ENF SIGNAL EXTRACTION FOR ROLLING-SHUTTER VIDEOS USING PERIODIC ZERO-PADDING

Jisoo Choi and Chau-Wai Wong
North Carolina State University, Raleigh, USA

ABSTRACT

Electric Network Frequency (ENF) analysis is a promising forensic technique for authenticating digital recordings and detecting tampering within the recordings. The validity of ENF analysis heavily relies on high-quality ENF signals extracted from multimedia recordings. In this paper, we propose an ENF signal extraction method for rolling shutter acquired videos using periodic zero-padding. Our analysis shows that the extracted ENF signals using the proposed method are not distorted and the component with the highest signal-to-noise ratio is located at the intrinsic frequency. The experimental results show that our proposed method can generate more precise ENF signals than those from the state-of-the-art method.

Index Terms—Electric Network Frequency (ENF) signal, frequency estimation, multirate analysis, rolling shutter

1. INTRODUCTION

Electric Network Frequency (ENF) analysis is a promising forensic technique for authenticating digital recordings [1] and detecting tampering within the recordings [2]. The ENF is the supply frequency of an electric grid. The supply frequency is not constant, but fluctuates around its nominal value of 60 Hz in North America or 50 Hz in most other regions of the world, due to the mismatch between the demand and the supply within the power network. Since different nodes within the grid are interconnected, the fluctuations in frequency at different locations of the same grid share similar patterns. The instantaneous values of ENF over time are regarded as the ENF signal, which can serve as a natural time stamp for authenticating multimedia recordings.

The ENF signal can be embedded into audio recordings via sensing acoustic vibrations or via interfering electromagnetically in sensing circuits [1]. Studies in [1, 3–6] show that ENF traces extracted from audio recordings can be used to assess the authenticity of the recordings. Studies in [7, 8] show that ENF extracted from the audio track of videos can be used to identify the locations that videos were recorded. The study in [9] shows that the ENF analysis can be extended beyond the realm of the forensic science to multimedia signal processing, e.g., the synchronization of videos without overlapped scenes, and the alignment of historical audio recordings.

More recent studies revolve around extracting ENF traces from the visual track of multimedia recordings [10–12]. Indoor lightings, such as fluorescent lights and incandescent bulbs, vary their light intensity at a frequency twice of that of the supply voltage, leading to near-invisible flickering in the illuminated environment. Consequently, cameras under the indoor illumination environment may capture videos that contain ENF signals. One of the major concerns encountered in extracting the embedded ENF signal from visual tracks is the aliasing effect due to limited sampling rate [10]. Specifically, when the nominal value of the ENF and the frame rate are 60 Hz and 30 frames per second (fps), respectively, it is challenging to extract the aliased ENF component because it is overlaid with the DC frequency component.

To tackle the issue of the limited sampling rate, the authors in [10, 11] demonstrated that the rolling shutter mechanism, which is traditionally considered detrimental in image and video analysis, can be exploited to increase the effective sampling rate. The rolling shutter acquires a video by sequentially reading and storing the pixel values of horizontal lines of each landscape frame. Since successive lines of a frame are acquired at different time points, the rolling shutter can foster the ENF signal extraction from the visual track by increasing the effective sampling rate by a factor of the number of lines [10]. In [11], the authors proposed an extraction method that directly concatenates the row signals of all frames by ignoring the idle periods occurred at the end of each frame. Multirate signal analysis was applied to show that the extracted signal approximates the desired ENF signal.

In this paper, we conduct a further study on exploiting the rolling shutter and propose a periodic zero-padding method for extracting the exact, undistorted ENF signals from the visual track. The proposed extraction method conducts an equivalently uniform sampling along the time by returning zeros during the idle period at the end of each frame. Our analysis will show that the spectrogram around the frequency of interest is not distorted and the component with the highest signal-to-noise ratio is located at the intrinsic frequency.

The rest of the paper is organized as follows. In Section 2, we describe and analyze the proposed method for ENF signal extraction from videos. The experimental results are reported in Section 3 and conclusions are drawn in Section 4.
2. PROPOSED EXTRACTION METHOD USING PERIODIC ZERO-PADDING

Rolling Shutter Preliminaries  The rolling shutter acquires each image over the frame period $T_c$ during which each row of the frame is exposed followed by the idle period $T_{idle}$ that no row is captured [13]. The read-out time $T_{ro}$ [14] is defined as the amount of time required for the rows of a frame to be generated and is unique for each camera model. The instantaneous light intensity is captured by all pixels of the same row simultaneously. When the foreground is of uniform color or is removed by motion compensation, the average value of simultaneous light intensity is captured by all pixels of the same row [13].

In [11], the rolling shutter mechanism was exploited to extract ENF trace from the visual track. During the frame period $T_c$, the rolling shutter can capture $M$ samples to its capacity, among which only $L$ samples recorded during the read-out time $T_{ro}$ will be retained as the rows of each frame. The remaining $M - L$ samples exposed during the idle period $T_{idle}$ will be discarded, where $L \leq M$ and $T_c = T_{ro} + T_{idle}$. They concatenate $L$ samples into one single row signal and we refer to this method as the direct concatenation method. The rolling shutter’s effective sampling rate $F_s$ can be defined in the pair of $(M, T_c)$ or $(L, T_{ro})$ [14] as $F_s = 1/T_c = M/T_c = L/T_{ro}$, where $T_s$ is the amount of time between the start of sampling one row and the start of sampling the subsequent row. In this case, a perceptual sampling frequency $F_{ps} = L/T_s$ was adopted for the frequency domain analysis without the need of knowing the value of $T_{ro}$.

Proposed Method  In this paper, we propose a periodic zero-padding method for extracting ENF signals from the visual track. Fig. 1 illustrates timing schedule for acquiring the nonzero samples from row 1 to row $L$ and zero samples from row $L + 1$ to row $M$. Unlike the direct concatenation method, the estimation of $T_{ro}$ is needed for calculating the value of $M$ via $L \cdot T_c/T_{ro}$. We denote the time domain input and output signals by $x(n)$ and $y(n)$, respectively, as shown in Fig. 2(a). Out of all $M$ points of the input signal $x(n)$ that corresponds to one frame, we zero out the last $M - L$ samples that correspond to the idle period and denote the resulting output signal as $y(n)$. Such input-output relationship can be modeled by a $L$-branch filter bank with $M$-factor downsamplers and upsamplers and unit delay operators, as shown in Fig. 2(b). The DTFT of the $l$th branch signal can be expressed as follows:

$$Y_l(\omega) = \frac{1}{M} \left( \sum_{m=0}^{M-1} X \left( \omega - \frac{2\pi m}{M} \right) e^{j\left(\omega - \frac{2\pi m}{M}\right)l} \right) e^{-j\omega l}. \quad (1)$$

Hence, the final output $Y(\omega)$ as a summation of all $L$ branches is

$$Y(\omega) = \sum_{l=0}^{L-1} Y_l(\omega) = \sum_{m=0}^{M-1} A_m \cdot X \left( \omega - \frac{2\pi m}{M} \right). \quad (2)$$

where

$$A_m = \frac{1}{M} \sum_{l=0}^{L-1} \exp \left( -j \frac{2\pi ml}{M} \right) = \frac{L}{M} \text{asinc}_L \left( \frac{2\pi m}{M} \right) \cos \left( -j \frac{\pi m}{M} (L - 1) \right), \quad (3)$$

in which $\text{asinc}_L(x) = \frac{\sin(Lx/2)}{Lx/2}$ is the aliased sinc function or Dirichlet function [15]. It is a periodic function that has zero crossings at integer multiples of $\frac{2\pi}{T_c}$ within one cycle and its maximum values are located at integer multiples of $2\pi$. Replacing the normalized angular frequency $\omega$ in (2) by $\frac{2\pi F}{T_c}$, we obtain the output as a function of the analog frequency $F$ as follows:

$$Y(F; F_s) = \sum_{m=0}^{M-1} A_m \cdot X \left( F - mF_s; F_s \right), \quad (4)$$

where $g(F; F_s) \overset{\text{def}}{=} g \left( \frac{2\pi F}{F_s} \right)$ for $g = X$ and $Y$. Eq. (4) shows that $Y(F; F_s)$ is the weighted sum of $X(F; F_s)$’s that are
shifted by integer multiples of the frame rate \( F_c = 1/T_c \). \( Y(F; F_c) \) has nonzero ENF frequency components at

\[
F = \pm F_c + mF_c + \nu F_s,
\]

where \( m, \nu \in \mathbb{Z} \) and \( F_s \) is defined as the doubled frequency of the nominal ENF. Note that in Eq. (4), the 0th summation term is the ideal input source signal \( X(F; F_c) \) scaled by \( A_0 \). Observe that the shift step \( F_c = F_s/M \), which means that the shifted copies \( X(F - mF_c; F_s) \), \( m = 0, \cdots, M - 1 \) will evenly occupy the full sampling range \([0, F_s]\) Hz. The frequency components that used to be aliased under the slow sampling rate \( F_s \) are now well within the range under the rolling shutter’s effective sampling frequency \( F_c \).

In order to obtain a high quality extracted ENF signal in terms of the signal-to-noise ratio (SNR), one simple but effective strategy is to pick the frequency component of the highest SNR strength, out of the \( M \) copies \( A_mX(F - mF_c; F_s) \) for \( m = 0, \cdots, M - 1 \). This is because the noise power is a constant value \((L/M)\sigma^2\), for all frequency components when the input signal \( x(n) \) is corrupted by additive white Gaussian noise with power \( \sigma^2 \). Since \( |A_m| = \frac{\pi}{F_s} |\text{asinc}(\frac{2\pi m}{M})| \), the set of elements \( m^* \) and \( \nu^* \) which maximizes the SNR or equivalently, \( |A_m| \), can be found by the following optimization problem:

\[
\min_{m, \nu \in \mathbb{Z}} \frac{2\pi m}{M} - 2\pi \nu, \tag{6}
\]

which leads to the solution \( m^* = Mk^* \). Substituting \( m^* \) into (5), we obtain the frequency components of the strongest ENF signals as follows:

\[
F_{(m^*, k^*)} = \pm F_c + (k^* + \nu^*) F_s, \tag{7}
\]

which implies that our proposed method always results in the strongest ENF signal at \( F_c \) within the range of \([0, 0.5F_s]\) Hz.

Due to the space limitation, we skip the proof and directly provide the analytic form of the positive frequency component achieving the highest SNR for the direct concatenation method as follows:

\[
F_{(m^*, k^*, \nu^*)} = F_c + \text{round}((T_m - T_c)F_s)F_s + (k^* + \nu^*) F_p, \tag{8}
\]

where \( \text{round}(x) \) returns the nearest integer of \( x \). It reveals that the direct concatenation method needs both \( T_c \) and \( T_m \) to determine the location of the shifted frequency component with the strongest ENF signal. Table 1 shows the calculated frequencies that achieve the maximum SNR for the direct concatenation method and our proposed method when \( F_c = 120 \) Hz and \( T_m = 19.8 \) ms. The proposed method guarantees that the ENF trace with the highest SNR can always be extracted at the doubled frequency of the nominal ENF, i.e., \( F_c = 120 \) Hz.

Another advantage of the proposed method over the direct concatenation method is that the Fourier spectrum and hence the extracted ENF signals are not distorted. Our proposed method has a scaling factor \( A_m \) that uniformly scales the magnitude of the ENF signal along the frequency axis, whereas the direct concatenation method has a scaling function \( A_{m'}(F) \) that distorts the spectrum along the frequency. However, one should note that such distortion has a minor impact on the quality of the extracted ENF signal as the frequency spread of ENF signals is usually narrow.

### Table 1: Frequency band achieving the maximum SNR.

<table>
<thead>
<tr>
<th>Case</th>
<th>Frame rate (fps)</th>
<th>Frequency achieving maximum SNR (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30.0000</td>
<td>Method in [11]</td>
</tr>
<tr>
<td>2</td>
<td>23.0002</td>
<td>Proposed</td>
</tr>
<tr>
<td>3</td>
<td>16.0032</td>
<td></td>
</tr>
</tbody>
</table>

#### 3. Extraction of ENF Traces

**Experimental Conditions** We conducted experiments using the back camera of iPhone 6s in an indoor environment with electric lighting in Raleigh, USA. The videos were acquired by facing the mobile camera toward a white wall, and we used the read-out time 19.8 ms reported in [16]. We recorded the power mains signal in parallel and treated it as the ground-truth ENF signal. To calculate a spectrogram for the ENF signal, we split the signal into frames of 12 seconds with 90% overlap. Quadratic interpolation was used to refine the locations of detected spectral peaks [17].

We calculate the SNR at a frequency of interest, \( F \), using the ratio of estimated signal and noise power from the empirical power spectrum within a small neighborhood around \( F \). We estimate the noise power \( P_{\text{noise}} \) as the averaged power within the frequency ranges \((F - \delta_1, F - \delta_2)\) and \((F + \delta_2, F + \delta_1)\). We estimate the signal power \( P_{\text{signal}} \) by subtracting the estimate of \( P_{\text{noise}} \) from the peak power spectrum value at \( F \). In our experiment, we empirically chose \( \delta_1 = 2 \) Hz and \( \delta_2 = 0.5 \) Hz.

**Results and Discussions** Fig. 3 illustrates the spectrogram of the row signal generated by the proposed periodic zero-padding method. In this example, the frame rate of the camera is set to be \( F_c = 23.0062 \) fps under the nominal value of ENF at 60 Hz to verify whether its spectral composition matches the theoretical prediction. The choice of \( F_c = 23.0062 \) fps is to avoid the mix of DC and ENF components. In the case of \( F_c/F_p \in \mathbb{Z} \), e.g., \( F_c = 120 \) Hz and \( F_p = 30 \) fps, DC and ENF from different components will overlap. In Fig. 3(a), we can observe the spectral distribution resulted from the periodic zero-padding method where the multiple DC components appear at intervals of \( F_p \) in thick red strips and multiple copies of ENF signal are in faint red strands. Fig. 3(b) shows a zoomed-in region around 120 Hz, where we can see the ENF signals and DC signals appear consistent with the prediction from our theoretical model in Section 2.
Fig. 3: (a) A spectrogram for the row signal of $F_e = 120$ Hz and $F_c = 23.0062$ fps generated by the proposed periodic zero-padding method. The thick straps are DC components, and the faint straps are the ENF components. (b) Zoomed-in version around 120 Hz.

Fig. 4: The ENF signals extracted using the periodic zero-padding method from a static white-wall test video when $F_e = 120$ Hz and $F_c = 30$ fps: [Middle] The ENF signal extracted around $F_e = 120$ Hz achieves the highest SNR; [Top & Bottom] The adjacent ENF signals at 150 Hz and 90 Hz, respectively.

Fig. 5: The ENF trace (black solid) from the maximum SNR of Fig. 4 and its ground-truth ENF signal from power (red dotted). The black ENF trace was properly shifted and scaled to be aligned with the red ENF trace.

Table 2: Comparison of the proposed method with prior art under the low SNR conditions.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>Average of correlations</th>
<th>Method in [11]</th>
<th>Proposed</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-10$</td>
<td>0.959</td>
<td>0.959</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>$-20$</td>
<td>0.858</td>
<td>0.892</td>
<td>0.01**</td>
<td></td>
</tr>
<tr>
<td>$-23$</td>
<td>0.754</td>
<td>0.839</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>$-25$</td>
<td>0.459</td>
<td>0.551</td>
<td>0.01**</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 5 shows extracted ENF signals from different frequencies when the frame rate is 30 fps, a commonly used sampling rate. The figure reveals that the ENF signal with the highest SNR is achieved at 120 Hz, and the top and bottom plots extracted from the nearest bands at 150 Hz and 90 Hz have lower SNRs. The observation is consistent with the prediction from our theoretical result. It is interesting to note that DC components do not have a strong negative effect on the accuracy of the frequency estimation when overlapping with the ENF components. We plan to investigate the characteristics of the DC components in future work.

To quantitatively compare the extracted ENF signal around 120 Hz of Fig. 4 to the ground-truth ENF signal, we use three matching criteria: the normalized cross-correlation (NCC), the mean squared deviation (MSD), and the mean absolute deviation (MAD). The calculated NCC, MSD, and MAD with respect to the reference signal are shown in Fig. 5. Based on the matching criteria, we confirm that the ENF signal estimate by the periodic zero-padding method is similar to the power reference ENF signal.

To compare the performance of our proposed method with the direct concatenation method [11], we extract ENF signals using the two methods and split each signal into 6 segments to obtain multiple correlation coefficients. We then do a t-test to see if the group means of the correlation coefficients are significantly different in the statistical sense. Table 2 reveals that our proposed method is statistically better than the direct concatenation method at $-20$ dB and $-25$ dB, and comparable at $-10$ dB and $-23$ dB.

4. CONCLUSION AND FUTURE WORK

In this paper, we have proposed an alternative method for extracting ENF signals by continuing sampling during the idle period. We have analyzed the proposed method using the multirate filter bank model, which shows that the extracted ENF signals are not distorted and the highest SNR component is located at the intrinsic frequency. The experimental results support the theoretical analysis and show that our proposed method can generate more precise ENF signals than the state-of-the-art method. Future work would be conducting extensive evaluations on a large dataset that contains different camera models and SNR conditions.
5. REFERENCES


