Journal of Energy and Power Engineering 10 (2016) 504-512
doi: 10.17265/1934-8975/2016.08.007

A Parametric Autoregressive Model for the Extraction of Electric Network Frequency Fluctuations in Audio Forensic Authentication

Tarek E. Gemayel and Martin Bouchard

School of Electrical Engineering and Computer Science, University of Ottawa, 800 King Edward, Ottawa K1N 6N5, Canada

Received: May 25, 2016 / Accepted: June 7, 2016 / Published: August 31, 2016.

Abstract: This paper proposes a new method for extracting ENF (electric network frequency) fluctuations from digital audio recordings for the purpose of forensic authentication. It is shown that the extraction of ENF components from audio recordings is realizable by applying a parametric approach based on an AR (autoregressive) model. The proposed method is compared to the existing STFT (short-time Fourier transform) based ENF extraction method. Experimental results from recorded electrical grid signals and recorded audio signals show that the proposed approach can improve the time resolution in the extracted ENF fluctuations and improve the detection of tampering with short alterations in longer audio recordings.

Key words: Audio forensic authentication, electric network frequency fluctuations, autoregressive modeling, tampering and discontinuity detection.

1. Introduction

ENF (electric network frequency) signals are increasingly being used in the field of forensic digital audio authentication. Indeed, it is nowadays often possible to detect alterations made in audio recordings or to determine when and where a recording was made. In order to achieve this, a solution was proposed by taking into consideration the analysis of ENF signals embedded within an audio file and then comparing it to a reference database constructed by recording the ENF fluctuations of the electrical grid directly from a power outlet [1-3]. In Ref. [1] the extraction process of the ENF component was done either by implementing a STFT (short-time Fourier transform) method for long recordings (i.e., 1 hour or longer) or by using a zero-crossings method to tackle shorter recordings (i.e., 15 minutes or less), where the zero-crossings method measures the time differences between every two consecutive crossings in the time waveform of a band pass filtered signal around the ENF frequency. For the case of long audio recordings, the STFT method provides good performance. For shorter signals or short alterations in long signals, the STFT cannot provide good performance because of the use of long windows and the resulting poor time resolution (this will be illustrated in this paper). The zero-crossing method can provide good results for short audio recordings, but it was found to suffer from two shortcomings. Firstly, it has to operate at a high sampling frequency (e.g. 44.1 kHz or at least 8 kHz). By comparison, the STFT method employed in this paper works at a downsampled sampling rate of 0.2 Hz, which greatly reduces the computations and required memory. Secondly, the zero-crossing method was reported in Ref. [1] to be sensitive with results that can vary significantly depending on the analysis window size, window overlap, etc. The method is also non-linear in the sense that different results can be obtained when different sampling rates are used, unlike the STFT

Corresponding author: Martin Bouchard, Ph.D., professor, research field: signal processing and its applications.

methods which produce similar results at different sampling rates (as long as the Nyquist sampling criteria is met). The focus of this paper is to implement an AR (autoregressive) modeling method that works at the reduced sampling rate of 0.2 Hz. Under some conditions such AR methods are known to produce a "super-resolution", i.e., a better frequency resolution than SFTF methods when a small number of samples are available, or alternatively an equivalent frequency resolution using less samples than STFT methods (thus improving the time resolution).

This paper is organized as follows. Section 2 will present the basics of ENF fluctuation extraction and STFT methods. An outline of the proposed AR modeling method is presented in Section 3. Section 4 discusses the recording of the electrical grid signals for our experiments. Section 5 presents some extraction results with the STFT and AR methods, while Section 6 provides a conclusion.

2. Overview of ENF Fluctuations Extraction and STFT Extraction Methods

Generators rotating at the speed of 60 cycles per second in North America and 50 cycles per second in Europe produce an AC (alternative current) that travels along transmission lines. Ideally the frequency is fixed at a constant value of 60 Hz or 50 Hz, depending on the region. However, since electricity production is contingent on power demand, it must be generated accordingly. Higher demand will cause the frequency to drop while lower demand will cause the frequency to rise momentarily. As an example, the Canadian territory is divided into three major transmission lines: the Western interconnection, the Eastern interconnection and the Quebec interconnection [4, 5]. The fluctuations introduced in the grid are random, non-predictable and are most importantly uniform throughout the entire transmission line. For our work, an ENF database was built by capturing the electrical signal directly from a power outlet via a probe based on the design in Ref. [6], and connecting the probe to a computer and a sound board for recording purposes. Audio recordings were made using a battery-powered Olympus WS210S audio device. Since audio recording devices also capture some ENF interference from nearby appliances via electromagnetic fields, obviously the closer the recording device was to the emitting appliance the higher the ENF component was in the audio file. But this was not found to be an important factor in our system, since the level of ENF signal was always found to be sufficient to perform ENF extraction. The capturing of the reference electrical grid signal was done in a computer room at the University of Ottawa and the audio files were recorded in an apartment room located at a distance of about 2 km from the University campus. In Ref. [1] the STFT method and the FFT (fast Fourier transform) method were presented as two distinct methods but they can also be seen as equivalent, since they only differ in the presentation of the results (i.e. spectrogram for STFT vs. spectrum magnitude for the FFT). Both of them are based on a discrete Fourier transform and windowing of the signal, and are computed with FFTs in practice. In this paper we define the STFT as simply being the magnitude of the FFT taken on windowed portions of the signal, with a window of 200 seconds length moved in order to cover the whole signal as shown in Fig. 1. The number of windows required to scan the entire signal depends on the signal length N, the window size L and the window shift M applied between each consecutive window: $(N-L+M)/M$.

In our implementation of the STFT ENF extraction approach, we first downsample the recording's original

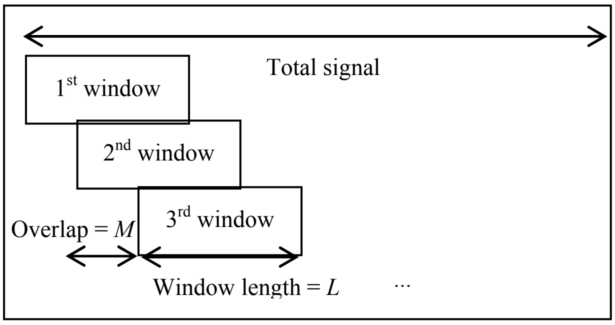


Fig. 1 STFT method with sliding window.

rate (i.e. usually ranging from 8 kHz to 44.1 kHz) to approximately 140 Hz, which is a bit higher than twice the 60 Hz component (to respect the Nyquist sampling criteria). Furthermore, a narrowband zoom or downsampling to 0.2 Hz around the ENF frequency of interest is performed, resulting in a FFT applied on the downsampled signal with a window of 40 samples (i.e. equivalent to 200 seconds), yielding a computational resolution of 5 mHz on a range of 0.2 Hz (normally from 59.9 Hz to 60.1 Hz although this may vary depending on the sampling clock frequency bias of the recording device). Fig. 2 shows an example of the resulting spectrum obtained from a STFT, resulting in 40 frequency points over a 0.2 Hz range. A weighted average can be computed over the 40 frequency points to extract the ENF sample, or alternatively the frequency of the maximum value can be taken as the ENF sample. A single ENF sample is thus produced for each analysis window. The process described was implemented for each consecutive window subject to a 5 seconds interval shift. This shift value could be increased up to 100 seconds or even 200 seconds to reduce the computational complexity and the data size, but using a smaller shift of 5 seconds was convenient because it also corresponds to the window shift used in the AR method to be described later in Section 3, allowing a better comparison and compatibility between the extracted ENF signals from the two different methods.

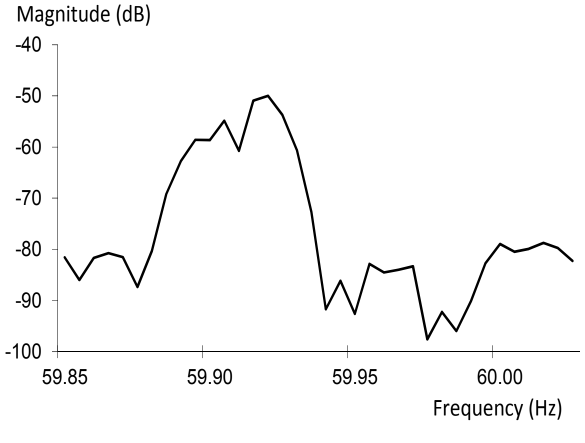


Fig. 2 FFT magnitude of the first window from STFT.

3. Autoregressive Approach

As previously mentioned, the AR approach was developed to cover the case of short audio recordings (e.g. less than 15 minutes) or short alterations in longer recordings. The zero-crossings method is the method that has previously been developed for such cases [1], however it requires much higher sampling rates (e.g. 8 kHz or 44.1 kHz as opposed to 0.2 Hz) and the computing complexity and storage capacity required are high so an alternative method is needed. A parametric model is defined on the assumption of a particular structure (“model”) for the signal being analyzed. This is in opposition to non-parametric methods that do not assume any structure in the signal, e.g. Fourier transforms which are applicable to any signals.

Autoregressive parametric models or pole-only models are suitable for signals with peaks or resonance frequencies, thus they are a good fit for the ENF modeling. The parameters or filter coefficients in an AR model can be estimated by finding the coefficients of a filter whose purpose is to predict the current sample of a random process by using a finite number of previous samples from the same process. As shown in Fig. 3, a signal $x[n]$ modeled by an AR process is produced by the output of an all-pole IIR (infinite impulse response) filter $H(z)$ of order M, with a white noise input $v[n]$ with zero mean μ_v. The filter response in the z-transform domain is described by Eq. (1) and in the time domain the system is described by the difference equation of Eq. (2):

$$H(z) = \frac{X(z)}{V(z)} = \frac{1}{1 - \sum_{k=1}^{M} w_k z^{-k}} \tag{1}$$

$$x[n] = v[n] + \sum_{k=1}^{M} w_k x[n-k] \tag{2}$$

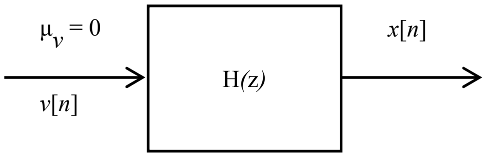


Fig. 3 Basic model for an AR process.

In general, the main objective of AR modeling is to compute the AR coefficients w_k in order to obtain an estimation of the PSD (power spectral density) $S_{xx}(e^{j\omega})$ of the output signal, which can be directly evaluated using Eq. (4) knowing the coefficients w_k and thus the resulting filter $H(z)$ in Eq. (1) and its magnitude frequency response $|H(e^{j\omega})|^2$. The variance σ_v^2 of the white noise input signal $v[n]$ is found from Eq. (5), producing the flat PSD $S_{vv}(e^{j\omega})$ for the input signal of the model:

$$S_{xx}(e^{j\omega}) = S_{vv}(e^{j\omega})\left|H(e^{j\omega})\right|^2 \qquad (3)$$

$$S_{vv}(e^{j\omega}) = \sigma_v^2 = r_x(0) - \sum_{k=1}^{M} w_k r_x(k) \qquad (4)$$

From $S_{xx}(e^{j\omega})$ the ENF frequency can be found by evaluating the peak position. Note that in order to extract the ENF frequency the knowledge of $S_{vv}(e^{j\omega})$ is not required, i.e., knowing $|H(e^{j\omega})|^2$ or $|H(e^{j\omega})|$ is sufficient. In our implementation a model of order M = 1 was found to be appropriate in order to detect a single peak at the final downsampled rate of 0.2 Hz. Many methods could be implemented for the parametric approach (e.g. Levinson-Durbin recursion) but for the purpose of our simple M = 1 model a direct programming of the Yule-Walker equations was performed. In this implementation, the sequence of samples $x[n]$ (in our case a window of 4 samples or equivalently 20 seconds) is used to determine the autocorrelation $r_x[n]$. These autocorrelation values are used in the set of Yule-Walker equations defined below. The size of the autocorrelation matrix R_x and the number of AR coefficients represented by the coefficient vector $\mathbf{w} = [w_1\ w_2 \cdots w_M]^T$ are dependent on the filter order M:

$$\mathbf{R}_x = \begin{bmatrix} r_x(0) & r_x(1) & \cdots & r_x(M-1) \\ r_x(1) & r_x(0) & \cdots & r_x(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(M-1) & r_x(M-2) & \cdots & r_x(0) \end{bmatrix} \qquad (5)$$

$$\mathbf{R}_x \mathbf{w} = \mathbf{r}_x \qquad (6)$$

$$\mathbf{w} = \mathbf{R}_x^{-1} \mathbf{r}_x \qquad (7)$$

with $\mathbf{r}_x = [r_x(1)\ r_x(2) \cdots r_x(M)]^T$. The following simple experiment illustrates how the ENF signal is extracted with the AR method. A digital audio recording of 5 minutes was sampled at 8 kHz, then downsampled to a rate of 0.2 Hz around the ENF frequency and analyzed with a window of 20 seconds (i.e., 4 samples at 0.2 Hz) with shifts of 5 seconds (i.e., 1 sample at 0.2 Hz). The magnitude frequency response of the filter produced by the AR model for the first window of 20 seconds is shown in Fig. 4. The frequency at which the maximum peak is located determines the ENF fluctuation sample. The algorithm was then executed for all the available windows, producing 57 ENF samples from the 5 minutes of audio recording, as depicted in Fig. 5. Comparisons between ENF signals extracted using the STFT method and the AR method will be presented in Section 5.

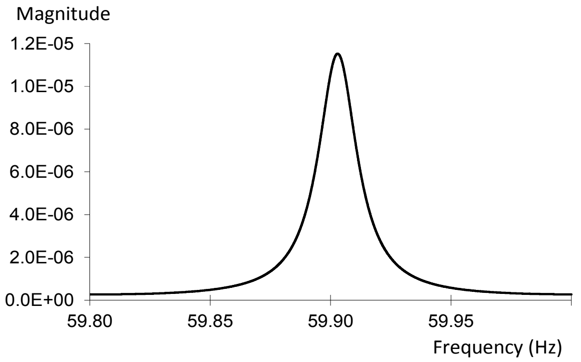


Fig. 4 Magnitude frequency response of the AR model for the first window (20 seconds of data).

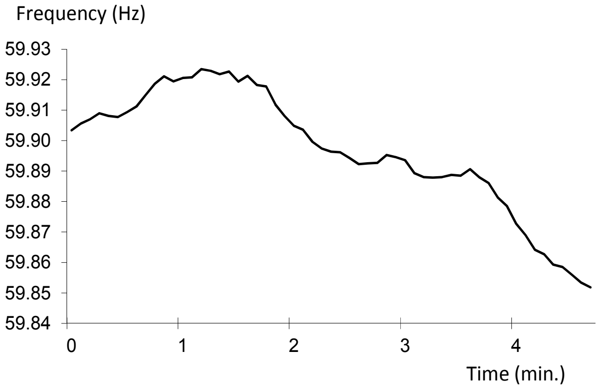


Fig. 5 Extracted ENF signal of 57 samples from 5 minutes of audio recording, using AR method.

4. Grid Signal Database

In order to build an electrical grid signal database, the grid information was captured via a probe and stored on a PC using a basic sound board. The grid database serves as a reference mark in the authentication process of a digital audio recording. The reference files constructing the database were saved in blocks of 12 hours. Even though the AR method is intended for short recordings, it is of course capable of handling longer signals. To illustrate this, proceeding with the same steps as the ones described in the previous section for an audio recording, the 60 Hz ENF signal extracted from 12 hours of grid reference signal is depicted in Fig. 6. If each original electrical grid data sample has 2 bytes, with mono-channel recording and a 44.1 kHz sampling rate (typical sound board setting), a day of recorded grid signal will require 7.6 GB of storage. Over one year this would require 2,774 GB. Table 1 illustrates the required storage space for different sampling rates. The database size can obviously be greatly reduced if the sampling rate of the recording process is lowered, i.e., if the sound board or recording device supports lower sampling rates, or if further offline downsampling processing is done. The use of a high sampling rate is only required if the method of zero-crossing is to be used.

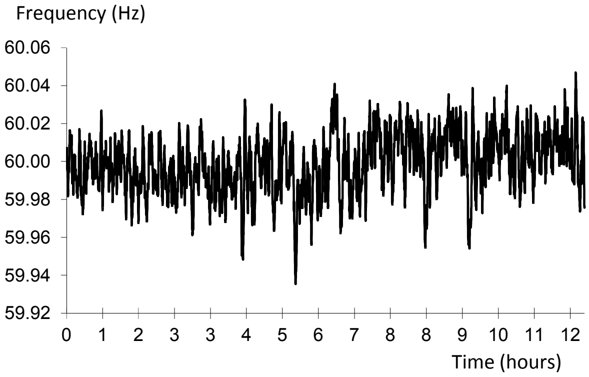


Fig. 6 Extracted grid ENF signal of 8,637 samples from 12 hours of recording, using AR method.

Table 1 Comparison of required storage for different sampling rates.

Sampling rate	Data size for 1 day of recording	Data size for 1 year of recording
44.1 kHz	~7.6 GB	~2.8 TB
32 kHz	~5.5 GB	~2 TB
8 kHz	~1.3 GB	~0.5 TB

5. Testing and Results

5.1 Correlation Approach

Correlating the ENF signal extracted from an audio signal with the ENF signal extracted from the grid database determines the audio ENF signal's "best match" in the grid ENF signal. Normalizing the correlation by the square root of the product of the total energy from both matching segments will produce a normalized correlation value between + 1 and – 1. A strong correlation between the audio ENF and the grid ENF means that the correlation value will be close to + 1, whereas a value closer to 0 represents a weak correlation between the signals. A correlation value close to – 1 would mean similar ENF waveforms with opposite signs, which is not an acceptable match in our application. Therefore, for our application it is only required to look for the maximum positive correlation coefficient value, indicating the location at which the best possible match occurs. The time where the maximum correlation value occurs is extracted in hours, minutes and seconds. This information can be used to obtain the exact date and time at which the audio recording has been made, by adding the detected maximum correlation time to the timestamp of the corresponding grid ENF signal (the timestamp was created by the recording computer and saved along the reference grid signal file, when capturing and recording the grid signal). If reference ENF signals from different electrical grids are available, it is then also possible to determine from which region the audio recording was made.

5.2 ENF Discontinuity Detection Technique

One of the most common ways of altering audio files is by applying the technique of "butt-splicing" [7, 8], removing a small audio part somewhere within the original digital audio recording and replacing it with another. In Ref. [8] it is stated that when an audio waveform is butt-spliced, the amplitude of the sample either side of the splice point is unlikely to be in equilibrium producing an abrupt change or discontinuity. However, meticulous alterations and manipulations render the audio signal hard to classify as unauthentic by simply looking at the time waveform or listening to the audio file. Although correctly butt-splicing a recording in the time domain is feasible, correctly placing an ENF frequency component can be much more laborious and burdensome. Moreover, combining signals recorded from different recording devices having different frequency bias will yield an ENF signal that is "discontinuous" in frequency at the time of the edit, therefore allowing detection of the butt-splicing. A basic first difference approach was implemented to detect sudden jumps in amplitude between two consecutive ENF samples. In this approach, Eq. 8 is applied on the audio ENF samples $x[n]$ of length N and $x'[n]$ are the first difference values:

$$x'[n] = \left| x[n] - x[n-1] \right| \quad 2 \le n \le N \qquad (8)$$

As an example, a first difference signal obtained from an audio signal that has been butt-spliced is depicted in Fig. 7 where the unusual peak corresponds to a sudden change in amplitude between two consecutive ENF samples. If the difference is larger than a set threshold the algorithm will label the audio as unauthentic. If several such altered segments are detected, a list of segment locations as well as the most likely alteration (highest first difference value) is provided. The audio recording is labelled as unaltered when the amplitude never exceeds the threshold. More results are presented in the next section.

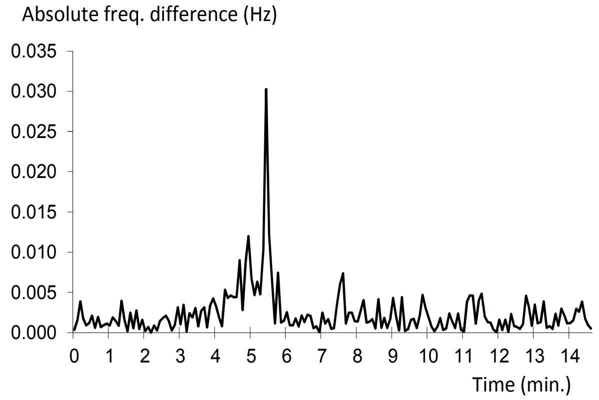


Fig. 7 Example of a first difference signal from an audio signal that has been butt-spliced for 10 sec. at approx. 5 minutes 30 sec.

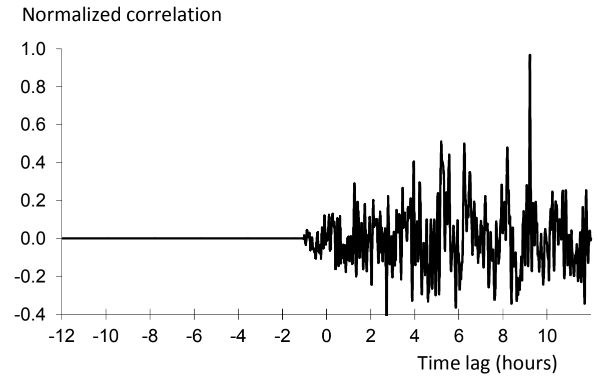
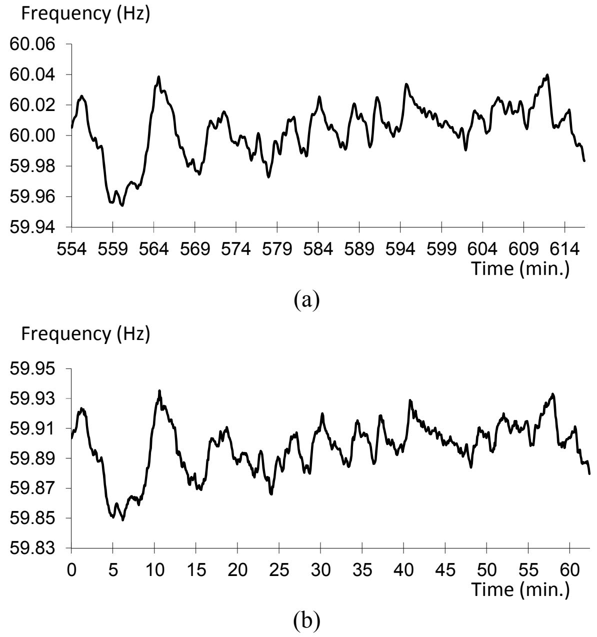


Fig. 8 Correlation graph obtained with AR method.

5.3 Experimental Results Comparing STFT and AR Methods

5.3.1 Case 1: Unaltered Audio Recording, 60 Minutes Length

The correlation graph for the ENF signals extracted from the audio and grid signals using the AR approach is presented in Fig. 8, where the maximum value of 0.967 occurs at a time around 9 hours 14 minutes with respect to the start of the grid file. A visual comparison of the corresponding grid ENF and audio ENF signals is shown in Fig. 9. Due to the visual similarity in the waveforms, and if no ENF discontinuity or butt-splicing is automatically detected in the audio ENF signal, it can be concluded that the audio recording is authentic. The correlation graph for the ENF signals extracted using the STFT method is shown in Fig. 10 where the maximum correlation value of 0.956 occurs again at a time around 9 hours 14 minutes with respect to the start of the grid file. Due to the longer length of the windows used in the STFT method (200 s vs. 20 s), the corresponding ENF waveforms will be much smoother (i.e., less noisy but also with less details or less time



(a)

(b)

Fig. 9 Grid ENF at 9 hours 14 minutes (top) and audio ENF (bottom) obtained with AR method.

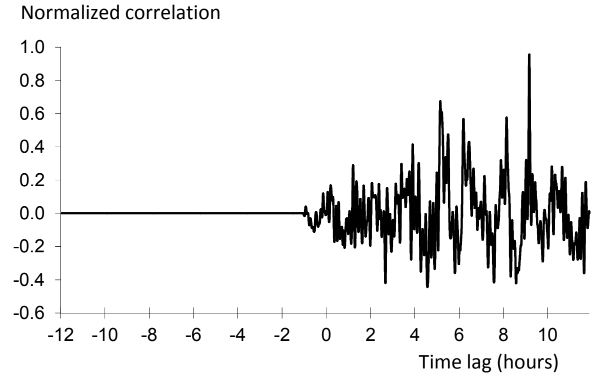
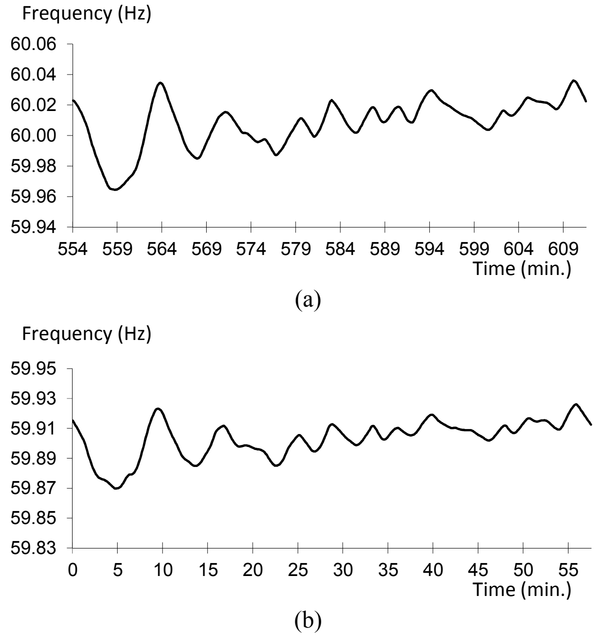


Fig. 10 Correlation graph obtained with STFT method.



(a)

(b)

Fig. 11 Grid ENF at 9 hours 14 minutes (top) and audio ENF (bottom) obtained with STFT method.

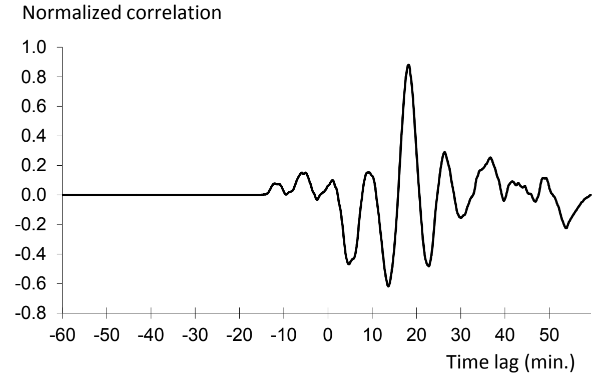
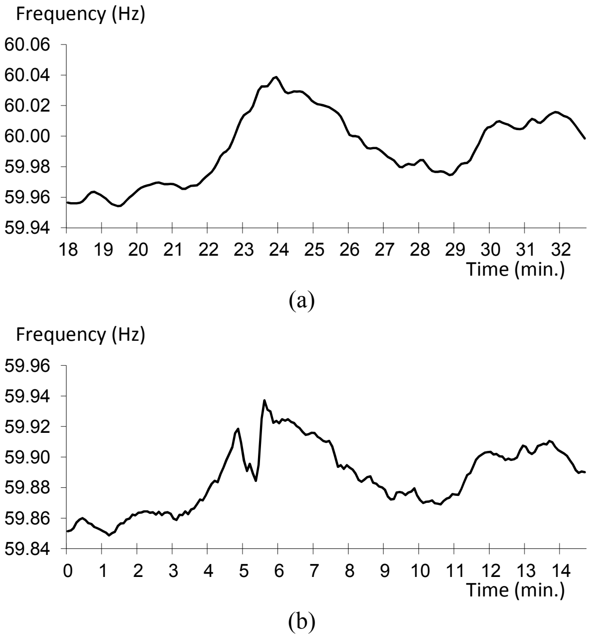


Fig. 12 Correlation graph obtained with AR method.

resolution) than the ENF waveforms obtained by applying the AR approach, as seen by comparing Figs. 11 and 9.

5.3.2 Case 2: Altered Audio Recording, 15 Minutes Length, 30 s of Content Modified

In this scenario, 30 s of audio were altered in a recording of length 15 minutes. The alteration was done using another segment of the same audio file. The ENF extraction from the audio and grid signals was first performed using the AR parametric approach. The maximum correlation value obtained is 0.879 occurring around 18 minutes from the start of the grid, as shown in Fig. 12. The lower correlation is already a first flag that the audio recording may not be authentic. Zooming-in on the ENF signals at the time corresponding to the maximum correlation value, it is clearly visible that the audio ENF has been tampered with, at approximately 5 minutes, as shown in Fig. 13. The discontinuity detection algorithm for butt-splicing also detects this (similar to Fig. 7, not shown here). When the STFT method is applied for extracting the audio and grid ENF signals, the maximum correlation value obtained is 0.911 at a time again around 18 minutes from the start of the grid, as seen in Fig. 14. However, the poor time resolution of the STFT method (because of the 200 s windows) makes it impossible to



(a)

(b)

Fig. 13 Grid ENF at 18 minutes (top) and audio ENF (bottom) obtained with AR method.

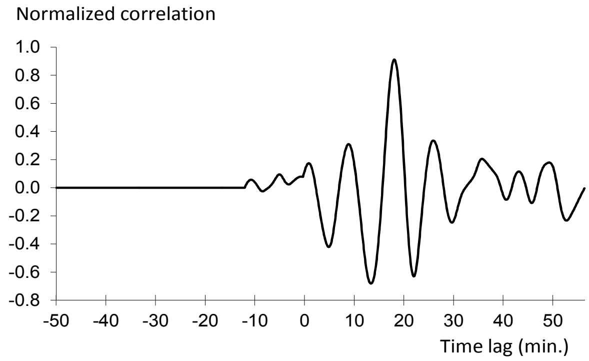
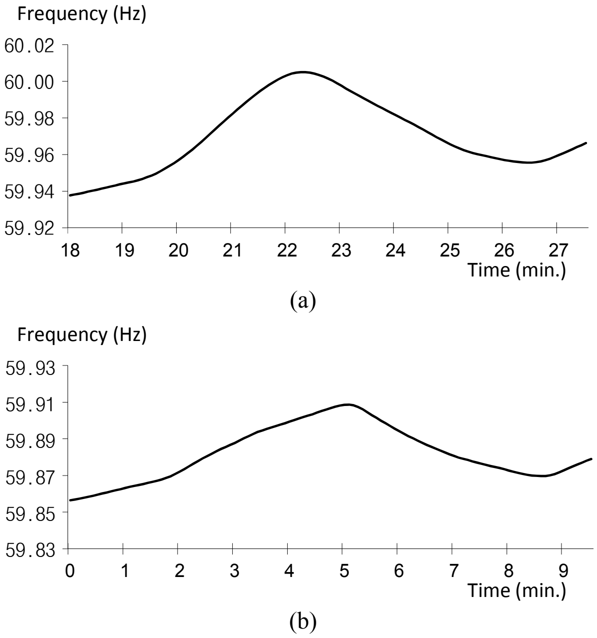


Fig. 14 Correlation graph obtained with STFT method.

perform a valid visual verification from Fig. 15, and from this smooth signal the discontinuity detection algorithm for butt-splicing also cannot detect any alteration. Therefore, we clearly see that for shorter audio recordings or alterations the better time resolution of the AR ENF extraction produces better results that the STFT ENF extraction. From the literature, it is expected that results similar to the AR ENF extraction method could possibly be obtained with the zero-crossings method [1], but as previously explained this method would be computationally expensive as it would operate at much higher sampling rates (e.g., 44.1 kHz or 8 kHz vs. 0.2 Hz for the AR method).



(a)

(b)

Fig. 15 Grid ENF at 18 minutes (top) and audio ENF (bottom) obtained with STFT method.

6. Conclusions

A new ENF extraction method based on AR modeling was developed, implemented and experimentally tested with real electrical grid recordings and audio recordings. The use of shorter windows in the AR method provided a better time resolution in the ENF signals and resulted in a better detection of short alterations in longer recordings, compared to the STFT method which uses longer windows. For a complete authentication, it was found that a combination of the correlation coefficient value, a visual verification of the strong matches between the audio and grid ENF signals and the use of a discontinuity detection algorithm for butt-splicing was required. From the results reported in Ref. [1], the zero-crossings ENF extraction method could possibly produce similar results to the proposed AR method, but it would also require more calculations and storage due to the fact that it operates at a high sampling rate (e.g., 44.1 kHz or 8 kHz vs. 0.2 Hz for the AR method). The zero-crossings ENF extraction method is also reported to be sensitive to the choice of parameters employed for the extraction [1]. Therefore, our proposed AR ENF extraction method can be an interesting alternative for extracting ENF fluctuations in audio forensic authentication.

Acknowledgements

This work was financially supported by the RCMP (Royal Canadian Mounted Police) and by a NSERC Discovery Grant. Special thanks to Dr. Matthew Case from RCMP for his collaboration.

References

[1] Grigoras, C. 2007. "Applications of ENF Criterion in Forensic Audio, Video, Computer and Telecommunication Analysis." *Forensic Science International* 167 (2-3): 136-45.

[2] Koenig, B. E., and Lacey, D. S. 2009. "Forensic Authentication of Digital Audio Recordings." *Journal of the Audio Engineering Society* 57 (9): 662-95.

[3] Liu, Y., Yuan, Z., Markham, P. N., Conners, R. W., and Liu, Y. 2012. "Application of Power System Frequency for Digital Audio Authentication." *IEEE Transactions on Power Delivery* 27 (4): 1820-28.

[4] North American Electric Reliability Corporation. NERC Interconnections. Accessed July 19, 2016. http://www.nerc.com/AboutNERC/keyplayers/Documents/NERC_Interconnections_Color_072512.jpg.

[5] North American Electric Reliability Corporation Resources Subcommittee. 2011. "Balancing and Frequency Control." NERC. Accessed July 19, 2016. http://www.nerc.com/docs/oc/rs/NERC%20Balancing%20and%20Frequency%20Control%20040520111.pdf.

[6] Grigoras, C., Smith, J. M., and Jenkins, C. W. 2011. "Advances in ENF Database Configuration for Forensic Authentication of Digital Media." In *Proceedings of the 131st Audio Engineering Society Convention,* 1-6.

[7] Grigoras, C., Rappaport, D., and Smith, J. M. 2012. "Analytical Framework for Digital Audio Authentication." In *Proceedings of the AES 46th International Conference-Audio Forensics: Recording, Recovery, Analysis and Interpretation,* 123-33.

[8] Cooper, A. J. 2011. "Detecting Butt-Spliced Edits in Forensic Digital Audio Recordings." In *Proceedings of the 39th International AES Conference-Audio Forensics: Practices and Challenges,* 11-21.