Improved Edit Detection in Speech via ENF Patterns

Paulo A. A. Esquef
National Laboratory for Scientific Computing
Petrópolis, Brazil
Email: pesquef@lncc.br

José A. Apolinário Jr.
Military Institute of Engineering
Rio de Janeiro, Brazil
Email: apolin@ieee.org

Luiz W. P. Biscainho
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
Email: wagner@smt.ufrj.br

***Abstract*—In a recent paper published in the IEEE TIFS, we proposed an edit detection method based on the instantaneous variations of the Electrical Network Frequency (ENF). In this work we modify the detection criteria of that method by taking advantage of the typical pattern of ENF variations elicited by audio edits. We describe the implemented modifications and directly confront the performance of both methods using two distinct signal databases that contain real-life speech recordings. The experimental results demonstrate that the new proposition has an improved performance in terms of lower equal error rates when compared to its former version.**

I. Introduction

Electric Network Frequency (ENF) analysis finds use in many audio forensic tasks such as audio authentication, time-of-recording estimation, and audio edit detection [1]–[7]. In the latter, measures of spectral distance [8] and phase changes of the ENF over time [9]–[11] can be used as indicators of edits in a speech signal under test (SUT).

In a recent paper devoted to edit detection in speech signals [12], we implemented a series of modifications to the method introduced in [9]. In brief terms, we reported how edits (cuts and inserts) made in a SUT containing an ENF component (ENFC) show up as local anomalous increases in the usual variations of the instantaneous ENF. An edit detection criterion was then devised by confronting the measured magnitude of the instantaneous ENF variations against a signal-dependent threshold that set a limit for the typical variations of the instantaneous ENF of unedited signals. Variations of the instantaneous ENF exceeding the threshold were then deemed as evidences of edits in the signal.

Typical occurrences of false positives in the method of [12] were due to other sources of interferences to the ENFC, such as contamination by broadband noise and impulsive disturbances. A natural question was whether the pattern of the anomalous instantaneous ENF variations elicited by edits in the signal would differ from those evoked by other types of sources.

In this paper, we show that edits in speech signals do evoke specific patterns for the anomalous instantaneous ENF variations. Therefore, we propose a modification to the edit detection method previously introduced in [12] such that, in addition to the threshold-based detection strategy, a verification of the pattern of the anomalous instantaneous ENF variations is carried out. This way, we render the edit detector less prone to false positives and, consequently, improve the overall detection performance, in terms of the equal error rate (EER), i.e., when the percentages of false positives and negatives are equal.

The subsequent sections are organized as follows. In Section II, we review the original method of [12] and highlight the novel detection criterion that leads to an improved performance. In Section III-A, we briefly outline the main characteristics of the two signal databases used in the experiments. In Section III-B, we specify the conducted experiments. In Section IV, we report the attained results in a direct confrontation with the method of [12]. Concluding remarks are drawn in Section V .

II. Proposed Method

The proposed method is based on the edit detector reported in [12]. Therefore, it inherits the same assumptions on the types of signals it can tackle. In brief terms, any SUT should be available in digital PCM format and contain an ENF component that should be stable and the energy-dominant one around the nominal ENF. We also assume that edits (cuts or inserts) made in the signal are inaudible and with their initial and end points located at voice-inactive passages of the signal. For edits made in voice-active parts, detection methods based on spectral difference [8] or other resources [13], [14] can be more effective than our proposition.

The processing steps of the proposed edit detection method are listed in Table I. As can be seen, the signal processing steps (1 to 6) are identical to those of the method of [12]. Moreover, in order to make it possible to directly compare the performances of the original and the novel proposition, the processing parameters are kept unchanged. For a detailed description of the first 6 processing steps given in Table I the reader is referred to [12]. Just to render the paper self-contained, we briefly review here the key-points involving those processing steps. The main focus will be on the detailed description of steps 7 to 10, which form the core of the novel edit detection criterion. These steps involve not only the detection of anomalous intensities of the instantaneous ENF variations, but also a further verification of the pattern of these variations, by matching with specific templates.

Step 1 is carried out by a conventional method for sampling rate reduction by a rational factor [16]. The Voice Activity Detector (VAD) we employed in step 2 is a simple energy-based detector followed by heuristics to enforce predefined limits for the minimum voice-activity duration (set to 60 ms) and minimum voice-inactivity duration (set to 150 ms) [12]. A demonstration of the VAD is available from [17].

In step 3 we used zero-phase filtering with a fourth-order elliptic bandpass filter with 2.8 Hz bandwidth centered at the nominal ENF, 0.5 dB maximum ripple in the passband, and

978-1-4673-6802-5/15/$31.00 ©2015 IEEE

Table I. PROCESSING STEPS OF THE IMPROVED 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 [15].
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	Selection of all local maxima of $d[n]$ that are above $t[n]$ and within a voice-inactive region of $x_{\rm d}[n]$.
8	Computation of two prototypical detection signals $d_0[n]$ for specified cases of edits afflicting a synthetically generated ENFC, to be used as shape templates.
9	Computation of a set of cross correlation coefficients among $d[n]$ segments centered at each local maxima selected in step 7 and the two templates of step 8.
10	Evaluation of the following detection criterion: a SUT is considered edited if the maximum norm of the set of cross correlation coefficients is greater or equal to a pre-defined λ. Otherwise, the signal is considered unedited.

Steps 7 to 10 constitute the proposed modifications to the method of [12].

minimum 100 dB attenuation. The instantaneous ENF is estimated in step 4 via Hilbert's method [15] adapted to discrete-time signals. The advantages and drawbacks of using Hilbert's analytic signal method for instantaneous frequency estimation are discussed in [12]. In order to reduce numerical errors due to the first-order approximation of the phase derivative, we smooth out the instantaneous frequency estimate by zero-phase filtering it through a fifth-order elliptic lowpass filter with passband of about 20 Hz, maximum 0.5 dB ripple, and stopband with minimum attenuation of about 64 dB.

The main assumption behind the detector introduced in [12] was that edits in the SUT evoked anomalous ENF variations around the nominal value. Thus, in step 5 we formed the detection signal $d[n]$ by first subtracting the nominal ENF value (estimated as the median value of the instantaneous ENF $f[n]$) and then taking the magnitude of the result. Anomalous local increases in $d[n]$ are considered as evidence of an edit in the signal. Edit detection is then accomplished by checking whether $d[n]$ exceeds a variable threshold $t[n]$ that represents the acceptable extent of normal variations in $d[n]$ for unedited cases. For computing the threshold $t[n]$ we adapted a method of background spectrum estimation called Two-Pass Split-Window (TPSW) [18]. A demonstration of the TPSW is available from [17]. Here, as in [12], 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.

Denoting $b[n] = \text{TPSW}\{d[n]\}$, the threshold is computed as $t[n] = b[n] + Gm_{\rm d}$, where G is a scalar gain and $m_{\rm d} = \text{median}\{b[n]\}$. In [12](Figure 5) one can verify the effect of using the TPSW to compute $t[n]$ in comparison with a simple moving mean or median filtering. For unedited signals, the value of G is the parameter the user needs to adjust to make $t[n]$ float above $d[n]$ as a raised envelope. Anomalous variations in $d[n]$ are expected to go above $t[n]$ flagging an edit event. Moreover, the value of G will be one of the two parameters to adjust in order to attain the EER condition.

Now, we move to the novel part of the proposed edit detector. In step 7, we collect all local maxima (if any) of $d[n]$ that occur above $t[n]$ and within voice-inactive parts of the signal. The latter criterion is used because we assumed that edit points occur in these regions. Of course, if no local maximum meeting the above conditions is found, the signal is deemed unedited. Otherwise, we proceed by checking whether $d[n]$ around each collected maxima is similar to patterns evoked by edits in the signal, as described in the following.

For the computation of the templates mentioned in step 8, we simply generate a test signal $x[n]$ that solely contains an idealized sinusoidal ENFC. Its frequency and sampling rate should match those of signals in the tested databases. Then, one single cut with length prescribed as a fraction (0.1 to 0.5) of one ENF cycle is made in $x[n]$. We then run steps 1 to 5 as in Table I to obtain the corresponding detection signal $d_0[n]$. Finally, each of the resulting templates is defined as a portion of $d_0[n]$ around the global maximum, which is elicited by the edit.

By way of illustration, a set of templates $d_0[n]$ for cuts in $x[n]$ with lengths equal to 0.1, 0.2, 0.3, 0.4, and 0.5 times the duration of a 50 Hz cycle ($L_{\rm 50Hz}$) is seen in Figure 1. The latter template is shown separately in panel (b) since its magnitude is way larger than those in panel (a). The cuts have been made in the middle of the ideally generated ENF component. We emphasize that the effect on $d[n]$ due to the cut duration as a fraction (from 0 to 1) of $L_{\rm 50Hz}$ is symmetrical w.r.t. half $L_{\rm 50Hz}$. For instance, cuts of $(0.3)L_{\rm 50Hz}$ and of $(k-0.3)L_{\rm 50Hz}$, with $k \in \mathbb{N}$, produce the same template for $d[n]$. This explains our choice of fractions restricted to the range from 0.1 to 0.5 times $L_{\rm 50Hz}$, when setting the length of the cuts, for template generation.

From Figure 1 we notice that the magnitude and the width of the main lobe of the templates increase as the cut duration increases from 0.1 to 0.5 times $L_{\rm 50Hz}$. Moreover, we see that the cut of $(0.5)L_{\rm 50Hz}$ elicits a $d[n]$ with a large number of zeros around the global maximum, in contrast with the pattern observed in $d[n]$ for shorter cuts. From these observations we experimentally found that two templates $d_0[n]$ suffice to the task of pattern verification: one related to the cut-length $(0.5)L_{\rm 50Hz}$ and another to the cut-length $(0.4)L_{\rm 50Hz}$. The latter serves well to smaller fractions of $L_{\rm 50Hz}$. As regards the length of the two templates, we experimentally found that about 4000 samples at the sampling rate of $d[n]$ is an adequate choice.

In order to show that edits in $x[n]$ give rise to specific templates in $d[n]$, we create another artificial example in which, instead of a cut, a burst of 150 samples of zero-mean white Gaussian noise is added to $x[n]$ in the middle of the signal. The attained $d[n]$, both for the clean version and that with a noise burst are depicted in Figure 2. As can be seen, the effect of the noise burst on $d[n]$ is a double peak pattern, which differs from the single peak one elicited by an edit in $x[n]$. In [12](Figures 8 and 9) one sees other examples of double-peak patterns produced by edits implemented in voice-active parts of a SUT.

In step 9, first the global maximum of the template is aligned in time with one of the local maxima of $d[n]$ selected in step 7. Then, the computation of the cross-correlation coefficient is carried out in the region where the two signals

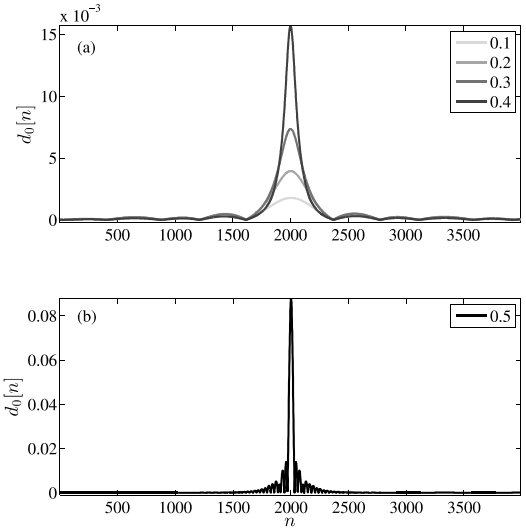

Figure 1. Panel (a): templates $d_0[n]$ for cuts in $x[n]$ with lengths equal to 0.1, 0.2, 0.3, and 0.4 times the duration of a 50 Hz cycle. Panel (b): templates $d_0[n]$ for a cut in $x[n]$ of 0.5 times the duration of a 50 Hz cycle.

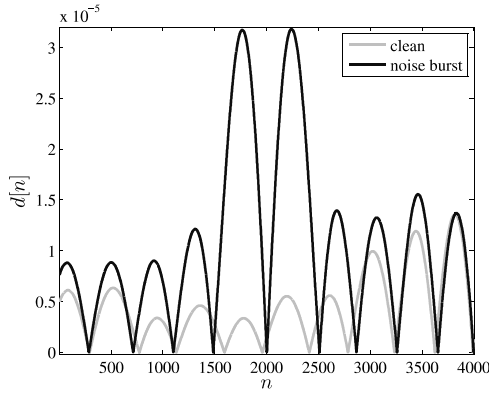

Figure 2. Signals $d[n]$ for a clean ENF component and its version corrupted with a central noise burst.

overlap in time. The procedure is repeated for every local maxima of $d[n]$ selected in step 7 and for the two templates designed in step 8. For the detection criterion defined in step 10, we found experimentally that the value of λ should lie around 0.7.

III. Quantitative Performance Evaluation

In order to compare under equal conditions the performance of the new edit detector with that described in [12], we tested the proposed method with the same experimental setup employed in [12]. More specifically, we adopted the same processing parameters, signal databases, and levels/types of additional distortions enforced to the database signals, as reviewed below.

A. Carioca 1 and Spanish Speech Databases

The Carioca 1 database features speech recordings of authorized PSTN phone calls in real-life office environments. All signals, which are sampled at 44.1 kHz with 16 bits wordlength, contain a 60 Hz nominal ENF component. Their durations range from 19s to 35s. The database contains a set of 100 unedited signals and another set of 100 signals, which are edited versions of those in the unedited set, with one cut or one insert per signal. In its original state the signals are noisy: the measured signal-to-noise ratio (SNR) at voice-active regions ranges from 16 dB to 30 dB (22.3 dB on average). The Carioca 1 database can be downloaded from [17].

The Spanish database contains recordings of female speakers via a microphone placed on a table in front of the subjects. All signals, which are sampled at 16 kHz with 16 bits wordlength, contain a 50 Hz nominal ENF component. The database contains a set of 100 unedited signals and another set of 100 signals, which are edited versions (with one cut per signal) of those in the unedited set. The signals are originally noisy with measured SNRs at voice-active regions from 10.1 dB to 30 dB (20.2 dB on average).

For all edited signals in both databases, the edit points have been chosen to start and end in voice-inactive parts of the signal. Edit durations have not been adjusted in any way to facilitate edit detection. The fractions of one ENF cycle are approximately uniformly distributed in the interval [0,1).

B. Experimental Setup

We aim at comparing the method reported in [12], which we use as a baseline here, with our improved edit detector. In order to facilitate referencing to the methods, henceforth we will call the edit detector of [12] by ED_b and the proposition with a novel detection criterion by ED_n. The subscripts b and n are mnemonics for baseline and novel.

For both ED_b and ED_n, the processing parameters used in steps 1 to 6 are the same (see Section II). For a more detailed description the reader is referred to [12](Section II.C). Our implementation code of the ED_n is available from [19].

In the evaluation of ED_b for optimal EER performance, the only adjustable parameter was the gain G that controls the height of $t[n]$ in step 6 (see Section II). For ED_n we do the same but also allow $\lambda \in \{0.7, 0.75\}$ in step 10. Experimentally, we found that setting λ around 0.7 yields quasi-optimal EER performance.

As in [12], we evaluate here ED_n for EER performance using Carioca 1 and Spanish databases, both in their original state as well as with further signal contamination by amplitude clipping and broadband background noise. Amplitude clipping is enforced at the original sampling rate, prior to signal analysis. The chosen percentages of clipped samples are $\{0.2, 0.5, 0.75, 1, 2, 3, 4\}$ w.r.t. the voice-active parts of each input signal.

We used three different types of additive processes to impose extra signal corruption by broadband background noise: zero-mean white (W) Gaussian noise as well as lowpass(LP)- and highpass(HP)-filtered versions of it. The LP and HP noise processes were produced by filtering W through first-order filters with transfer functions, respectively, $H_{\text{LP}}(z) = 1/(1-0.9z^{-1})$ and $H_{\text{HP}}(z) = 1/(1+0.9z^{-1})$. The prescribed SNR values (in dB) were $\{30, 25, 20, 15, 10, 5\}$ w.r.t. the voice-active parts of each signal. Note that only the signals with primitive SNR higher than a prescribed SNR have been modified. Furthermore, noise addition is enforced at the original sampling rate of the database signals.

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

Clipped Samples (in %)	Opt. G $\lambda = 0.7$	ED_n EER (in %)	ED_b
0.0	**3.3**	**2.0**	**4.0**
0.2	16.9	6.0	12.0
0.5	15.4	8.0	12.0
0.75	14.0	8.0	13.0
1.0	10.5	10.0	13.0
2.0	11.0	13.0	14.0
3.0	11.0	14.0	17.0
4.0	10.0	15.0	18.0

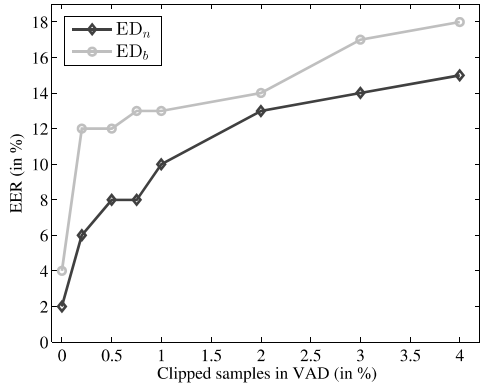

Figure 3. Comparative performance results between ED_b and ED_n in EER percentages, for the Carioca 1 database in its original and amplitude clipped versions.

Table III. PERFORMANCE OF THE EDIT DETECTION METHOD WITH THE SPANISH DATABASE DEGRADED WITH AMPLITUDE CLIPPING.

Clipped Samples (in %)	Opt. G $\lambda = 0.75$	ED_n EER (in %)	ED_b
0	**4.3** ($\lambda = 0.7$)	**1.0**	**6.0**
0.2	4.5	6.0	15.0
0.5	6.0	10.0	18.0
0.75	5.5	12.0	18.0
1.0	5.5	14.5	21.0
2.0	5.0	21.0	25.0
3.0	4.8	24.0	27.0
4.0	4.1	27.0	26.0

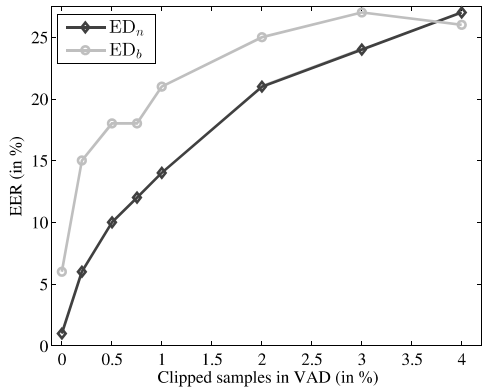

Figure 4. Comparative performance results between ED_b and ED_n in EER percentages, for the Spanish database in its original and amplitude clipped versions.

To ensure results with statistical significance, we generated an ensemble of 10 noisy versions of the two databases, for each type of background noise (W, LP, and HP) and prescribed SNR. Then, for each realization of the noisy databases, and for a given λ, we run ED_n for different values of G until achieving an average EER performance for the database ensemble.

As regards the EER performance, we computed the percentage of false positives as the number of signals the detector classifies as edited in the subset of unedited signals in the database, divided by the number of the latter. The same holds true for the percentage of false negatives, which is computed as the number of signals the detector classifies as unedited in the subset of edited signals in the database, divided by the number of the latter.

For a given value of λ, there is usually a range of values of G that ensure EER performance. The reported optimum value of G is the maximum value of the range.

IV. RESULTS

A. EER Performance for Amplitude Clipped Databases

We first show in Table II and Figure 3 the attained edit detection performance results for the original Carioca 1 database and its versions further degraded by amplitude clipping.

As can be seen for the Carioca 1 database, the EER percentages of the novel detector ED_n are consistently lower than ED_b. For the original database (0% clipping), we observe a reduction from 4% EER to 2% EER performance, respectively, from ED_b to ED_n. Reduction in EER with the ED_n is more prominent in the range from 0% to 1% clipping, where we see a decrease from 3 to 6 points in the percentage of the EER.

The detector error tradeoff, between false negatives (FP) and positives (FP), as functions of G and λ can be anticipated from the roles they play in the detector. Since G controls the overall height of $d[n]$, increasing G (for a fixed λ) tends to increase the rate of FN and decrease that of FP. Setting λ near 1 requires the observed pattern in $d[n]$ around a detected local maxima to be very similar to the templates to validade an edit detection. Doing so will tend to increase the rate of FN and decrease that of FP. These behaviors are confirmed by supplementary results available from the companion webpage [19].

Now we report in Table III and Figure 4 the detection performance results for the original Spanish database and its versions further degraded by amplitude clipping. We observe a similar behavior to that of the results related to the Carioca 1 database. For the original Spanish database (0% clipping) the best performance of ED_n is achieved with $\lambda = 0.70$ and we observe a reduction from 6% EER to 1% EER performance, respectively, from ED_b to ED_n. When dealing with amplitude clipped signals, except for the case of 4% clipping, ED_n (with $\lambda = 0.75$) yields consistently lower EER percentages than ED_b. For the Spanish database, reduction in EER with the ED_n is more prominent in the range from 0% to 2% clipping, where we see a decrease from 4 to 9 points in the EER percentages.

The results seen in Table II and III demonstrate that, for amplitude clipped signals, the novel detector ED_n consistently provides a more effective edit detection solution than the baseline ED_b. Since the database signals are already noisy in their original forms, the evaluations reported here take into account the combined effect of moderate broadband background noise and amplitude clipping.

Table IV. PERFORMANCE OF THE EDIT DETECTION METHOD WITH THE CARIOCA 1 DATABASE DEGRADED WITH WHITE GAUSSIAN NOISE.

White SNR (in dB)	Opt. G $\lambda = 0.7$	ED_n EER (in %)	ED_b
30	2.9	2.3	4.0
25	2.8	4.7	5.3
20	3.3	12.2	12.3
15	7.2	22.9	24.7
10	18.0	36.3	38.7
5	24.0	43.9	45.0

Table V. PERFORMANCE OF THE EDIT DETECTION METHOD WITH THE CARIOCA 1 DATABASE DEGRADED WITH LOWPASS GAUSSIAN NOISE.

Lowpass SNR (in dB)	Opt. G $\lambda = 0.7$	ED_n EER (in %)	ED_b
30	2.8	3.5	5.8
25	9.0	18.9	20.5
20	11.8	34.8	36.8
15	15.5	44.4	47.0
10	22.0	50.2	48.7
5	31.0	49.9	49.5

Table VI. PERFORMANCE OF THE EDIT DETECTION METHOD WITH THE CARIOCA 1 DATABASE DEGRADED WITH HIGHPASS GAUSSIAN NOISE.

Highpass SNR (in dB)	Opt. G $\lambda = 0.7$	ED_n EER (in %)	ED_b
30	3.0	3.1	4.1
25	3.0	2.1	3.4
20	2.9	3.2	4.0
15	2.7	6.5	5.9
10	2.6	10.0	10.1
5	24.9	15.5	17.2

B. EER Performance for Databases Corrupted with Extra Broadband Noise

We report comparative performance results between ED_n and ED_b, separately for each type of broadband noise, and for each database. Here, for practical reasons, we set $\lambda = 0.7$ for all evaluation tests. It is possible that, for specific cases, higher values of λ yield lower EER percentages than those reported below. Extra experimental results available from the companion webpage [19] reveal that λ has minor impact on the performance of ED_n. Typically, as λ is increased, the value of G that produces the EER condition tends to decrease.

In Tables IV, V, and VI we organize the attained edit detection performance results of ED_n and ED_b, respectively, for the versions of the Carioca 1 database further degraded by noise types W, LP, and HP (see Section III-B). As can be seen, ED_n tends to yield lower EER scores. To illustrate more clearly the differences between the performance figures of ED_n and ED_b, we show in Figure 5, the values of ΔEER defined as the EER percentages of the ED_n subtracted from those of ED_b. Hence, positive values of ΔEER indicate an improvement in performance, on the side of ED_n.

As can be seen from Figure 5, for the Carioca 1 database corrupted with noise type W, ED_n always outperforms ED_b. That is not always the case for signals degraded by noise types LP, and HP. It is clear that the ratio ΔEER/EER tends to decrease w.r.t the SNR. It is also clear that the levels of EER reduction when using ED_n are lower than those observed for the database signals degraded with amplitude clipping.

In the sequel, we report in Tables VII, VIII, and IX the attained edit detection performance results of ED_n and ED_b,

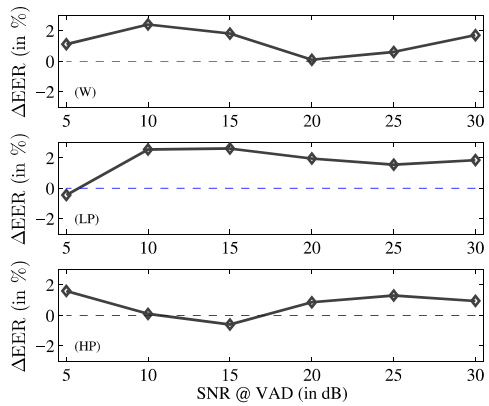

Figure 5. Comparative performance results between ED_b and ED_n in EER percentages, for the Carioca 1 database and its versions corrupted with broadband noise types W, LP, and HP.

Table VII. PERFORMANCE OF THE EDIT DETECTION METHOD WITH THE SPANISH DATABASE DEGRADED WITH WHITE GAUSSIAN NOISE.

White SNR (in dB)	Opt. G $\lambda = 0.7$	ED_n EER (in %)	ED_b
30	3.7	1.2	6.0
25	3.2	3.2	8.8
20	3.3	13.0	16.5
15	13.0	41.4	42.3
10	26.3	47.0	47.9
5	26.8	48.5	50.6

Table VIII. PERFORMANCE OF THE EDIT DETECTION METHOD WITH THE SPANISH DATABASE DEGRADED WITH LOWPASS GAUSSIAN NOISE.

Lowpass SNR (in dB)	Opt. G $\lambda = 0.7$	ED_n EER (in %)	ED_b
30	4.1	1.8	6.5
25	5.6	7.0	12.4
20	20.6	30.3	32.3
15	25.8	46.5	46.9
10	27.3	49.4	49.6
5	29.4	51.0	50.7

Table IX. PERFORMANCE OF THE EDIT DETECTION METHOD WITH THE SPANISH DATABASE DEGRADED WITH HIGHPASS GAUSSIAN NOISE.

Highpass SNR (in dB)	Opt. G $\lambda = 0.7$	ED_n EER (in %)	ED_b
30	3.4	1.0	6.0
25	3.8	1.3	6.0
20	2.4	4.3	8.6
15	2.5	10.5	11.6
10	4.3	22.5	26.8
5	9.6	41.2	42.0

respectively, for the versions of the Spanish database further degraded by noise types W, LP, and HP (see Section III-B). We also plot in Figure 6 the values of ΔEER as previously defined. We observe that, in general, the performance gains of the ED_n in relation to ED_b are higher for the Spanish database than for the Carioca 1 database. Except for the case with the Spanish database degraded by lowpass noise at 5dB SNR, ED_n outperforms ED_b. It is also clear that the most substantial performance gains are attained for signals with SNR above 20 dB. We stress that, despite the performance improvements, none of the detectors can be considered reliable for signals with SNRs equal or below 20 dB.

As regards edit detection using the third harmonic of the ENF [11], the performance of the ED_n for the Carioca 1

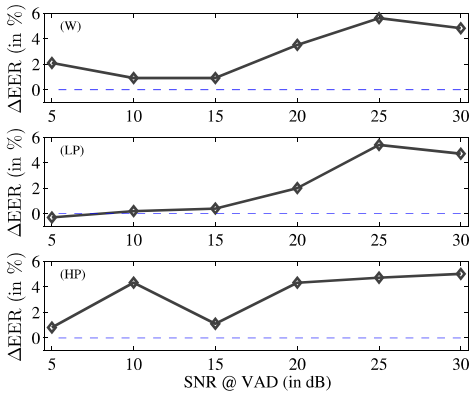

Figure 6. Comparative performance results between ED_b and ED_n in EER percentages, for the Spanish database and its versions corrupted with broadband noise types W, LP, and HP.

database was 22% EER ($G = 3.12$ and $\lambda = 0.85$). Thus, ED_n also shows some improvement in detection performance over the 24% EER reported in [11], and the 28% EER reported in [12]. Despite the improvement, a 22% EER performance cannot be considered reliable, in this case. Hence, edit detection via ENF higher harmonics alone still remains a challenge.

V. Conclusions

In this paper, we proposed a modification to the edit detection criterion of the ENF-based method of [12], in which anomalies in the instantaneous magnitude of ENF variations are taken as evidence of edits in a SUT. The introduced novelty here is an additional step to check whether a given detected anomaly follows specific patterns of ENF variations elicited by edits in a SUT. We devised a couple of templates that represent the effect of typical edits in a SUT on a detection signal. Then, for each detected anomaly in the instantaneous ENF we run a confirmation test to assert that it was indeed caused by an edit in the SUT. For that, the adopted criterion was that the cross-correlation coefficient between the ENF anomaly and one of the two templates was higher than a minimum level (typically, 0.7).

We confronted the proposed method (ED_n) directly with the baseline method of [12] (ED_b) and demonstrated experimentally that ED_n is more reliable since it tends to yield lower percentages of equal error rate (EER) detection than ED_b. For the Carioca 1 database, we observed a reduction from 4% EER (ED_b) to 2% EER (ED_n) detection. As for the Spanish database, we measured a reduction from 6% EER (ED_b) to 1% EER (ED_n) detection. For both databases degraded by amplitude clipping, ED_n almost always outperforms ED_b. The same holds true for the Spanish database corrupted with broadband noise. For the Carioca 1 database degraded with broadband noise, we see for ED_n a tendency of improved edit detection performance for SNR higher than 20 dB. We hope to have provided clear evidences to support the claim that, overall, the proposed edit detection method is superior to that introduced in [12].

Acknowledgments

The authors acknowledge and thank the financial support of CNPq-Brazil via Grants Nos. 475566/2012-2 and 304800/2013-9, as well as of CAPES Pró-defesa via Grant No. 23038.009094/2013-83.

References

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

[2] E. B. Brixen, "Further investigation into the ENF criterion for forensic authentication," Presented at the 123rd Convention of the Audio Engineering Society, Oct. 2007, preprint 7275.

[3] A. J. Cooper, "The electric network frequency (ENF) as an aid to authenticating forensic digital audio recordings – an automated approach," in *Proceedings of the AES 33rd International Conference: Audio Forensic, Theory and Practice*, Denver, USA, June 2008.

[4] R. W. Sanders, "Digital authenticity using the electric network frequency," in *Proceedings of the AES 33rd International Conference: Audio Forensic, Theory and Practice*, Denver, USA, June 2008.

[5] B. E. Koenig and D. S. Lacey, "Forensic authentication of digital audio recordings," *Journal of the Audio Engineering Society*, vol. 57, no. 9, pp. 662–695, Sept. 2009.

[6] R. C. Maher, "Audio forensic examination: Authenticity, enhancement, and interpretation," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 84–94, Mar. 2009.

[7] F. Rumsey, "Electric network frequency analysis for forensic audio," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 852–855, Oct. 2012.

[8] D. P. Nicolalde-Rodriguez and J. A. Apolinário Jr., "Evaluating digital audio authenticity with spectral distances and ENF phase change," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 1417–1420.

[9] D. P. Nicolalde-Rodríguez, J. A. Apolinário Jr., and L. W. P. Biscainho, "Audio authenticity: Detecting ENF discontinuity with high precision phase analysis," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 534–543, Sept. 2010.

[10] S. Coetzee, "Phase and amplitude analysis of the ENF for digital audio authentication," in *Proceedings of the AES 46th International Conference*, Denver, USA, June 2012, pp. 1–5.

[11] D. P. Nicolalde-Rodríguez, J. A. Apolinário Jr., and L. W. P. Biscainho, "Audio authenticity based on the discontinuity of ENF higher harmonics," in *in Proceedings of the 21st European Signal Processing Conference*, Marrakech, Maroc, Sept. 2013, pp. 1–5.

[12] P. A. A. Esquef, J. A. Apolinário Jr., and L. W. P. Biscainho, "Edit detection in speech recordings via instantaneous electric network frequency variations," *IEEE Transactions on Information Forensics & Security*, vol. 9, no. 12, pp. 2314–232, Dec. 2014.

[13] X. Pan, X. Zhang, and S. Lyu, "Detecting splicing in digital audios using local noise level estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. Kyoto, Japan: IEEE, March 2012, pp. 1841–1844.

[14] H. Zhao, Y. Chen, R. Wang, and H. Malik, "Audio source authentication and splicing detection using acoustic environmental signature," in *Proceedings of the 2nd ACM workshop on Information hiding and multimedia security*, ACM, Ed., New York, USA, 2014, pp. 159–164.

[15] L. Cohen, *Time Frequency Analysis: Theory and Applications*, 1st ed. Prentice Hall, 1994.

[16] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Prentice Hall, 2009.

[17] P. A. A. Esquef, "Companion web-page of the paper," lps.lncc.br/index.php/demonstracoes/tifs2014, Mar. 2014.

[18] W. A. Struzinski and E. D. Lowe, "A Performance Comparison of Four Noise Background Normalization Schemes Proposed for Signal Detection Systems," *Journal of the Acoustical Society of America*, vol. 76, no. 6, pp. 1738–1742, Dec. 1984.

[19] P. A. A. Esquef, "Companion web-page of the paper," http://lps.lncc.br/index.php/demonstracoes/wifs15, Mar. 2015.