. Detailed Description of the Processing Stages

As regards step 1, a conventional method for sampling rate reduction (decimation) by a rational factor P/Q has been used [31]. In general, this consists of an upsampler by an integer factor P in series with a lowpass filter followed by a downsampler by an integer factor Q. For instance, suppose that a SUT with an ENF $= 50$ Hz is originally sampled at 16 kHz. After decimation, since $\omega_0 f_{\rm s,d}/2\pi = 50$ Hz, with $\omega_0 = \pi/10$ rad/sample, one must set the new sample rate $f_{\rm s,d} = 20 \times 50\,{\rm Hz} = 1000\,{\rm Hz}$. Hence, the decimation in step 1 has a ratio of 1/16 and is carried out by applying first an anti-aliasing lowpass filter with unit gain and cutoff frequency

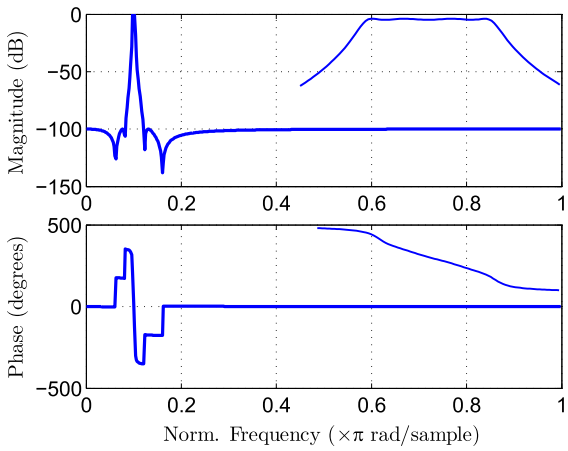

Fig. 1. Frequency response of the 4th-order IIR filter used in step 3. Details of the responses in the passband region are shown as inlays (out of scale) in the upper-right corners of the plots.

of 500 Hz and then downsampling the resulting signal by an integer factor of 16.

For step 2, instead of the algorithm described in [32], a simple energy-based Voice Activity Detector (VAD) has been employed. First, $|x_{\rm d}[n]|$ is median filtered with a window of about 8 ms. The resulting profile is lifted up by an offset equal to its own median value m. Then a running-maximum filter with length of 25 ms is applied to produce the final magnitude envelope $e[n]$. A proto-indicator of voice activity $v[n]$ is then obtained by logically evaluating, for all n, whether $e[n]$ is greater than $3m$. True values of $v[n]$ imply samples of $x_{\rm d}[n]$ in voice-active passages. Otherwise, samples of $x_{\rm d}[n]$ are in voice-inactive regions. Finally, $v[n]$ is post-processed to enforce predefined limits for the minimum activity duration (set to 60 ms) and minimum inactivity duration (set to 150 ms). First, voice-active passages shorter than 60 ms are discarded and incorporated to voice-inactive regions. Then, voice-inactive segments shorter than 150 ms are incorporated into adjacent voice-active passages.

The parameters of the proposed VAD have been found experimentally and turn out to be robust to the sampling frequency of the SUT. Thus, computational costs can be significantly reduced if the VAD is applied to the downsampled signal $x_{\rm d}[n]$ instead of the original $x[n]$. Implementation details can be looked at from the demonstration code provided in the companion web-page of this paper, see [30].

The bandpass filter in step 3 has been designed as a fourth-order elliptic filter. Taking as reference the sampling frequency of $x_{\rm d}[n]$ in Hz, the filter was specified with 2.8 Hz bandwidth centered at the nominal ENF, 0.5 dB maximum ripple in the passband and minimum 100 dB attenuation. Figure 1 displays the frequency response of the designed filter. Note the details of the passband region in which one sees that the phase is approximately linear. Given the high selectivity of the bandpass filter, phase linearity becomes irrelevant for the application tackled here. Therefore, it is justifiable to replace the high-order FIR filter used in [23] with an IIR filter that does the required job at a way lower computational cost. Moreover, zero-phase filtering is employed to compensate for

time delays between $x_{\mathrm{d}}[n]$ and $x_{\mathrm{enfc}}[n]$. For edit detection purposes, zero-phase filtering is unnecessary, but for edit location in time, it may help attaining more accurate estimates.

ENF estimation in step 4 is carried out with Hilbert's method [29] adapted to discrete-time signals. First, the analytic version of $x_{\mathrm{enfc}}[n]$ is obtained as $x^{(\mathrm{a})}_{\mathrm{enfc}}[n] = x_{\mathrm{enfc}}[n] + i\mathcal{H}\{x_{\mathrm{enfc}}[n]\}$, where $i = \sqrt{-1}$ and $\mathcal{H}$ is the Hilbert operator [29]. Since $x_{\mathrm{enfc}}[n] = \cos(\varphi[n])$, with instantaneous phase $\varphi[n]$, is an adequate model for the ENFC, $x^{(\mathrm{a})}_{\mathrm{enfc}}[n] = \exp(i\varphi[n])$. For slow-varying $\varphi[n]$, an approximation of the IF, in rad/sample, is $f[n] = \angle(x^{(\mathrm{a})}_{\mathrm{enfc}}[n]x^{*(\mathrm{a})}_{\mathrm{enfc}}[n-1])$. If necessary, the IF can be expressed in Hz by $f[n]f_{\mathrm{s}}/2\pi$.

Hilbert's analytic signal method for instantaneous frequency estimation has the advantage of being a well established sample-by-sample technique for which fast algorithms exist. Furthermore, it is known for its high sensitivity to noise. For the objective of this work, this feature represents at the same time a benefit and a nuisance. By having high sensitivity to noise, Hilbert's method will respond to subtle variations in the ENFC caused by edits made in the SUT, thus facilitating the detection task. Hence, for signals with low to moderate background noise, as the ones contained in the two tested databases, one expects that the proposed method will perform better than the DFT-based detector introduced in [23]. However, if the SUT is severely contaminated by background noise, unreliable estimates of the ENF are unavoidable, making edit detection as difficult as in the method of [23].

The performance evaluations in terms of Equal Error Rate (EER) reported in Section IV corroborate the above reasoning. More specifically, for the Carioca 1 database (see Section IV-A1), which contains real-life noisy signals, the superior performance of the current proposition (4% EER) compared with the method in [23] (7% EER) can be explained by the high-sensitivity of Hilbert's method for frequency estimation: it gives the proposed method the ability to detect challenging edit cases, such as cuts whose durations are very near an integer number of ENF cycles (see Section III-A).

The IF estimate can be refined by using higher-order operators for the phase derivative. Alternatively, lowpass filtering $f[n]$ has been experimentally found to reduce spurious oscillations that appear in the computation of $f[n]$ due to numerical approximations. For that purpose, again using the sampling frequency of $x_{\mathrm{d}}[n]$ in Hz as a reference, a fifth-order elliptic lowpass filter has been designed with passband of about 20 Hz and maximum 0.5 dB ripple as well as stopband with minimum attenuation of about 64 dB. As before, zero-phase filtering was employed. The attained $f[n]$ has initial and end transients of about 1 s caused by all filtering employed up to this stage. Both must be discarded to avoid false positives in the detection outcomes. Naturally, this rules out edit detection in the beginning and end of the SUT.

For illustration purposes, Figure 2 shows three ENF estimates computed from a noisy unedited signal: one using the first-order difference of the ENFC phase; other using a second-order difference of the ENFC phase; and the post-processed ENF via the low-pass filter specified above. As can be seen, especially from the detailed view in the inlay plot, the proposed lowpass filtering removes the ripples from the ENF estimate (first-order difference) but preserves the longer time scale variations of the ENF, which are the information of interest to the edit detection task.

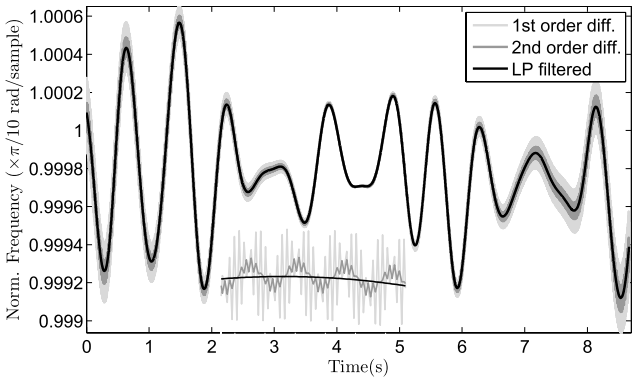

Fig. 2. ENF estimates obtained by Hilbert's method using first- and second-order phase difference as well as the former post-processed with the specified lowpass filter. Note the inlay plot with a detailed view of the estimates around time instant 8.1 s.

Following the rationale of anomalous ENF variations as indicators of edits in audio, in step 5 the detection signal $d[n]$ is simply defined as the modulus of the difference between $f[n]$ and its median value. To establish tolerance bounds for ordinary ENF variations, a method of background spectrum estimation called Two-Pass Split-Window (TPSW) [33] is resorted to. It is a simple nonlinear filtering that provides satisfactory estimates for the background spectral energy despite observed peaks in the spectrum. Here, sudden ENF variations take the role of spectral peaks, whereas normal and slower variations work as the background level.

The background level estimation could be also be carried out by median filtering or any other adequate tool available in the literature [34]. The TPSW-based estimator has been chosen because it is an effective yet simple background level estimator for peaky signals, being also highly efficient computationally (see [35, Sec. 1.4]). A comparison among the background estimates given by TPSW filter, moving-median filter, and moving-mean filter is shown later in Figure 5.

In brief terms, in the TPSW the input signal is applied to a filter whose finite impulse response is a rectangular window (moving-average) with a gap (zero-values) in the middle. The length of the gap should be as large as the typical duration of the spurious sudden signal variations, which will be called peaks here for simplicity. The influence of any peak present in the input signal on the filtered signal gets lateralized w.r.t the time location of the peak, since little effect on the output is observed when, during the convolution sum, the peak is inside the gap of the impulse response. As a result, the output of the first pass is an estimate of the long term tendency of the signal, except before and after peak locations, where the estimate is biased by the peaks. To compensate for that, these biased regions are replaced with the corresponding portions taken from the input signal. This can be automated by comparing the amplitudes of the input signal with those

of a magnitude scaled version of the first pass output. The resulting nonlinear modified signal is then passed to a regular moving-average filter to produce the final background estimate $b[n] = \text{TPSW}\{d[n]\}$. A demonstration movie of the TPSW can be downloaded from [30].

In this work, the duration of the split-window and moving-average filters of the TPSW has been set experimentally to 1 s, and that of the central gap to 125 ms. Furthermore, appropriate extensions via signal mirroring were made in the extremities of $d[n]$ to minimize filtering transients in $b[n]$.

The values of $b[n]$ can be viewed as a measure of the local half-excursion of $d[n]$ in unedited conditions. Given the purpose of the variable threshold $t[n]$ established in step 6 and the detection criterion defined in step 7, it seems justifiable to set $t[n] = b[n] + Gm_{\text{d}}$, where G is a scalar gain and $m_{\text{d}} = \text{median}\{b[n]\}$. In words, $t[n]$ is a version of $b[n]$ raised up by an offset of G times m_{d}, which represents the average half-excursion of $d[n]$. The value of G controls the overall height of $t[n]$ and will be the main parameter of the edit detection method. The use of $v[n]$ in the detection criterion gives the method robustness to signal defects that occur at voice-active regions, such as amplitude clipping, and that can trigger false detection. For distortion-free SUT, $v[n]$ could be ignored altogether in the method. Doing so would allow detecting edits in voice-active parts as well.

III. Qualitative Evaluation of the Proposed Edit Detection Method

To start with, analysis results of an unedited signal are confronted against those of edited versions of the same signal. The selected unedited signal, which belongs to the Carioca 1 database [23], is sampled at $f_{\text{s}} = 44.1$ kHz and contains an ENFC of 60 Hz. According to step 1 defined in Table I, it is initially downsampled to $f_{\text{s,d}} = 1200$ Hz and then subjected to steps 2–7, using the processing parameters mentioned in Section II-C. In addition, $G = 5$ has been set for all experiments shown in this section. For that setup, Figure 3 shows in panel (a) a portion of the unedited decimated SUT $x_{\text{d}}[n]$ and the corresponding active voice indicator $v[n]$. In panel (b) the isolated ENFC $x_{\text{enfc}}[n]$ is depicted. The 60 Hz oscillations are too fast to be seen, but from the envelope of $x_{\text{enfc}}[n]$, one verifies that the amplitude-modulated part of the ENFC is almost constant. In panel (c), one sees the detection signal $d[n]$, the variable threshold $t[n]$, and the voice-activity indicator $v[n]$ (scaled). As seen, for $G = 5$, the threshold $t[n]$ floats above $d[n]$ and no edit is detected. Note also the order of magnitude of the ENF variations, captured in $d[n]$.

A. Effects of Edit Duration

Now, the same analysis is carried out for the previous signal in three different edited versions: each one with a cut made at about 2.3 s but with durations respectively equal to $d_1 = 6.4/60$ s, $d_2 = 6.2/60$ s, and $d_3 = 6.02/60$ s. Here, the objective is to check the edit detectability w.r.t. the non-integer fraction of full 60 Hz cycles contained in the cut duration. Figure 4 displays the attained signals $d[n]$, $t[n]$, and $v[n]$ for

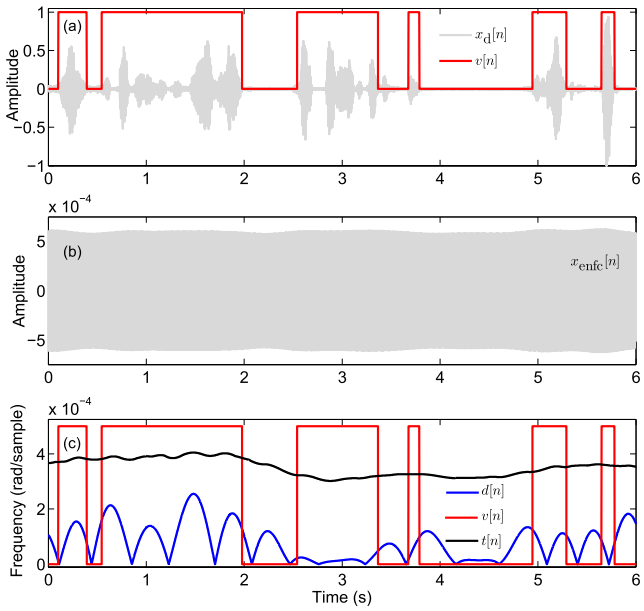

Fig. 3. Analysis of an unedited SUT. Panel (a): decimated signal $x_{\text{d}}[n]$ and corresponding active voice indicator $v[n]$. Panel (b): Isolated ENFC $x_{\text{enfc}}[n]$. Panel (c): detection signal $d[n]$, variable threshold $t[n]$, and $v[n]$.

each test condition. Note that, for the cut with duration d_1, the detection signal $d[n]$ shown in panel (a) is more peaky and way higher above the threshold $t[n]$ than the other $d[n]$ seen in panels (b) and (c). In the latter, since d_3 has 6 full 60 Hz cycles plus 2% of one 60 Hz cycle, $d[n]$ maxima is just a bit higher than $t[n]$. In general, the closer the fraction is to 50% of a 60 Hz cycle the easiest the detection. As the fraction approaches 0% or 100%, edit detection becomes harder. Of course, detection is impossible if a cut lasts an integer number of 60 Hz cycles. For signals with low to moderate levels of background level, experimental results have shown that edit detection is feasible for any cut whose duration difference from the nearest integer number of nominal ENF cycles is at most 2% of the nominal ENF cycle.

The thresholds $t[n]$ seen in Figure 4 are obtained via the TPSW-based background estimator described earlier in Section II-C. Alternatively, $t[n]$ could be computed by replacing the TPSW by a moving-median filter or a moving-mean filter. Just for illustration purposes, the cases (b) and (c) featured in Figure 4 are reanalyzed with these two types of filters, all having the same length as the TPSW filter. The results are displayed in Figure 5, where the type of filter used to compute the threshold $t[n]$ is indicated as a subscript.

As can be seen in Figure 5(a), for peaky $d[n]$, the threshold $t_{\text{mean}}[n]$ gets biased upward around the peak location. As expected the thresholds $t_{\text{TPSW}}[n]$ and $t_{\text{median}}[n]$ are more immune to the presence of peaks in $d[n]$ than $t_{\text{mean}}[n]$. Furthermore, the threshold $t_{\text{TPSW}}[n]$ has the advantage of being smoother and cheaper computationally than $t_{\text{median}}[n]$.

In Figure 5(b), where the edit-related peak in $d[n]$ is less prominent, the thresholds have similar behaviors. Yet, $t_{\text{TPSW}}[n]$ seems to be more stable and overall lower in level than $t_{\text{median}}[n]$ and $t_{\text{mean}}[n]$, for the same value of $G = 5$.

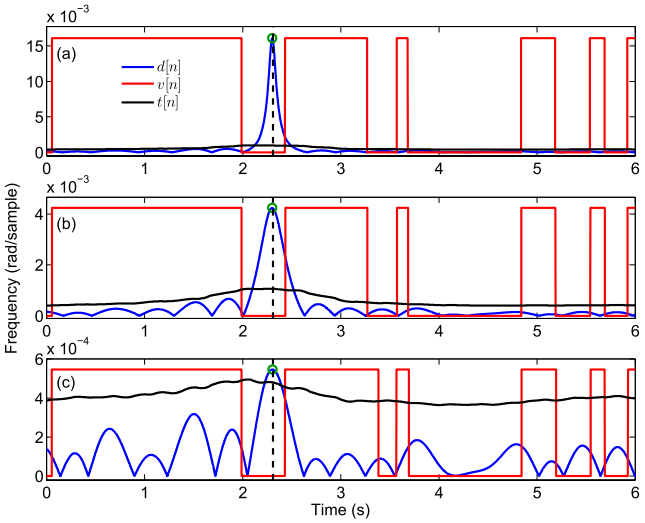

Fig. 4. Analysis of the SUT of Figure 3 with three different cuts. Detection signals $d[n]$, variable thresholds $t[n]$, and $v[n]$ for cuts with durations: (a) $d_1 = 6.4/60$ s; (b) $d_2 = 6.2/60$ s; and (c) $d_3 = 6.02/60$ s. The actual cut location is indicated with the vertical traced line. Note that the time instants of the maxima of $d[n]$, marked by green circles, are excellent estimates for the edit locations.

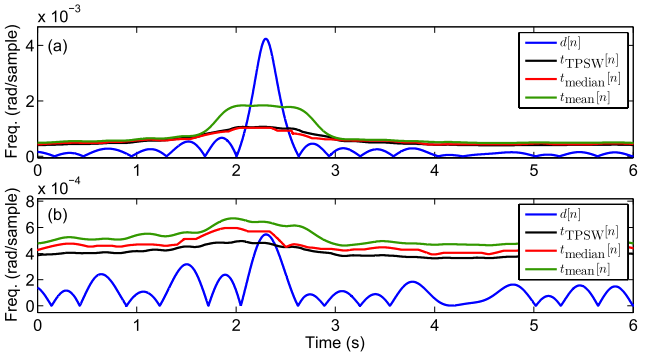

Fig. 5. Detection signals $d[n]$ against three different variable thresholds: $t_{\text{TPSW}}[n]$, $t_{\text{median}}[n]$, and $t_{\text{mean}}[n]$. All filters have the same length (1200 samples or 1 s). Panels (a) and (b) refer, respectively, to the same $d[n]$ of Figure 4(b) and (c).

In this example, only the $t_{\text{TPSW}}[n]$ allows correct edit detection. By reducing the value of G the other two threshold alternatives could also detect the edit-related peak in $d[n]$. However, this would increase the risks of false positive detections.

In order to gain insight on the behavior of $d[n]$ observed from Figure 4, the corresponding $x_{\text{enfc}}[n]$ are depicted in Figure 6. As can be seen in panel (a), for duration d_1, a drop in the amplitude envelope of $x_{\text{enfc}}[n]$ occurs around the cut location. This time localized fast variation in the amplitude modulated part of $x_{\text{enfc}}[n]$ will contribute a locally larger variation of the ENF $f[n]$ [29], explaining the observed peak in $d[n]$ around the same location. From the results seen in panels (b) and (c), it is clear that the extent of the level drop toward zero decreases as the cut duration approaches the nearest integer number of 60 Hz cycles. For differences as small as 2% of a 60 Hz cycle as in panel (c), the variation in the envelope of $x_{\text{enfc}}[n]$ is barely noticeable by visual means. Yet, the resulting subtle variation in $d[n]$ forms a peak that goes above the detection threshold $t[n]$.

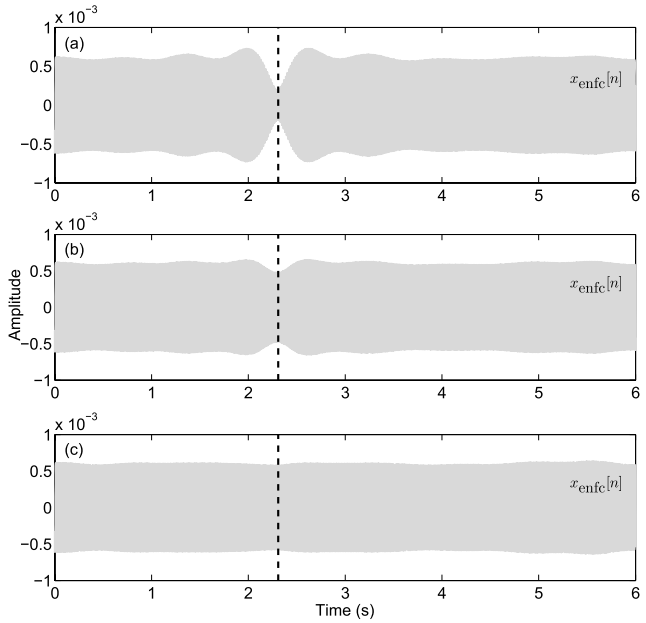

Fig. 6. Isolated ENFC $x_{\text{enfc}}[n]$ related to signals of Figure 4 with cut durations: (a) $d_1 = 6.4/60$ s; (b) $d_2 = 6.2/60$ s; and (c) $d_3 = 6.02/60$ s. The actual cut locations are indicated with vertical traced lines.

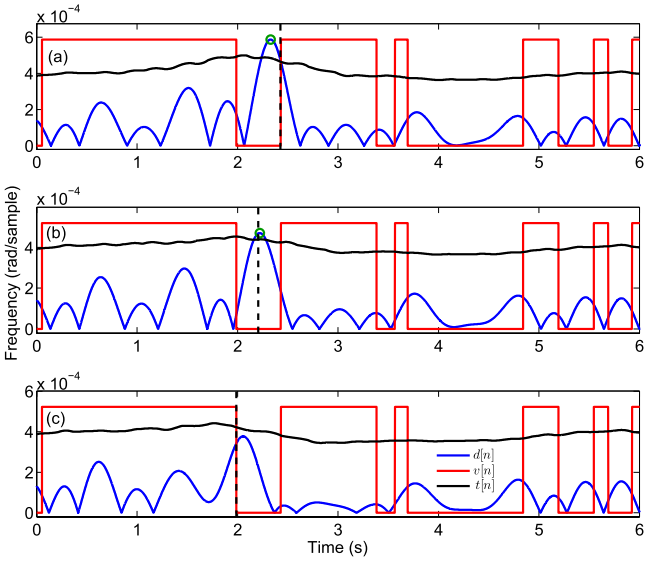

Fig. 7. Analysis of the SUT of Figure 3 with cuts of $d_3 = 6.02/60$ s at three locations: (a) at the beginning; (b) in the middle; and (c) at the end of the first voice-inactive region. Same conventions of Figure 4 apply here.

B. Effects of Edit Location

In this section, the novel edit detection method is evaluated w.r.t. the location of cuts made within a voice-inactive region of the SUT. For that, a challenging edit detection condition is chosen, i.e., the SUT will have cuts with duration d_3, which differs only 2% from the nearest integer number of full 60 Hz cycles. Three cut locations are set: in the middle of the voice-inactive region and at its extremities. The attained results, with same processing parameters as before, are organized in Figure 7. As can be observed from panels (a) and (c), for cuts occurring at the extremities of a voice-inactive

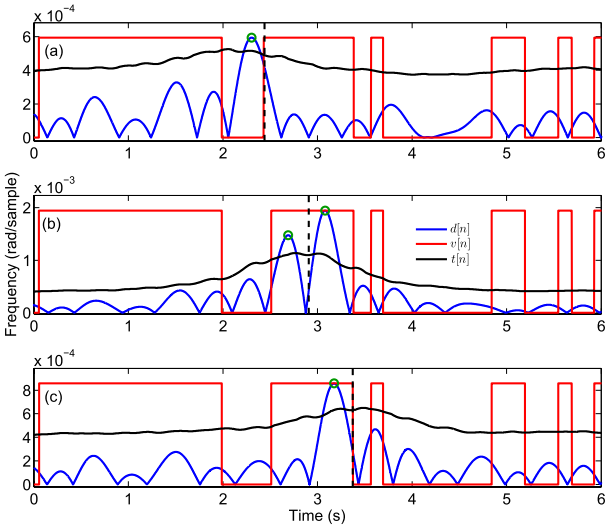

Fig. 8. Analysis of the SUT of Figure 3 with cuts of $d_3 = 6.02/60$ s at three locations: (a) at the beginning; (b) in the middle; and (c) at the end of the second voice-active region. Same conventions of Figure 4 apply here.

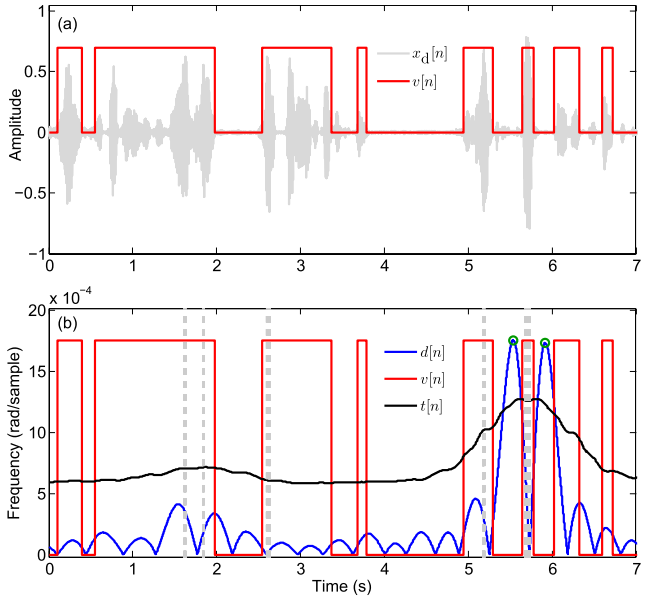

Fig. 9. Analysis of the SUT of Figure 3 with symmetrical amplitude clipping at level 0.7: (a) $x_{\mathrm{d}}[n]$ and $v[n]$; (b) $d[n]$, $t[n]$, and $v[n]$. The dashed vertical gray bands mark the time instants of the amplitude clipped samples.

region, their locations, as estimated by the abscissas of the nearest maxima of $d[n]$, tend to be biased. In these specific examples, the temporal biases happen to be inward the voice-inactive region. In panel (a) one sees that the cut is correctly detected, since there is a peak of $d[n]$ above $t[n]$, still within a voice-inactive part of the SUT. However, a false negative is observed in the example shown in panel (c), for there is no peak of $d[n]$ above $t[n]$. Note that false negatives would also happen if the temporal bias moved the maxima of $d[n]$ to voice-active parts of the SUT.

The observed temporal biases have nothing to do with the direction of the processing, since non-causal zero-delay filters are used. The explanation lies in the fact that a cut made between voice-active and inactive regions is bound to produce an abrupt amplitude jump in the speech signal. This generates a time-localized energy increase across the entire frequency spectrum, creating a perturbation in $d[n]$. Hence, two effects to the ENFC take place: an energy decrease due to the ENF phase discontinuity (see Figure 6) and a sudden local energy increase due to an amplitude jump in the speech signal. As it will be seen in Figure 8(b), the latter produces two peaks in $d[n]$ around the cut location. Therefore, the temporal biases seen in Figure 7(a) and (c) may be manifestations of these peaks on $d[n]$.

Just for illustration purposes, Figure 8 shows that the edit detection is possible when cuts occur within voice-active regions. For that, the restriction of local maxima of $d[n]$ above $t[n]$ to lie inside voice-inactive portions must be relaxed. Again, the same processing parameters of the previous examples have been used. As can be seen in panels (a)–(c), temporal biases in edit location estimates happen. In all cases the method catches the edits, since there are peaks of $d[n]$ above $t[n]$. In panel (b), one notices a substantial increase in the magnitude of the ENF variations around the true cut location. This is more likely to be caused by a high amplitude jump in the signal than by a phase discontinuity of the ENFC produced by the cut.

C. Effects of Amplitude Clipping

Amplitude clipping is a common nonlinear distortion that may afflict telephone call recordings. Assessing its influence on the performance of the proposed method is the object of this section. For that purpose, the SUT used in Figure 3 is normalized to maximum amplitude equal to unity. Then, symmetrical amplitude clipping to 0.7 is arbitrarily enforced.

Analysis of the above distorted SUT, with the same processing parameters as before, produces the results shown in Figure 9. In contrast with the result of Figure 3(c), there can be seen an overall increase in the extent of ENF variations, as captured by $d[n]$, caused by mild amplitude clipping (first four occurrences) in the SUT. Nevertheless, the variable threshold $t[n]$ also follows this tendency, for the ENF variations are slow.

However, for the fifth and stronger amplitude clipping that occurs at about 5.7 s, the ENF variations exceed the limit imposed by $t[n]$. One notices that the variation pattern in $d[n]$ differs from that induced by a cut in the SUT, since two lateral peaks appear adjacent to the clipped region. In the specific situation shown, this region is inside a short-time voice-active passage. Therefore, the peaks above $t[n]$ happen to be within voice-inactive regions and will mistakenly produce false positive detections. Were the clipping located well inside a long voice-active passage, the lateral peaks would probably lie inside it. If they were, a correct inhibition of a false positive would be guaranteed, according to the detection criteria (step 7) defined in Table I.

In summary, the proposed edit detection method seems to be robust to moderate amplitude clipping as well as to strong clipping, provided it occurs well inside a long voice-active region of the SUT. Otherwise, a false positive is an unavoidable detection outcome.

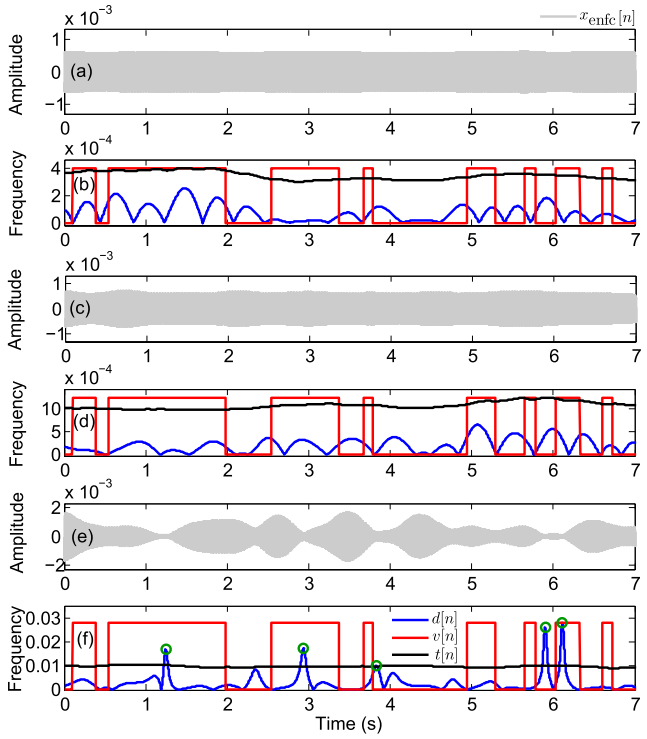

Fig. 10. Analysis of the SUT of Figure 10 corrupted by white Gaussian noise. Signals $\tilde{x}_{\text{enfc}}[n]$, $d[n]$, $t[n]$, and $v[n]$, respectively, for: original moderate background level – Panels (a) and (b); 20 dB local SNR – Panels (c) and (d); and 0 dB local SNR – Panels (e) and (f).

D. Effects of Broadband Background Noise

Another common source of distortion in audio is additive contamination with broadband background noise. If the SUT is corrupted with background noise of non negligible energy around the ENF, the bandpass filter (step 3 in Table I) used to isolate the ENFC will also preserve the noise energy within the filter's passband. As a consequence, the ENFC and its IF estimate are affected.

Given that the noise is additive, it is safe to assume that only the signal-to-noise ratio (SNR) observed in $x_{\text{enfc}}[n]$, i.e., within the passband of the bandpass filter, will matter to the performance of the edit detection method. This frequency-localized measure of the SNR will be called hereafter simply local SNR. In the experiments designed for this section, the SUT will be considered "noiseless" despite the original moderate background level it contains. Moreover, the extracted ENFC $x_{\text{enfc}}[n]$ will be treated as the signal of interest. First, a white Gaussian noise $e[n]$ is artificially generated at f_s sample-rate and then submitted to the same processing chain of the original SUT $x[n]$, up to the point of the bandpass filtering. After that, the power of this filtered noise $e_{\text{enfc}}[n]$ is adjusted to produce a desired local SNR. The artificially corrupted ENFC $\tilde{x}_{\text{enfc}}[n] = x_{\text{enfc}}[n] + e_{\text{enfc}}[n]$ is then submitted to the remaining steps of the processing chain.

Figure 10 shows the attained results for three different conditions: original background level and local SNRs of 20 dB and 0 dB, respectively. Here, it is more instructive to visualize $\tilde{x}_{\text{enfc}}[n]$ or its envelope and the related $d[n]$ and $t[n]$, for each of the defined conditions. As can be seen, in the original case, the envelope of $x_{\text{enfc}}[n]$ is practically constant over time, becoming increasingly more bumpy as the local SNR is reduced. As noise with more power is added to $x_{\text{enfc}}[n]$ the overall magnitude of ENF variations tends to increase: compare the maximum levels of $d[n]$ among panels (b), (d), and (f). For the 20 dB local SNR case seen in panels (c) and (d), the threshold $t[n]$ overtops $d[n]$, despite the overall magnitude increase of the latter when compared to the original case. However, for the 0 dB local SNR scenario, seen in panels (e) and (f), the more bumpy envelope of $\tilde{x}_{\text{enfc}}[n]$ affects more severely the ENF estimate and may produce sudden large variations in $d[n]$, especially when the envelope of $\tilde{x}_{\text{enfc}}[n]$ approaches zero. For example, observe the behavior of $\tilde{x}_{\text{enfc}}[n]$ and $d[n]$ around time instants 1.2s, 2.9s, 3.8s, 5.9s, and 6.1s, when peaks of $d[n]$ go above $t[n]$. Among these peaks, only those that occur in voice-inactive regions will trigger false positive detections.

E. Discussion

From the results seen in the previous sections, the following remarks can be made about the new edit detection method:

1) For signals edited in voice-inactive passages, false negative outcomes may happen if the edit removes from (or includes to) the SUT portions whose durations are very close to an integer number of ENFC cycles.
2) The occurrence of moderate amplitude clipping in the SUT may have little effect on detection performance.
3) Strong amplitude clipping in the SUT may result in false positives, if it is located in short-duration voice-active parts or not well inside those of long duration.
4) Although not previously exemplified, the negative effect of clipping in producing false positives is more prominent when the sampling frequency of the SUT is low. This is because the new high frequencies created by the clipping are more likely to exceed the Nyquist frequency ($f_s/2$) and be mirrored back to the low-frequency range to the point of interfering with the ENFC.
5) Both false positives and false negatives may result from contamination with background noise at low SNR around the ENF. False positives may happen as a consequence of sudden variations in $d[n]$ that go above $t[n]$. On the other hand, false negatives may occur because small variations in $d[n]$ due to an edit, detectable in a SUT with moderate background level, may be masked by the overall increase in the magnitude of $d[n]$ and $t[n]$ related to a more noisy $\tilde{x}_{\text{enfc}}[n]$.

The main objective of the experiments and results presented in Section III is to provide a forensic analyst with a clear understanding of the capabilities as well as the weaknesses of the proposed edit detection method. In knowing how and why more challenging real scenarios, with distorted SUTs, imply less reliable detection results, the analyst can go beyond the mere trusting of the detection outcome and have more useful elements to explore in order to make a more

informed judgment on whether a SUT has been edited or not.

IV. Quantitative Performance Evaluation

This section is devoted to report a more comprehensive evaluation of the proposed edit detection method. For that, a large number of unedited and edited signals taken from two distinct annotated speech databases will be used to assess the performance of the detection method. Furthermore, performance evaluations under more challenging conditions featuring SUTs corrupted with extra background noise or distorted by amplitude clipping will also be reported.

A. Description of the Databases

1) Carioca 1 Database: The database called Carioca 1, also used in [23], contains audio recordings of authorized PSTN phone calls. For that a simple analog circuit that uses a transformer was employed as an interface between the telephone line and the line-in input of an AC powered computer soundboard that recorded the audio from phone calls in PCM format.

All signals in the Carioca 1 database have been recorded at 44.1 kHz with 16 bits wordlength. Most of the energy related to the signals' speech content is within the range between 100 Hz and 4000 Hz. Thus, the 60 Hz ENFC that is present in the signals does not interfere with the speech information. The database contains in total 200 signals whose durations range from 19s to 35s. Out of the 100 unedited signals in the database, 50 are excerpts of male speech and 50 are excerpts of female speech. For each gender separately, from the 50 unedited recordings other 50 edited versions have been created: 25 with cuts and 25 with inserts. For each signal, just one cut or insert has been made. Moreover, the time instants (sample indices) and ranges of each edit have been carefully registered. The beginning and end of the cuts or inserts have been chosen to occur at voice-inactive parts of the signal. However, edit durations have not been checked or adjusted a priori in order to avoid biases toward more difficult or easier detection cases (see Section III-A).

Careful analysis of the signals in the Carioca 1 database reveals that, the individual SNRs measured at voice-active regions are distributed from 16 dB to 30 dB (22.3 dB on average). Hence, already in its original state the signals do have background noise at low to moderate levels. The strength of the ENFC does vary among the recordings, being about 20 dB lower for those of male speech in comparison with those of female speech. Several signals in the database suffer from intermittent noises, amplitude clipping, and spurious spikes. Despite that, no measures have been taken to fix the observed problems so that, by keeping the database intact, direct comparisons with the results of [23] could be made.

The individual SNR of each signal has been measured as

$$\mathrm{SNR}_{\mathrm{dB}}\{x[n]\} = 10\log_{10}\left\{\frac{\sum_n |x[n]\upsilon[n]|^2 - \sum_n |x[n]\bar{\upsilon}[n]|^2}{\sum_n |x[n]\bar{\upsilon}[n]|^2}\right\}, \quad (1)$$

where $\bar{\upsilon}[n] = 1 - \upsilon[n]$, with $\upsilon[n]$ being a binary indicator of voice activity, i.e., if $x[n_0]$ is in a voice active region, then $\upsilon[n_0] = 1$.

2) New Spanish Database: Contrarily to the Carioca 1 database, the Spanish database used in [23] was not readily available. To remedy that, a new Spanish database has been produced based on a subset of signals of the Gaudi Biometric Database – Gaudi-25 [36]. They comprise recordings of female subjects who enunciate sequences of numbers as well as prescribed or free sentences. These signals have not passed through a telephone channel, but were recorded using a microphone (Target CPT3GX) placed on a table in front of the subjects. Nevertheless, all of them contain a clear and stable 50 Hz ENFC. All signals are originally sampled at 16 kHz with 16-bit wordlength.

The Spanish database contains 100 unedited signals whose durations range from 16s to 42s. For each signal, a single cut has been made to produce an edited version. Given that the detection objective is to decide whether a signal is edited or not, regardless of the type of the edit, the absence of signals edited via inserts was not considered a shortcoming. On the contrary, inserts may facilitate the detection task, for two instances of ENFC phase discontinuities are possibly created. As with the Carioca 1 database, the cuts have been selected to begin and end at voice-inactive parts of the signal. Cut durations have not been adjusted in any way to facilitate edit detection (see Section III-A). Therefore, cut durations may be in excess of an integer number of 50 Hz ENF cycles by a fraction uniformly distributed within [0,1).

For the signals in the new Spanish database, the individual SNRs at voice-active regions range from 10.1 dB to 30 dB (20.2 dB on average). Hence, as with the Carioca 1 database signals, background noise is present at low to moderate levels. The strength of the ENFC also changes among the signals. Moreover, the database contains about 10 signals distorted by amplitude clipping, thumps, and impulsive noise, including dry-mouth noises. All these real-world events interfere with the ENFC and, thus, are likely to produce false positives.

B. Parameters of the Detection Method

For the quantitative performance evaluation using the Carioca 1 database, the processing parameters of the novel edit detection method are set exactly as described in Section II-C. Note that they have been chosen experimentally using a few test signals and well-informed decisions, but without any formal optimization procedure over a training database. The parameter G, which controls the overall height of the threshold $t[n]$, will be the only free parameter to be adjusted to achieve an EER detection, i.e., that with same number of false positives and negatives. An exhaustive search is made among values of G chosen from a coarse grid in a range wide enough. If necessary, more values of G in a finer grid around the EER can be tested.

For processing the signals of the Spanish database, only step 1 of the procedure described in Table I is adapted to account for a reduced sampling rate. The remaining steps and parameters are kept the same as before.

C. Experimental Setup

Performance evaluation of the proposed edit detection method, in terms of EER, will be carried out under three main conditions: Database signals as they were originally recorded; amplitude clipping forced upon the database signals; extra broadband background noise added to the database signals. It shall be stressed that, the database signals are far from being noiseless. They are recorded from typical office environment and thus contain background noise that has been estimated to produce SNRs no higher than 30 dB. The extra artificial degradations of the database signals are detailed below.

1) Signals With Amplitude Clipping: In order to distort the signals of both databases with amplitude clipping, a procedure similar to that used in [23] is implemented. A symmetrical amplitude clipping level is adjusted so as to reach prescribed percentages of clipped samples w.r.t the number of samples within voice-active regions of the signal. Here, the VAD described in Section II-C is employed and clipping is carried out at the original sampling rate of the signal. The set of prescribed percentages of clipped samples was {0.2, 0.5, 0.75, 1, 2, 3, 4}.

For each of the above percentages, all signals in the database are distorted before being submitted to the edit detection method. Then, several runs with different values of G are performed until the EER is achieved.

2) Signals Corrupted With Extra Background Noise: As in [23], signal corruption with extra broadband background noise has been carried out with three different types of additive noise processes: zero-mean white Gaussian as well as lowpass- and highpass-filtered versions of it. For that, respectively, the first-order filters with transfer functions $H_{\mathrm{LP}}(z) = 1/(1-0.9z^{-1})$ and $H_{\mathrm{HP}}(z) = 1/(1+0.9z^{-1})$ have been used.

Extra background noise has been added to each signal in the two databases to force a prescribed SNR, whenever this was lower than the primitive individual SNR. Reproducing the conditions used in [23], the SNR was measured w.r.t to the voice-active parts of each signal. The set of prescribed SNR values (in dB) was {30, 25, 20, 15, 10, 5}.

For each type of background noise and prescribed SNR, an ensemble of 10 noisy versions of the whole database has been generated. Then, for each realization of the noisy database, the edit detection was run with different values of G until achieving an average EER detection for the database ensemble.

D. Results for Original and Amplitude-Clipped Signals

The performance indicators in terms of EER for the original Carioca 1 database as well as its version with amplitude-clipped signals are organized in Table II.

As can be seen from Table II, for the Carioca 1 database in its original condition, the proposed edit detection method attained 4% EER, thus outperforming the method introduced in [23], for which the EER is 7%. Careful analysis of the metadata of the Carioca 1 database, related only to signals edited by cuts, reveals that in about 6 of them, the difference between the cut duration and an integer number of 60 Hz cycles is less than 5% of a 60 Hz cycle. Therefore, as seen in Section III-A, in these signals cut detection is hard and bound to yield false negatives. The above results suggest that the proposed method does a better job than that of [23] in detecting cuts with duration very close to an integer number of ENFC cycles.

TABLE II
PERFORMANCE OF THE EDIT DETECTION METHOD WITH THE CARIOCA 1 DATABASE DEGRADED WITH AMPLITUDE CLIPPING

Clipped Samples (in %)	Optimal G	EER (in %)
0.0	5.8	4.0
0.2	19.3	12.0
0.5	17.9	12.0
0.75	18.0	13.0
1.0	14.8	13.0
2.0	14.6	14.0
3.0	12.4	17.0
4.0	11.4	18.0

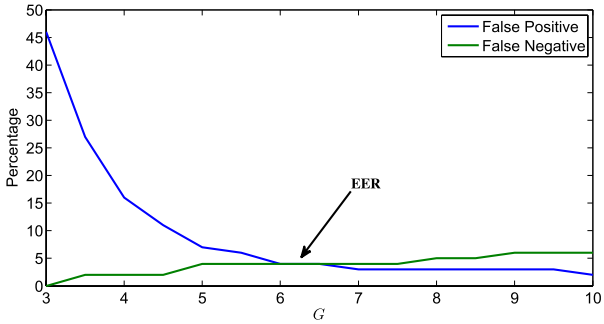

Fig. 11. Behavior of the percentages of false positives and false negatives as a function of G, measured for the Carioca 1 database. The EER region is indicated by an arrow.

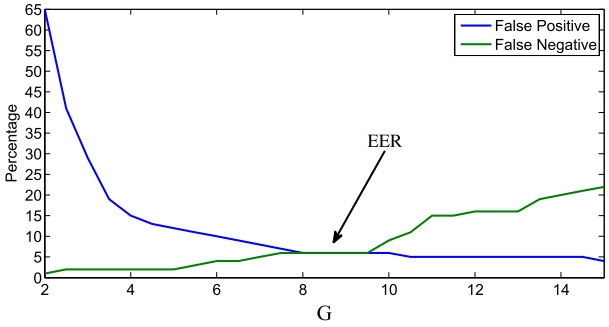

Fig. 12. Behavior of the percentages of false positives and false negatives as a function of G, measured for the new Spanish database. The EER region is indicated by an arrow.

For the Carioca 1 database signals, Figure 11 shows the behavior of the attained percentages of false positives and false negatives as a function of G, with values in the range between 3 and 10, in steps of 0.5. The EER is indicated with an arrow. As seen, the percentage of false positives tends to increase as G is reduced, whereas the percentage of false negatives tends to increase with G. Note that the EER lies in a narrow region instead of a single point. This suggests the best performance in terms of EER is robust w.r.t. the choice of G.

As regards performance as function of the percentage of clipped samples, the attained results with the proposed method follow the profile seen in [23, Fig. 12]. However, since these

TABLE III

PERFORMANCE OF THE EDIT DETECTION METHOD WITH THE NEW SPANISH DATABASE DEGRADED WITH AMPLITUDE CLIPPING

Signal Degradations in the Training Database		
Clipped Samples (in %)	Optimal G	EER (in %)
0.0	8.0	6
0.2	10.7	15
0.5	10.95	18
0.75	9.6	18
1.0	10.2	21
2.0	8.9	25
3.0	8.6	27
4.0	8.0	26

results refer to the original Spanish database, they cannot be directly compared with those of the proposed method. Despite of that, it is clear that the performance is impoverished by the presence of amplitude clipping in the signals.

The performance results attained with the new Spanish database are displayed in Table III. As can be seen, for the new Spanish database in its original condition, the performance of the proposed method happens to be equal to that reported in [23] for the original Spanish database. Since the databases are not the same, the results cannot be directly compared. The same holds true for the amplitude clipped versions, since different VAD algorithms have been used. Nevertheless, even for percentages of clipping as low as 0.5% the attained performance becomes too poor to the point of rendering impractical the use of the edit detection method.

For the new Spanish database signals, Figure 12 shows the behavior of the attained percentages of false positives and false negatives as a function of G, with values in the range between 2 and 15, in steps of 0.5. The EER is indicated with an arrow. A similar behavior to that seen in Figure 11 is observed here. Again, as the EER lies in a region, the best performance is robust w.r.t. the choice of G.

By comparing Tables II and III it becomes clear that the performance degradation due to amplitude clipping in the new Spanish database is more severe than that observed in the Carioca 1 database, for the same percentages of clipped samples in voice-active regions. A possible explanation for that lies in the lower sampling rate of the signals in the Spanish database. Since clipping produces localized discontinuities in the signal, high-frequency content will be generated, exceeding the Nyquist frequency. Owing to spectral aliasing, these spurious components may appear in the low end of the spectrum, thus inducing spurious variations to the ENF. The above problem is less critical when the sampling rate is higher, for less aliasing occurs.

E. Results for Signals With Extra Background Noise

Performance figures of the proposed edit detection method with the Carioca 1 database corrupted with broadband background noise are organized in Table IV. A consolidated view of these results is also given in Figure 13. It can be noticed that the mean EERs as functions of SNR follow the same profile reported in [23, Fig. 10], which refers to the original Spanish database. The observed behavior is in line

TABLE IV

PERFORMANCE OF THE EDIT DETECTION METHOD WITH THE CARIOCA 1 DATABASE DEGRADED WHITE GAUSSIAN NOISE (W) AND ITS LOWPASS (LP)- AND HIGHPASS (HP)-FILTERED VERSIONS

	Optimal G			Mean EER (%)		
SNR (dB)	W	LP	HP	W	LP	HP
30	5.1	5.1	5.8	4.0	5.7	4.1
25	4.3	15.7	4.9	5.4	20.4	3.4
20	4.6	21.1	4.4	12.3	36.7	4.0
15	8.8	28.6	3,8	24.7	47.0	5.9
10	18.1	31.6	3.9	38.7	48.7	10.1
5	24.1	31.2	6.2	45.0	49.4	17.1

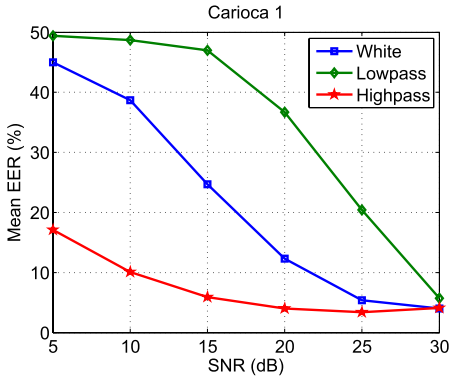

Fig. 13. Edit detection performance with Carioca 1 database corrupted with zero-mean white Gaussian noise as well as, lowpass- and highpass-filtered versions of it.

TABLE V

PERFORMANCE OF THE EDIT DETECTION METHOD WITH THE SPANISH DATABASE DEGRADED WHITE GAUSSIAN NOISE (W) AND ITS LOWPASS (LP)- AND HIGHPASS (HP)-FILTERED VERSIONS

	Optimal G			Mean EER (%)		
SNR (dB)	W	LP	HP	W	LP	HP
30	7.5	8.4	8.1	6	6.5	6
25	6.5	10	8.3	8.75	12.35	6
20	6	25.6	6.1	16.45	32.3	8.55
15	16	29.1	4.35	42.3	46.95	11.55
10	29	29.9	5.7	47.95	49.6	26.8
5	29.5	30.9	13	50.6	50.65	42

with the qualitative analysis made in Section III-D, in that what really matters is the local SNR in the narrow bandpass region around the ENFC. In fact, for a same overall SNR in voice-active passages, the local SNR due to the white Gaussian noise is higher than that due to lowpass-filtered noise but lower than that due to highpass-filtered noise.

The performance results of the proposed edit detection method with the new Spanish database degraded with broadband background noise are summarized in Table V and in Figure 14. Again, the same overall behavior seen in Figure 13 is observed, despite a poorer performance of the Spanish database in comparison with the Carioca 1 corrupted with white and highpass-filtered noise at low SNR. Note that the results seen in Figures 13 and 14 reflect the specific characteristics of the signals in the Carioca 1 and Spanish databases. For instance, the performance may change

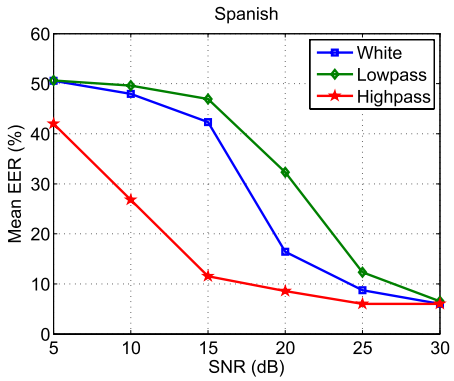

Fig. 14. Edit detection performance with Spanish database corrupted with zero-mean white Gaussian noise as well as, lowpass- and highpass-filtered versions of it.

depending on the strength of the ENFC in the signals. Here, artificial scenarios with additive noise have been set, whereas in real-world situations, background noise may be also attenuated at low-frequencies by the communication channel, thus easing the edit detection task.

From Figures 13 and 14, it is evident that the edit detection method has unsatisfactory performance for signals with SNR lower than 25 dB, especially when corrupted with white Gaussian and lowpass noise. This calls for further investigations in order to improve the performance of the method when dealing with noisy signals. A possibility worth checking is the use of parametric models to estimate the ENF, as proposed in [14].

F. Performance Results When Using the 3rd ENF Harmonic

In [25] the authors evaluate the method described in [23] for the Carioca 1 database when the third harmonic of the ENF is used instead of the original 60 Hz component. They report a 24% EER detection performance of their method.

To verify the performance of the proposed method under the same conditions, it has been adapted to work with the 180 Hz component of the ENF. For that, the passband filter used in Step 3 of Table I is redesigned as a fourth-order elliptic filter with center frequency at 180 Hz and a bandwidth of 2.8 Hz. All other processing parameters and steps are kept unchanged. Since there can be substantial variations of speech energy around the 180 Hz component, it will suffer more severe perturbations than the 60 Hz ENFC. Therefore, the novel edit detection method is expected to perform more poorly than for the 60 Hz ENFC. In fact, a 28% EER detection performance has been measured for the proposed edit detection method, a bit worser than the 24% EER reported in [25]. It is evident that none of the propositions can provide reliable detection results when using only the third harmonic of the ENFC.

V. Conclusions

A novel edit detection method for forensic audio analysis was proposed and described in this paper. It included modifications on a previous method introduced in [23], both in the ENF analysis and in the detection criterion, which employed a data-driven threshold-based strategy to catch out anomalous variations of the ENF related to edit events. After a qualitative evaluation of the influences of edit duration and location as well as noise contamination on the detection ability, a quantitative performance assessment of the method was carried out, in terms of the equal error rate (EER) detection. For that, two distinct annotated databases that contain both unedited and edited signals were used. The Carioca 1 was the same one used in [23] and a new Spanish database was produced.

For the Carioca 1 database, 4% EER was measured for the proposed method in contrast with 7% EER reported for the performance of the edit detector presented in [23]. As regards the Spanish database, 6% EER was achieved.

Tests using database signals distorted with amplitude clipping and additive broadband noise were also conducted. The attained performance results of the novel method are in line with those published in [23]. However, a direct comparison between the performance of the two methods is ruled out, since the noisy signals involved are not exactly the same. It seems evident from the results that amplitude clipping and additive broadband noise severely affect the performance of the proposed edit detection method. The same holds true when the third harmonic of the ENFC is used. In light of these limitations, further research is needed to improve detection performance in more challenging scenarios. Options worth investigating in the future are the use of parametric ENF analysis [14] and ENF estimators based on multitone models [15], [16] to mitigate the influence of corrupting noise on edit detection performance.

Acknowledgment

The authors would like to thank the Reviewers for the invaluable contributions given to improve the quality of this work.

References

[1] C. Grigoras, "Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis," *Forensic Sci. Int.*, vol. 167, nos. 2–3, pp. 136–145, Apr. 2007.

[2] E. B. Brixen, "Further investigation into the ENF criterion for forensic authentication," presented at the 123rd Conv. Audio Eng. Soc., New York, NY, USA, Oct. 2007.

[3] 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 Pract.*, Denver, CO, USA, Jun. 2008, pp. 1–10.

[4] R. W. Sanders, "Digital audio authenticity using the electric network frequency," in *Proc. AES 33rd Int. Conf., Audio Forensic, Theory Pract.*, Denver, CO, USA, Jun. 2008, pp 1–11.

[5] 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.

[6] R. C. Maher, "Audio forensic examination," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 84–94, Mar. 2009.

[7] F. Rumsey, "Electric network frequency analysis for forensic audio," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 852–855, Oct. 2012.

[8] C. Grigoras, J. M. Smith, and C. W. Jenkins, "Advances in ENF database configuration for forensic authentication of digital media," presented at the 131st Conv. Audio Eng. Soc., New York, NY, USA, Oct. 2011.

[9] A. Hajj-Ahmad, R. Garg, and M. Wu, "Instantaneous frequency estimation and localization for ENF signals," in *Proc. Signal Inf. Process. Assoc. Annu. Summit Conf.*, Hollywood, CA, USA, Dec. 2012, pp. 1–10.

[10] Y. Liu, J. Chai, B. Greene, R. Conners, and Y. Liu, "A study of the accuracy and precision of quadratic frequency interpolation for ENF estimation," in *Proc. AES 46th Int. Conf.*, Denver, CO, USA, Jun. 2012, pp. 1–5.

[11] O. Ojowu, J. Karlsson, J. Li, and Y. Liu, "ENF extraction from digital recordings using adaptive techniques and frequency tracking," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 4, pp. 1330–1338, Aug. 2012.
[12] Y. Liu, Z. Yuan, P. N. Markham, R. W. Conners, and Y. Liu, "Application of power system frequency for digital audio authentication," *IEEE Trans. Power Del.*, vol. 27, no. 4, pp. 1820–1828, Oct. 2012.
[13] Z. Yuan, Y. Liu, R. Conners, and Y. Liu, "Effects of oscillator errors on electric network frequency analysis," in *Proc. AES 46th Int. Conf.*, Denver, CO, USA, Jun. 2012, pp. 1–4.
[14] R. Garg, A. L. Varna, and M. Wu, "Modeling and analysis of electric network frequency signal for timestamp verification," in *Proc. IEEE Int. Workshop Inf. Forensics Security*, Tenerife, Spain, Dec. 2012, pp. 67–72.
[15] D. Bykhovsky and A. Cohen, "Electrical network frequency (ENF) maximum-likelihood estimation via a multitone harmonic model," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 5, pp. 744–753, May 2013.
[16] A. Hajj-Ahmad, R. Garg, and M. Wu, "Spectrum combining for ENF signal estimation," *IEEE Signal Process. Lett.*, vol. 20, no. 9, pp. 885–888, Sep. 2013.
[17] Z. Lv, Y. Hu, C.-T. Li, and B.-B. Liu, "Audio forensic authentication based on MOCC between ENF and reference signals," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, Beijing, China, Jul. 2013, pp. 427–431.
[18] H. Su, R. Garg, A. Hajj-Ahmad, and M. Wu, "ENF analysis on recaptured audio recordings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, USA, May 2013, pp. 3018–3022.
[19] M. Huijbregtse and Z. Geradts, "Using the ENF criterion for determining the time of recording of short digital audio recordings," in *Proc. 3rd Int. Workshop Comput. Forensics*, The Hague, The Netherlands, Aug. 2009, pp. 116–124.
[20] J. Chai, F. Liu, Z. Yuan, R. W. Conners, and Y. Liu, "Source of ENF in battery-powered digital recordings," presented at the 135th Conv. Audio Eng. Soc., New York, NY, USA, Oct. 2013, pp. 1–7.
[21] W.-H. Chuang, R. Garg, and M. Wu, "Anti-forensics and countermeasures of electrical network frequency analysis," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 12, pp. 2073–2088, Dec. 2013.
[22] D. P. Nicolalde and J. A. Apolinário, Jr., "Evaluating digital audio authenticity with spectral distances and ENF phase change," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 1417–1420.
[23] D. P. N. Rodríguez, J. A. Apolinário, Jr., and L. W. P. Biscainho, "Audio authenticity: Detecting ENF discontinuity with high precision phase analysis," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 3, pp. 534–543, Sep. 2010.
[24] S. Coetzee, "Phase and amplitude analysis of the ENF for digital audio authentication," in *Proc. AES 46th Int. Conf.*, Denver, CO, USA, Jun. 2012, pp. 1–5.
[25] D. P. Nicolalde-Rodriguez, J. A. Apolinário, Jr., and L. W. P. Biscainho, "Audio authenticity based on the discontinuity of ENF higher harmonics," in *Proc. 21st Eur. Signal Process. Conf.*, Marrakech, Morocco, Sep. 2013, pp. 1–5. [Online]. Available: http://www.eurasip.org/Proceedings/Eusipco/Eusipco2013/
[26] 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, 2009.
[27] R. Garg, A. L. Varna, A. Hajj-Ahmad, and M. Wu, "'Seeing' ENF: Power-signature-based timestamp for digital multimedia via optical sensing and signal processing," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 9, pp. 1417–1432, Sep. 2013.
[28] *Perceptual Evaluation of Speech Quality (PESQ): Objective Method for End-to-End Speech Quality Assessment of Narrow Band Telephone Networks and Speech Codecs*, document ITU-T P.862, Technical Recommendation, Geneva, Switzerland, 2005.
[29] L. Cohen, *Time Frequency Analysis: Theory and Applications*, 1st ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1994.
[30] P. A. A. Esquef. (Mar. 2014). *Companion Web-Page of the Paper*. [Online]. Available: http://lps.lncc.br/index.php/demonstracoes/tifs2014
[31] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2009.
[32] A. Benyassine, E. Shlomot, S. Huan-Yu, D. Massaloux, C. Lamblin, and J.-P. Petit, "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.
[33] W. A. Struzinski and E. D. Lowe, "A performance comparison of four noise background normalization schemes proposed for signal detection systems," *J. Acoust. Soc. Amer.*, vol. 76, no. 6, pp. 1738–1742, Dec. 1984.
[34] L. O. Nunes, P. A. A. Esquef, and L. W. P. Biscainho, "Evaluation of threshold-based algorithms for detection of spectral peaks in audio," in *Proc. 5th AES-Brazil Conf.*, São Paulo, Brazil, May 2007, pp. 66–73.
[35] P. A. A. Esquef, L. W. P. Biscainho, and V. Välimäki, "An efficient algorithm for the restoration of audio signals corrupted with low-frequency pulses," *J. Audio Eng. Soc.*, vol. 51, no. 6, pp. 502–517, Jun. 2003.
[36] Biometric Recognition Group—ATVS. *Gaudi Biometric Database*. [Online]. Available: http://atvs.ii.uam.es/databases.jsp, accessed Mar. 2013.

Paulo Antonio Andrade Esquef was born in Brazil in 1973. He received the Engineering degree from the Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil, in 1997, the M.Sc. degree from the Alberto Luiz Coimbra Institute of Post-Graduation and Research in Engineering (COPPE), UFRJ, in 1999, and the D.Sc. (Tech.) degree from Aalto University, Espoo, Finland, in 2004, all in electrical engineering.

In 2008, he joined the Signal Processing Laboratory at COPPE, UFRJ, as a Post-Doctoral Researcher.

Dr. Esquef has been an Associate Researcher with the National Laboratory for Scientific Computing, Rio de Janeiro, since 2009. His research interests are in audio DSP applications, such as sound analysis and synthesis, audio restoration, audio quality measurements, and sound source modeling.

José Antonio Apolinário, Jr. was born in Taubaté, Brazil, in 1960.

He graduated (signal corps officer) from the Military Academy of Agulhas Negras, Resende, Brazil, in 1981. He also 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, Brasília, Brazil, in 1993, and the D.Sc. degree from the Alberto Luiz Coimbra Institute of Post-Graduation and Research in Engineering, Federal University of Rio de Janeiro, Rio de Janeiro, in 1998, all in electrical engineering.

He is currently an Associate Professor with the Department of Electrical Engineering, IME, where he has already served as the Head of the Department of Electrical Engineering and the Vice Rector for Study and Research. He was a Visiting Professor with the Escuela Politécnica del Ejército, Quito, Ecuador, from 1999 to 2000, and a Visiting Researcher and twice a Visiting Professor with Aalto University, Espoo, 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). He has organized and been the first Chair of the Rio de Janeiro Chapter of the IEEE Communications Society. He was the Finance Chair of the IEEE International Symposium on Circuits and Systems in Rio de Janeiro, in 2011.

Luiz W. P. Biscainho (M'03) was born in Rio de Janeiro, Brazil, in 1962. He received the Electronics Engineering (*magna cum laude*) degree from the Department of Electronics Engineering, Polythechnic School, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, in 1985, and the M.Sc. and D.Sc. degrees from the Alberto Luiz Coimbra Institute of Post-Graduation and Research in Engineering (COPPE), UFRJ, in 1990 and 2000, respectively, all in electrical engineering.

Having worked in the telecommunication industry between 1985 and 1993, he is now an Associate Professor with the Department of Electronics and Computer Engineering, Polythechnic School, Electrical Engineering Program, COPPE. His research area is digital audio processing.

He is currently a member of the Audio Engineering Society, the Brazilian Telecommunications Society, and the Brazilian Computer Society.