EUSIPCO 2013 1569728299

AUDIO AUTHENTICITY BASED ON THE DISCONTINUITY OF ENF HIGHER HARMONICS

Daniel P. Nicolalde-Rodríguez, José A. Apolinário Jr.[1]*, and Luiz W. P. Biscainho*[2]

[1]**Military Institute of Engineering (IME)**
Department of Electrical Engineering (SE/3)
Program of Defense Engineering (PGED)
Rio de Janeiro, Brazil
emails: danielnicolalde@hotmail.com and apolin@ime.eb.br

[2]**Federal University of Rio de Janeiro (UFRJ)**
Dept. of Electronics and Computer Engineering (DEL/Poli)
Program of Electrical Engineering (PEE/COPPE)
Rio de Janeiro, Brazil
email: wagner@lps.ufrj.br

ABSTRACT

This work deals with the use of electrical network frequency (ENF) higher harmonics to assess audio authenticity. It is assumed that ENF has corrupted the signal under analysis and also that, due to some non-linearity inherent to the sound recording process, higher harmonics are present. After down-sampling the audio signal and band-filtering it around a higher harmonic frequency, we estimate the phase of the resulting signal; the result provides a visual aid to check for possible audio tampering indicated by abrupt phase changes. As in a previous work based on the nominal ENF, we here use a feature extracted from the signal estimated phase to perform an automatic audio authenticity test. The result, although not as accurate as when the nominal ENF is used, provides a useful hint to audio forensic analyst whenever, for some reason, the nominal ENF has been removed from the signal.

1. INTRODUCTION

In recent years, an intensive use of the electrical network frequency (ENF) in forensic audio analysis has been observed [1]–[7]. Hence the popularity of this topic leads us to suspect that even a non-specialist can resort to anti-forensic techniques with the help of commercial audio software. For instance, someone can quite easily filter the nominal ENF out of the audio signal—and even fill the gap with noise—in an attempt to avoid audio authenticity analysis. However, when it happens, it is possible to still use information from the ENF present in its higher harmonics.

In practice, sound recording devices such as pre-amplifiers, analogue recorders, and A/D converters are not ideal and present some degree of non-linearity. A signal, in particular a quasi-single tone, passing through a non-linear transfer system produces additional harmonics, i.e., integer multiples of the fundamental frequency. This paper investigates the use of an ENF higher harmonic in audio authenticity analysis.

This paper is organized as follows. In the next section, we motivate our particular interest in the ENF third harmonic. In Section 3, we adapt a method based on phase discontinuity which tests for digital tampering [4, 5] to work with this harmonic. The resulting method is evaluated over a real-world *corpus* in Section 4. Conclusions are provided in Section 5.

2. CHOOSING THE HARMONIC

Let us assume that the (first harmonic of the) ENF originally embedded in a given audio recording has been filtered out on purpose in order to avoid forensic analysis [7]. Due to the nonlinearities tipically found in the recording devices themselves, we can expect to find (although weaker) higher harmonics of the ENF in the signal, which will obviously follow any frequency fluctuations of the fundamental.

Ideally, the best harmonic to use should present high energy and not be mixed with the speech signal. Firstly, it is reasonable to assume that recordings are edited in places with no voice activity in order to keep the tampering imperceptible. Furthermore, since the most common nonlinearities in recording systems can be modeled as odd functions, our search can be restricted to the odd harmonics. In order to aid this choice, we designed a simple experiment: a soft saturation (modeled by a hyperbolic tangent) was applied to a sinusoidal signal, which was further corrupted by additive white noise. The scheme and result are presented in Fig.1, where we can observe the prominent peaks located at odd harmonics, decreasing with frequency—thus suggesting the third harmonic is the best candidate for authenticity analysis whenever we do not find the ENF itself in the signal under study.

Corroborating this experiment, we have found the ENF third harmonic in most of the signals contained in the *corpus* used in [5], which was prepared without taking into account any nonlinear or ENF harmonics analysis. Since the presence of this harmonic is critical to the authenticity analysis, we describe in the following a procedure to detect this frequency.

3. THE PROPOSED METHOD

Our method, which was based on the relation between abrupt phase changes of ENF harmonics and edits in the audio signals, comprises three actions. The first one is to check for the

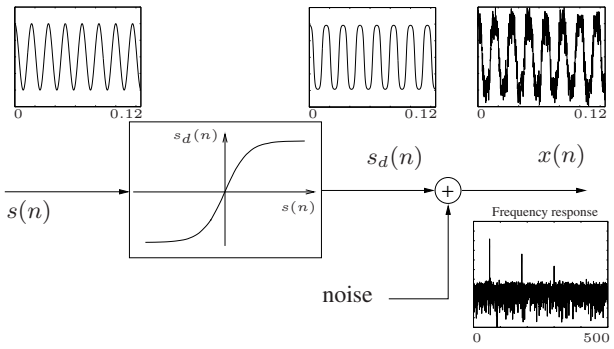

Fig. 1. A sinusoidal signal modified by a typical recording non-linearity and corrupted by additive noise (SNR = 10 dB). Note: in the time-domain plots, the horizontal axis is time, displayed in seconds; in the frequency response, the horizontal axis is frequency, displayed in Hz.

presence of the ENF third harmonic. The second one is to visualize its phase behavior. Finally, the last one is to determine, with the help of a decision metric, whether a signal can be taken as authentic or not.

3.1. ENF Third Harmonic Detection

A basic way to detect the presence of ENF third harmonic is verifying the presence of strong spectral components around 180 Hz (in this work, we assume that the ENF is 60 Hz). Specifically, we should devise some procedure to automatically verify if this component is embedded in the audio signal; a successful method is explained below.

(a) Down-sample the audio signal to a frequency of 1200 Hz in order to reduce the computational load.

(b) Apply a very sharp linear-phase FIR bandpass filter to the down-sampled signal. We have used in our experiments a 1.2 Hz bandwidth filter centered at 180 Hz with 10,000-coefficients (zero-phase filtering performed by the Matlab® function *filtfilt*). The filtered signal is denoted as s_{NB}.

(c) Apply a filter similar to that in item (b) but with a bandwidth of 10 Hz to the down-sampled signal. The filtered signal is denoted as s_{WB}.

(d) Compute the logarithmic ratio between the variance of the narrow-band filtered signal (1.2 Hz) and the variance of the wider band filtered signal (10 Hz):

$$R = 10 \log \left\{ \frac{\text{var}\left[s_{\text{NB}}(n)\right]}{\text{var}\left[s_{\text{WB}}(n)\right]} \right\}. \qquad (1)$$

When the ENF third harmonic **is present** in audio signals, R should result close to 0 dB ($\text{var}\left[s_{\text{NB}}(n)\right] \approx \text{var}\left[s_{\text{WB}}(n)\right]$, since that harmonic is a dominant peak). On the other hand, when the ENF third harmonic **is not present**, R tends to decrease ($\text{var}\left[s_{\text{NB}}(n)\right] < \text{var}\left[s_{\text{WB}}(n)\right]$).

(e) Decide whether the ENF third harmonic is embedded in the signal based on the decision rule:

$$R \underset{H_{\overline{3\text{H}}}}{\overset{H_{3\text{H}}}{\gtrless}} \gamma_{3\text{H}}, \qquad (2)$$

where $H_{3\text{H}}$ and $H_{\overline{3\text{H}}}$ represent the hypotheses that the ENF third harmonic is or is not embedded in the audio signal, respectively, and $\gamma_{3\text{H}}$ is the decision threshold. For values of R greater than $\gamma_{3\text{H}}$, we decide that the ENF third harmonic is embedded in the audio signal.

For determining the threshold $\gamma_{3\text{H}}$, we can resort to two sets of audio signals (corpora) recorded in a typical application scenario (a telephone conversation, for instance): one containing and the other one not containing the ENF third harmonic. Subsequently, the values of R for these signals can be computed and from their histograms we should be able to choose a proper $\gamma_{3\text{H}}$. As an additional suggestion, one could adjust two PDFs to fit the histograms and choose $\gamma_{3\text{H}}$ such that we have equal error (miss and false alarm) rates.

3.2. Visual Inspection

In order to help an end-user in the task of performing authenticity examination of an audio signal, a visual tool, also used in [5], is described in the following steps. Although a subjective task, the visual inspection should be always used for it takes into account the expertise of the forensic analyst.

(a) Down-sample the audio signal to a frequency of 3600 Hz.

(b) Apply to the down-sampled signal a filter similar to that in item (b) of Subsection 3.1, but with a bandwidth of 0.8 Hz (this bandwidth can be modified depending of the ENF tolerance of the local power company); we have used a filter with 30,000 coefficients in our experiments.

(c) Divide the filtered signal in blocks of ten periods of the third harmonic of the nominal ENF, $T = \frac{1}{180}$ s. Each block overlaps the former by nine periods.

(d) Estimate the phase of each segmented block obtained in item (c) using the DFT^1 method [5, 8]. In this case, we have used a 2,000-point DFT in the DFT^1 method. Let $\hat{\phi}(n_{\text{b}})$ be the phase estimate for the block index n_{b}.

(e) Plot $\hat{\phi}(n_{\text{b}})$ in degrees versus n_{b}. Each increment in n_{b} represents a duration of $T = \frac{1}{180} s$ in the time domain. Abrupt phase changes in the resulting plot can then be interpreted by the analyst as a possible signal edition (deletion or insertion).

3.3. Automatic Detection

Following the same principle used in [5], a feature is required to characterize abrupt phase changes in the ENF third harmonic. These abrupt changes are related to possible editions imposed to the audio signals. The automatic detection provides a fast first assessment which, in real cases, should always be supported by visual inspection.

The variation of the estimated phase every block n_b is the chosen criterion. This variation is represented by:

$$\hat{\phi}'(n_\mathrm{b}) \quad = \quad \hat{\phi}(n_\mathrm{b}) - \hat{\phi}(n_\mathrm{b}-1), \tag{3}$$

for $2 \leq n_\mathrm{b} \leq N_\mathrm{Block}$.

As in [5], a feature F is defined as:

$$F = 100 \log \left\{ \frac{1}{N_\mathrm{Block}-1} \sum_{n_\mathrm{b}=2}^{N_\mathrm{Block}} \left[\hat{\phi}'(n_\mathrm{b}) - m_{\hat{\phi}'} \right]^2 \right\}, \tag{4}$$

where $m_{\hat{\phi}'}$ is the mean value of $\hat{\phi}'(n_\mathrm{b})$.

The automatic decision is based on the rule:

$$F \underset{H_\mathrm{O}}{\overset{H_\mathrm{E}}{\gtrless}} \gamma, \tag{5}$$

where γ is the threshold for the final decision, H_O is the hypothesis that the signal has not been tampered, and H_E is the hypothesis that the signal has been edited. For values of F greater than γ, one decides that the audio signal has been edited.

Let $\{P_D, P_F, P_M\}$ be a probability group, where P_D represents the probability of detection (the audio signal is considered edited, when it has indeed been edited); P_F represents the probability of false alarm (the audio signal is considered edited, but it has not in fact been edited); and P_M represents the probability of missing (the audio signal is considered as not edited, but it really has been edited). Such probabilities can be expressed as:

$$\begin{aligned} P_\mathrm{D} &= P\left(\hat{H} = H_\mathrm{E} | H_\mathrm{E}\right) = P\left(F > \gamma | H_\mathrm{E}\right), \\ P_\mathrm{F} &= P\left(\hat{H} = H_\mathrm{E} | H_\mathrm{O}\right) = P\left(F > \gamma | H_\mathrm{O}\right), \text{ and} \\ P_\mathrm{M} &= P\left(\hat{H} = H_\mathrm{O} | H_\mathrm{E}\right) = P\left(F < \gamma | H_\mathrm{E}\right). \end{aligned}$$

In an attempt to attain an efficient detection, we can set γ to the value that ensures $P_M = P_F$. This point is known as EER (Equal Error Rate) in a Detection Error Tradeoff (DET) curve (P_M versus P_F for varying γ) [9]. In order to obtain the DET curve, it is necessary to prepare a database including both edited and unedited versions of a large set of audio signals, and evaluate them with the proposed method for an extended range of γ. More detailed information on this procedure can be found in [9].

4. EXPERIMENTAL RESULTS

The proposed method was evaluated using a set of signals obtained from two telephone calls, each one with an approximate duration of one hour, recorded in the city of Rio de Janeiro, Brazil. A total of 50 signals between 30- and 60-s length were extracted from each recording. We have edited half of these signals with a fragment deletion and the other half with a fragment insertion; the edition points were chosen during intervals of silence such that even an attentive listening of the edited signals would not reveal tampering. The final database is therefore composed by 100 unedited and 100 edited audio signals. All signals, although with embedded ENF, presented low background noise and no noticeable saturation—the recording process was carried out through a regular procedure, without any attempt to induce nonlinearities. Yet, higher ENF harmonics were present. This database shall hereafter be referred to as the CARIOCA *corpus*.

4.1. Detecting the ENF Third Harmonic

For the proper detection of the ENF third harmonic, it is necessary (as explained in Subsection 3.1) to obtain an adequate value for $\gamma_{3\mathrm{H}}$. In this work, we want to detect the presence of a strong 180 Hz component in the signals (assuming an ENF nominal value of 60 Hz). We have used the signals from the CARIOCA *corpus* as the reference signals with the 180-Hz harmonic embedded; as the reference signals without the 180-Hz harmonic embedded, we have used signals from two public databases (AHUMADA and GAUDI) from Spain (where the nominal ENF is 50 Hz) [10].

In Fig. 2, the histograms of R for signals both with and without 180 Hz are plotted. Additionally, it can be seen that a threshold for ENF third harmonic detection was set to $\gamma_{3\mathrm{H}} = -6.3890$.

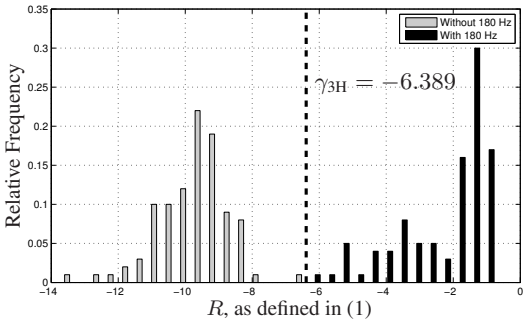

Fig. 2. Histogram of feature F for signals with and without the third ENF harmonic.

4.2. Visualizing the results

In this Subsection, two examples of the proposed visual method (Section 3.2) are presented.

We start with the example of an audio fragment deletion as presented in Fig. 3, where points P_1 and P_2 bound the deleted portion of the original signal. This figure presents the phase estimation of the ENF as well as the phase estimation of its third harmonic (for the original and the edited signals). In both cases, we can relate the abrupt phase change at the edited point P_1 with a possible signal edition.

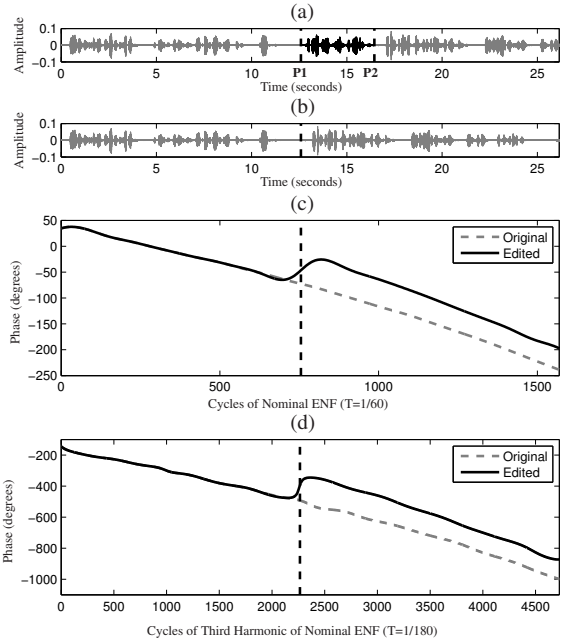

Fig. 3. Example of a fragment deletion. The limits of the deleted fragment are points P_1 and P_2. (a) Original Signal. (b) Edited Signal. (c) Phase estimation of ENF computed with the DFT[1] method with a window size of 10 cycles of nominal ENF ($T = \frac{1}{60}$) and $N_{\text{DFT}} = 2{,}000$ points. (d) Phase estimation of the ENF third harmonic computed with the DFT[1] method with a window size of 10 cycles of the ENF third harmonic ($T = \frac{1}{180}$) and $N_{\text{DFT}} = 2{,}000$ points.

Fig. 4 presents an example of a fragment insertion of length $P_4 - P_3$ in an audio signal where points P_3 and P_4 are the editing points. As in Fig. 3, the phase estimation of the ENF and the phase estimation of the ENF third harmonic (for the original and the edited signals) are shown. Again, abrupt phase changes are symptoms of signal edition.

4.3. Threshold for Automatic Detection and EER

Automatic detection, as explained in Subsection 3.3, aims at allowing the computer to decide whether the audio signal has been edited or not. This automatic decision depends on the

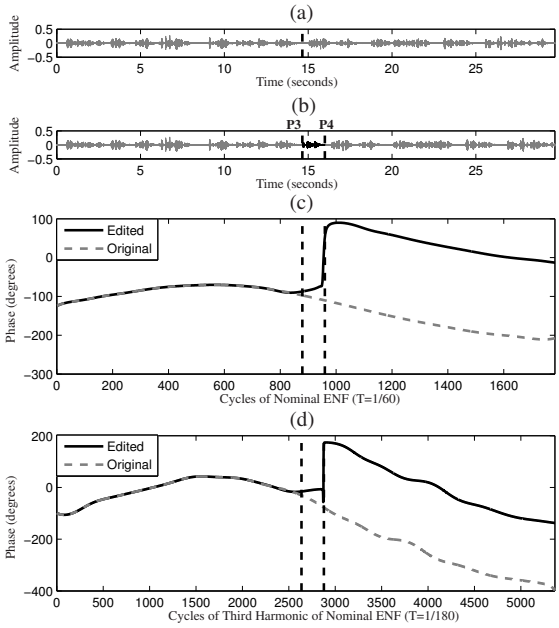

Fig. 4. Example of a fragment insertion. The limits of the inserted fragment are points P_3 and P_4. (a) Original Signal. (b) Edited Signal. (c) Phase estimation of ENF computed with the DFT[1] method with a window size of 10 cycles of nominal ENF ($T = \frac{1}{60}$) and $N_{\text{DFT}} = 2{,}000$ points. (d) Phase estimation of the ENF third harmonic computed with the DFT[1] method with a window size of 10 cycles of the ENF third harmonic ($T = \frac{1}{180}$) and $N_{\text{DFT}} = 2{,}000$ points.

value of F. We have obtained F for all signals of the CARIOCA *corpus* and the histograms of these values (for both original and edited signals) are depicted in Fig. 5.

For an automatic detection, we need to set an operating point (a value for γ in a DET curve). The DET curve for our experimental data is presented in Fig. 6. The threshold that results in $P_M = P_F$ ($\gamma = -182.96$, also indicated in Fig. 5) is shown in Fig. 6 and corresponds to the EER point (0.24, 0.24) in the DET curve. That means that that the detection probability is $P_D = 76\ \%$ with error probability (both miss and false alarm) equal to $24\ \%$.

Carrying out the same detection process with the ENF (first instead of third harmonic) for the same signals, an EER of 0.07 was obtained in [5]; this corresponds to a detection rate of $P_D = 93\ \%$ with error (both miss and false alarme) equal to $7\ \%$.

5. CONCLUSIONS

In this work, we have addressed the use of a superior harmonic of the ENF signal to evaluate audio authenticity by means of the occurrence of abrupt phase changes. We have seen that the third harmonic (180 Hz) was the most prominent

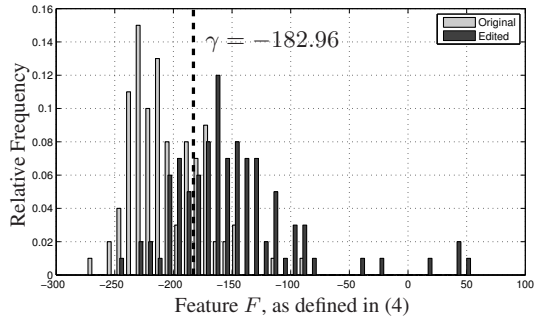

Fig. 5. Histogram of the values of feature F for original and edited signals.

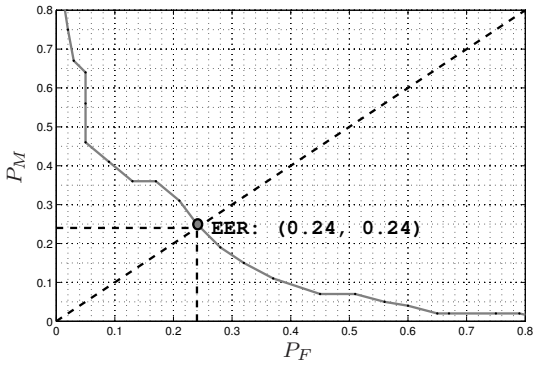

Fig. 6. The DET curve: $P_M \times P_F$.

candidate to support this analysis. We have presented a visual tool as well as an automatic detection method for the target task. From the experiments carried out with this technique, the equal error rate obtained for the automatic detection was $24\ \%$, as opposed to the EER of $7\ \%$ obtained directly with the ENF signal. A larger error rate could be expected due to the fact that the third harmonic is usually weaker than its fundamental, besides being closer to the frequency range occupied by speech. Nevertheless, the visual method may result in a useful tool for the audio authenticity examiner. The proposed method applies whenever, for any reason, the ENF (first harmonic) is not present; in case it is present, a combination of both signals (fundamental and third harmonic) could be devised to improve the detection rate. Moreover, alternative schemes employing odd harmonics beyond the third can be envisioned, provided they convey additional information on the potential tampering.

Acknowledgements

The authors express their gratitude to the Brazilian agencies CAPES and CNPq for partially funding their research.

6. REFERENCES

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

[2] 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* 33^{rd} *International Conference: Audio Forensic, Theory and Practice*, Denver, USA, June 2008.

[3] C. Grigoras, "Applications of ENF analysis in forensic authentication of digital audio and video recordings," *Journal of Audio Engineering Society*, vol. 57, no. 9, pp. 643–661, Sept. 2009.

[4] D. P. Nicolalde-Rodríguez 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 (ICASSP 2009)*, Taipei, Taiwan, Apr. 2009.

[5] 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.

[6] R. Garg, A. L. Varna, and M. Wu, ""Seeing" ENF: natural time stamp for digital video via optical sensing and signal processing," in *Proceedings of the 19th ACM International Conference on Multimedia*, Scottsdale, USA, Nov. 2011.

[7] W.-H. Chuang, R. Garg, and M. Wu, "How secure are power network signature based time stamps?," in *Proceedings of the ACM Conference on Computer and Communications Security*, Raleigh, USA, Oct. 2012.

[8] M. Desainte-Catherine, M. Desainte-Catherine, and S. Marchand, "High-precision fourier analysis of sounds using signal derivatives," *Journal of Audio Engineering Society*, vol. 48, no. 7/8, pp. 654–667, July/Aug. 2000.

[9] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of the European Conference on Speech Communication and Technology*, Rhodes, Greece, Sept. 1997.

[10] J. Ortega-García, J. González-Rodríguez, and V. Marrero-Aguiar, "Ahumada, a large speech *corpus* in spanish for speaker characterization and identification," *Elsevier Speech Communication*, vol. 31, pp. 255–264, June 2000.