Response to the AE and to the Reviewers

Manuscript ID: T-IFS-06339-2016

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

The authors would like to thank the Associate Editor and the Reviewers for volunteering their time in reviewing our manuscript and providing us with valuable comments which allowed us to significantly improve the quality and presentation of the paper. We have carefully revised the manuscript based on the reviewers' comments and suggestions. All changes in the paper have been marked with a red color. The following is our point-by-point responses to the raised comments.

I -REPLY TO REVIEWER #1

The paper is sound and the results important due to the improved performance (when compared to similar techniques) of the proposed scheme whenever the audio signal is corrupted by noise or non-linearity.

REPLY: Thanks a lot for your positive evaluation of our work.

Yet, I am not comfortable for recommending its publication as is. First of all, the text must be improved (too many details to list them all).

REPLY: Thank you for pointing this to us. We did our best to improve the text.

Secondly, it seems to me that the main possible use of the proposed method is regarding a class of recorded signals widely used nowadays, those audio signals recorded and already stored as MP3 files. Therefore, an experiment with edited MP3 files is recommended.

REPLY: We are grateful with your suggestion. We made experiments with edited and unedited MP3 audio files as suggested, and added the Subsection VI-C where we summarize the results.

I.1 - The manuscript needs a detailed review for typos, grammar issues and use of English (how to use "the" correctly, for instance)

REPLY: Thank you a lot for alerting us about it. We did our best to include all necessary corrections.

I.2 - SPHINS is not exactly an acronym but more like a nick-name. Authors could mention that the technique shall be designated herein as SPHINS (and not between parentheses implicating it corresponds to an acronym).

REPLY: We are thankful for your observation. We adopted the suggestion using SPHINS as a nick-name and no longer between parentheses.

I.3 - Section I (paragraph starting with "Despite the large"): The ENF is not deterministic, it may be slow varying but definitely not deterministic. Authors should avoid this concept associated to the ENF signal.

REPLY: Thank you for pointing out this to us. We excluded the improper word "deterministic" from the text, keeping only the mention to the well-behaved characteristics of the ENF signal, as follows:

> The high availability associated with the well behaved characteristics make it an attractive feature under the forensic point of view, which explains the wide use of it.

*(4) Equation (5) is not actually 100% accurate since x(n)*h_{bp}(n) is not exactly x(n) but a filtered version of it.*

REPLY: We appreciate your comment. In order to make clear that this is indeed an approximation, we have changed the equations as follows:

$$\hat{x}(n) \;=\; h_{\mathrm{bp}}(n) * s_{\mathrm{ut}}(n),$$

$$\hat{x}(n) \;=\; h_{\mathrm{bp}}(n) * x(n) + h_{\mathrm{bp}}(n) * [v(n) + e(n)],$$

$$h_{\mathrm{bp}}(n) * x(n) \;\approx\; x(n),$$

$$\hat{x}(n) \approx x(n) + z(n).$$

(5) Equation (7) is incorrect: a "j" is missing.

REPLY: We are grateful with your observation. The equation was corrected.

(6) Define "first derivative" of a sequence, say x(n): is it "x(n)-x(n-1)" or "(x(n)-x(n-1))/T"?

REPLY: Thank you for your comment. For clarity, we added in a footnote the definition used for the first derivative, as follows:

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

(7) Above Equation (14): "by the first and the last..." the second "the" is missing. "The rotational invariance property allow..." use "allows."

REPLY: Thank you for your correction. These errors are fixed. We made a detailed review for other typos and grammar issues.

(8) The explanation provided for Equation (14) is poor. It could be improved.

REPLY: We are thankful for your suggestion and we improve the explanation to this equation, showing the reason why it holds for the considered data model, as follows:

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

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

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

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

satisfying the rotational invariance property.

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

$$\mathbf{u_u}\phi = \mathbf{u_d},$$

where ϕ is a complex exponential whose argument corresponds to the desired angular frequency ω.

(9) Use "Block 1 in Fig. 1 ..." instead of "the block 1 in Fig. 1 ..."

REPLY: We adjust the text according to your valuable suggestion. Thank you very much.

(10) Subsection IV B: "As shown in 4, both estimators ..." Use "As shown in Fig. 4, both estimators ..." or "As shown in this figure, ..."

REPLY: We made the adjustments as suggested. Thank you very much for your hint.

(11) Last phrase on page 4: the word "Therefore" seemed to me not well placed in there; it lacks motivation to imply the logical link suggested by this word. Please review its use.

REPLY: We appreciate your hint. We reviewed the text in this sentence and made some changes, as follows:

> Considering these aspects, we propose a feature aimed to exploit jointly the Hilbert ENF estimates, which are more sensible to abrupt phase discontinuities, and the ESPRIT ENF estimates, which are more robust to noise.

(12) Some scholars say that "To classify automatically ..." is preferred to "To automatically classify ..." (avoiding the so-called split infinitive).

REPLY: Thank you for your suggestion. We adopted the suggested form, avoiding the split infinitive.

(13) Fig. 4: horizontal and vertical axes must include some information. Moreover, the legends are not defined.

REPLY: Thank you for pointing this out to us. We fixed the error in this figure.

(14) Page 8: "A 10-fold cross validation strategy is also made and, even using a more realistic cross-validation tests, SPHINS achieves similar results." Please improve this phrase, starting by checking subject-verb agreement.

REPLY: We are grateful for your observation. We reviewed the text based in your comment, as follows:

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

(15) Conclusion: "... the proposed method ... performs a 4% EER, ... " not "perform". Also in this section, please improve paragraph starting with "The method in [11]] ..." for clarity.

REPLY: Thank you for your suggestion. We corrected the use of the verb "to perform", and reviewed the text based in your comment, as follows:

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

(16) The last suggestion for future works, audio embedded with video H-sync pulses, do authors have a reference to include here?

REPLY: To the best of our knowledge, there are no references on this topic. However, the presence of narrowband signals from the video H-sync pulses embedded in audio files is the result of the forensic working with audio/video evidences at the Brazilian National Institute of Criminalistics. Such narrowband signals can be exploited for tampering detection in future works.

(17) Is corpus "Carioca 1" publicly available (for download)? If so, a mention to the site would be welcome for other researchers (to compare the performances of their works).

REPLY: The corpus is publicly available for download. We mention the website in which the download can be done. Thank you for your hint.

(18) Reference [2] seemed to me out of order of appearance (using BIBTEX would avoid this kind of problem).

REPLY: Thank you for pointing this out to us. We fixed it.

(19) References (and short bios) could definitely be improved (many details to be fixed).

REPLY: Thank you for your observation. We made a detailed review and fixed several details.

II -REPLY TO REVIEWER #2

The work in this paper falls under the umbrella of works proposing new approaches to carry out ENF-based forensic applications. In this case, the authors propose a new technique for using ENF embedded in audio to detect tampering. The authors compute ENF estimates using the ESPRIT method, and the Hilbert method, and then use the sample kurtosis of the estimated ENF along with an SNR estimate to form a feature vector that is fed to an SVM system to be trained to indicate whether or not tampering has taken place. The authors compare their approach to previous approaches to do ENF-based tampering detection (and use the same Carioca 1 database) and show that their approach results in higher performance. I think the work is interesting and adds value. Using the same dataset as the previous work does allow for easier comparison and I appreciate the level of detail with the results.

REPLY: Thanks a lot for your summary and positive evaluation of our work.

I do have comments/questions which I am copying below:

- I don't think the authors make it clear early on that the ENF actually does not stay at its nominal value and actually changes around it due to the supply/demand mismatch. I think the first explicit mention of it is in Section III. Possibly to accommodate readers unfamiliar with ENF, it might be better to make this more explicit and clearer earlier in the manuscript, before the discussion on approaches to estimate ENF or earlier ENF authentication work.

REPLY: Thank you for your suggestion. Aiming to improve the presentation and taking into account your comments, we modified the second paragraph of Section II where the ENF signal is first presented, as following:

> Although the power grid signal is ideally a real sinusoidal signal that oscilates with a nominal frequency, the actual ENF signal presents variations in its instantaneous oscilation frequency due to the mismatch between energy supply and demand on the power grid. Indeed, these signals are quite well behaved, showing no abrupt frequency and phase changes over time. The good behavior of these signals is a mandatory feature to the correct operation of many electric and electronic equipments. Therefore, despite the existent variations, strict control mechanisms are used to maintain the frequency of power grid signal within very narrow limits.

- As the authors have mentioned, there are several approaches to estimate the ENF signal from an audio recorder, yet the authors only chose to use the ESPRIT and Hilbert estimators to be part of the feature vector, and not others (e.g., Spectrogram-based, MUSIC, etc.) Have the

authors tried including estimates from the other approaches in the feature vector? Would it be beneficial to do so?

REPLY: We have tried to combine the kurtosis of estimates obtained from different approaches, like MUSIC and Spectrogram-based ENF estimates. When inserting this information in the feature vector as a new dimension, there is a slightly performance degradation, probably since there is high correlation between this information and the other dimensions, and due to the so-called "curse of dimensionality". The inclusion of new features in extra dimensions tends to worsen the performance of a classifier for a fixed amount of training samples, especially when these new features add a little amount of information.

- I am not entirely comfortable with the title of Section IV being "state of the art" (Better title maybe something along the lines of "ENF Estimation Approaches"?) There have been several proposed approaches for ENF estimation over the past few years and I'm not sure if one can confidently refer to ESPRIT and Hilbert as the "state of the art". Another point here is that not all methods require a band-pass filter, e.g. can do spectrogram first and later focus on certain frequency ranges, and more importantly many times the ENF traces do not appear at the nominal frequency and one would need to examine the harmonics of the nominal frequency for ENF traces.

REPLY: We agreed and changed the referred title as suggested: "ENF Estimation Approaches". Thank you very much for your suggestion.

- I am unclear as to what exactly the acronym SPHINS stands for. I find that it should be made clear somewhere in the manuscript.

REPLY: Actually SPHINS is not an acronym. We agree that it is more like a nickname. To make it more clear, we changed the paper avoiding to use SPHINS between parentheses, and reinforcing that the method is designated by the authors as SPHINS. We are thankful for your comments.

- A comment on the SNR estimate: Shouldn't the P_{ENF} be actually max[P_{ds}(k)] minus the estimated noise floor P_{noise} rather than just max[P_{ds}(k)]? On a related note, I don't think the authors motivated why they were computing the SNR in the first place. I think it would be

helpful if the authors mentioned the intuition behind including the SNR in the feature vector similarly to how they explained why they included kurtosis.

REPLY: Thank you for pointing out this to us.. Indeed, there were a few typos in these equations, with negligible effects on results. We have reviewed the equations carefully and made the proper corrections, as follows:

—

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

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

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

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

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

Additionally, we have included in the last paragraph of Subsection V-B an explanation to justify the inclusion of SNR_{ENF} in the feature vector, as follows:

> Using the kurtosis as an outlier measure for both HEE and 3E estimates, the proposed feature vector is defined as:
>
> $$\mathbf{F} = [\kappa(\hat{\omega}_{\text{H}_\text{b}})\ \kappa(\hat{\omega}_{\text{E}})\ \text{SNR}_{\text{ENF}}],$$
>
> where $\kappa(\hat{\omega}_{\text{H}_\text{b}})$ and $\kappa(\hat{\omega}_{\text{E}})$ are the kurtosis of ENF estimates, and $\mathbf{F}\in\mathbb{R}^3$ is the feature vector that summarizes anomalous ENF variations for an arbitrary audio recording. Since the HEE and 3E estimates have different robustness to noise, we included SNR_{ENF} as an element in feature vector. This allows the SVM classifier learning an appropriate decision surface in the training stage, jointly considering the SNR_{ENF} value with the kurtosis of the ENF estimates.

- What is the block size that the authors ultimately used for ESPRIT ENF estimation? i.e., what is the size of the signal for which they computed the instantaneous ENF value.

REPLY: Thanks for your question. The block size used to compute the 3E estimates is determined by the *N* value in Equation (13). In Section-VI we evaluate the influence of the block size for different *N* values. For all other tests we use N = 200 samples (167 ms, for a sampling rate of 1200 Hz). We have changed the following paragraph in Section-VI to reinforce the correspondence between *N* and the block size:

To evaluate the influence of the 3E block size in our algorithm, we assess the method performance among several block sizes by conducting cross-validation tests for sets of $N = \{100, 200, 300, 400, 500, 600, 700, 800\}$ samples

- I think the authors use the word 'overlook' erroneously. My impression is that they meant to say 'give an overview' instead, as 'overlook' means 'fail to notice', which I don't suppose is their intended meaning when using it.

REPLY: We replace the word “overlook” as suggested. Thank you for your comments.

- The paper could perhaps benefit from a table showing the parameters/variables used in the modeling.

REPLY: We are grateful for your hint. We introduce Table I describing the main variables and parameters in Section I.

III -REPLY TO REVIEWER #3

This presents an interesting approach to using ENF to detect tampering in an audio recording. Before publication, however, several things must be addressed.

REPLY: Thanks a lot for your positive evaluation of our work.

1. The amplitude of the power grid signal in equations (2) and (7) is time-varying, and so should read A(n) instead of just A.

REPLY: Thanks for your comment. We have adopted this in our modeling.

2. Equation (7) is missing a 'j' in the exp argument.

REPLY: We fixed it. Thank you for pointing out this to us.

3. In your definition of the right singular vectors below eqn (13), it should say "the right singular values of X", not 'U'

REPLY: Thank you for pointing out this to us. We corrected this typo.

4. In the first paragraph of Section V, it is stated "... ensuring a sufficient SNR in this spectral neighborhood." What is sufficient? Please provide a quantitative or qualitative explanation.

REPLY: Thank you for your observation. We reviewed the text in this paragraph, resulting in the following sentence:

> Similarly to the state-of-the-art schemes, the proposed method requires the validity of certain assumptions. First, the power grid interference is the most intense signal in the narrow vicinity of nominal ENF, ensuring a signal to noise ratio (SNR) in this spectral neighborhood that allows a reliable estimate of the ENF. With further degradation of the SNR, there is a deterioration in the quality of ENF estimate, resulting in a worse performance in tampering detection.

5. When downsampling an audio recording from 44.1 kHz to 1.2 kHz, there must be some sort of interpolation happening since 44100/1200 is not an integer. Please comment on how the downsampling occurred and if this affects the accuracy of your classifier.

REPLY: Thank you for your suggestion. In fact, we do some interpolation before the decimation procedure to achieve the desired sampling rate. Considering that the interpolation and decimation procedures are executed in a proper manner, avoiding the occurrence of aliasing, there is no reason to expect major influences in the classifier accuracy. We included a proper explanation in the article as suggested, resulting in the following text:

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

6. In section V.A, I'm not convinced that you're calculating the SNR of the ENF relative to the noise, but rather the SNR of the (ENF+noise) to the noise. In your formulation, it seems you're assuming that the noise isn't present in the ENF Bandwidth. Please clarify.

REPLY: Thank you for your comment. Indeed, there were some typos in these equations, with negligible effects on results. We have reviewed the equations carefully and made the proper corrections, as follows:

—

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

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

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

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

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

7. Please justify why eqn (25) is used as the kernel func.

REPLY: Thank you for your question. We choose a Gaussian kernel function due to its simplicity, flexibility, mathematical tractability, and numerical stability. More specifically, the Gaussian kernel has advantages for being an universal kernel, i. e., a kernel function able to find Hilbert spaces with universal approximating capability, usually giving reasonable results. In fact, the choice of a proper kernel function is not a trivial problem, and, as a rule of thumb, the Gaussian kernel is usually a good choice if used with suitable parameters. In practice, the kernel is tuned in proper cross-validation tests. We included a proper explanation in the article as suggested, resulting in the following text:

> The proposed SPHINS method uses a Gaussian radial basis function as the kernel function, defined as the following:
>
> $$K(\mathbf{F}_i, \mathbf{F}_j) = \exp\left(-\frac{||\mathbf{F}_i - \mathbf{F}_j||^2}{2\sigma^2}\right).$$
>
> We choose a Gaussian kernel function due to its simplicity, mathematical tractability, numerical stability, and due to its advantages for being an universal kernel, i. e., a kernel function able to find Hilbert spaces with universal approximating capability, usually giving reasonable results [23]. The parameter σ is an empirically determined optimal parameter that controls the bias-variance trade-off [22]. Typically the optimal σ is determined in cross validation tests.
>
> *(...)*
>
> [23] W. Liu, J. Príncipe, and S. Haykin, *Kernel Adaptive Filtering: A Comprehensive Introduction*. New York: Wiley, 2010.

8. By the time Section VI is underway, there are too many 3-letter acronyms to keep track of. A legend at the front of the paper would be helpful.

REPLY: We are grateful with your hint. We introduce Table II describing the acronyms in Section I.

9. Since the methods are only applied to a single corpus of data, the results, although promising, could be interpreted as being too anecdotal. The tone of the results section must be adjusted to reflect this. For example, the SPHINS outperforming [14] in Fig. 7 is nice, but your phrase "SPHINS achieves a better performance than [14] in noise degraded scenarios" is too strong, since this claim can only be made for this particular data set. Can this gain in performance be formally proven, i.e., mathematically generalized? Otherwise, for a claim as general as you have made, you'll need to see repeated results over many sets of data.

REPLY: We are thankful with your observations and suggestions here. We reviewed the text in the results section to reflect the correct tone of the results as suggested, and we reinforce that in the conclusion. Moreover, the need to perform additional tests on other corpora is described in conclusion as a future work.

> (...) when applied to Carioca 1 database, the proposed SPHINS method outperforms [14] in digital saturation scenarios.

> (...) when applied to Carioca 1 database, SPHINS achieves a better performance than [14] in noise degraded scenarios.

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