ASSESSING SPECTRAL ESTIMATION METHODS FOR ELECTRIC NETWORK FREQUENCY EXTRACTION

Georgios Karantaidis
Aristotle University of Thessaloniki
Department of Computer Science
Thessaloniki, Greece
gkarantai@csd.auth.gr

Constantine Kotropoulos
Aristotle University of Thessaloniki
Department of Computer Science
Thessaloniki, Greece
costas@aiia.csd.auth.gr

ABSTRACT

The Electric Network Frequency (ENF) criterion provides useful forensic evidence for multimedia authentication. In this paper, a systematic study of non-parametric and parametric spectral estimation methods is conducted for ENF extraction. Fast implementations of the Capon method and the Iterative Adaptive Approach, which exploit the Gohberg-Semencul factorization of the inverse covariance matrix, are included as well. When long segments are used, a very high matching accuracy is achieved. That is, the maximum correlation-coefficient between the extracted ENF and the ground truth may exceed 99%. Similarly, the standard deviation of error may be as small as $1.069 \cdot 10^{-3}$. Non-parametric spectral estimation techniques are shown to be able to detect an alteration in an audio recording, where a short utterance recorded in Europe is replaced by the same content recorded in the US.

CCS CONCEPTS

• **Applied computing → Computer forensics**; **Investigation techniques**;

KEYWORDS

Electric Network Frequency, Spectral Estimation Methods, Fast Algorithms, Matching Procedures, Multimedia Authentication

ACM Reference Format:
Georgios Karantaidis and Constantine Kotropoulos. 2018. ASSESSING SPECTRAL ESTIMATION METHODS FOR ELECTRIC NETWORK FREQUENCY EXTRACTION. In *22nd Pan-Hellenic Conference on Informatics (PCI '18), November 29-December 1, 2018, Athens, Greece.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3291533.3291538

1 INTRODUCTION

Multimedia forensic analysis is widely used due to the rapidly increased volume of shared audio and video recordings. However, the multimedia content can be edited, altered, and modified for various purposes. In forensic sciences, authenticating digital content and determining the time and place of recording are critical tasks.

A forensic tool, which is used for forgery detection in multimedia recordings is the Electric Network Frequency (ENF) criterion [9]. ENF is the supply frequency in power distribution networks and its nominal value is 50 Hz in Europe and 60 Hz in the U.S. The major property of ENF signal is that its value is fluctuating in a random way around its nominal value. These fluctuations are assumed to be identical through an inter-connected network.

Adaptive techniques for ENF extraction are proposed in [15], where a detailed comparison between various techniques is made. In [11], different methods of ENF estimation are elaborated, addressing the problem of geo-location estimation from the ENF signal. A more precise and detailed study focusing on determining the intra-grid location of recordings is discussed in [3]. In [5], the ENF signal is modeled as an autoregressive process. Computationally efficient maximum-likelihood estimation via a multitone harmonic model is presented in [1]. Apart from digital audio recordings, the ENF signal can be extracted from digital video content recorded in indoor environments with the presence of fluorescent lighting [4] in order to estimate the time of recording and verify its authenticity.

In this paper, we assess case studies of ENF extraction, resorting to either non-parametric spectral estimation methods (e.g., periodogram and refined periodogram methods, such as Blackman-Tukey, Welch and Daniell, Capon spectral estimator, Interative Adaptive Approach [IAA]) or parametric ones (e.g., Estimation by Rotational Invariance Techniques [ESPRIT], Multiple Signal Classification [MUSIC]). Fast algorithms for Capon and IAA spectral estimation methods are included as well by exploiting the Gohberg-Semencul factorization of the inverse covariance matrix [7, 8, 19]. All methods are applied to consecutive frames of data recorded from the power mains as well as the audio recording as used in [15]. The fundamental ENF and its harmonics are estimated by tracking the maxima of the power spectrum, and applying quadratic interpolation in each frame [15]. Motivated by Professor Petre Stoica's "Spectral estimation is an art", here we put emphasis on the details of band-pass (BP) filtering of raw signal prior to spectral analysis and the fine tuning of parameters involved in spectral analysis techniques, which enable us to report more accurate results than those disclosed in [1, 15]. This is the first contribution of the paper. Besides the fundamental frequency, ENF extraction is carried out in its higher harmonics, which demonstrate a higher SNR than the fundamental one, yielding better results, as observed also in [14]. In addition to existing matching procedures between the extracted ENF time series and the ground truth one of equal length, efficient dynamic time warping (DTW) is employed and assessed, which allows the aforementioned time series to have different lengths and eliminates the need to downsample the ground truth ENF time series, as is tacitly assumed in [15]. This is the second contribution of the paper. When long stationary segments of the extracted ENF time series (e.g., having duration 20 sec or so)

PCI '18, November 29-December 1, 2018, Athens, Greece

ACM ISBN 978-1-4503-6610-6/18/11…$15.00
https://doi.org/10.1145/3291533.3291538

Table 1: Frame parameters (in sec)

Parameters	Data 1	Data 2
Time shift, T	1	1
Frame length, L_1[15]	20	33
Frame length, L_2	40	50

are used, a very high matching accuracy is achieved. Such long segments are not always available in practice. Accordingly, the third contribution of the paper is in assessing ENF extraction methods from short utterances. To do so, an alteration in an audio recording was devised by replacing a short utterance recorded in Europe by the same utterance recorded in the US with the latter being perceptually indistinguishable from the former in order to assess the limits of spectral estimation methods for ENF extraction. Simple non-parametric spectral estimation techniques, such as the periodogram and Daniell method, are shown to be able to detect the aforementioned alteration in the audio recording. To the best of authors' knowledge, such an experiment is conducted for first time.

The rest of the paper is organized as follows. In Section 2, the datasets used and the band-pass filtering of raw signals prior to spectral analysis are described. Spectral analysis methods for ENF estimation are briefly presented in Section 3. The results of all methods are demonstrated in Section 4. Section 5 concludes the paper and proposes topics of future research.

2 DATASET DESCRIPTION

The two datasets discussed in [15] and the ENF ground truth associated to them are used here as well. The first dataset (Data 1) was recorded by connecting an electric outlet directly to the internal sound card of a desktop computer and the second one (Data 2) was a speech recording captured by the internal microphone of a laptop computer. For both recordings the initial sampling frequency was $44.1\,kHz$. At first, the original recordings are downsampled to a frequency that contains the fundamental frequency and some of its higher harmonics, i.e., $F_s = 441\,Hz$. The next step includes band-pass filtering of the signal around the nominal ENF or its harmonics. For the signal recorded from the power mains, the band-pass edges of the filter are set at $59.9\,Hz$ and $60.1\,Hz$ for the fundamental frequency and the filter order is 1501. The second harmonic band-pass edges are set at $119.9\,Hz$ and $120.1\,Hz$, respectively. The filter order remains the same, i.e., 1501. The third harmonic band-pass edges are set at $179.9\,Hz$ and $180.1\,Hz$ and the filter order is equal to 1001. In each case, a Hamming window is used. For the audio recording, the band-pass edges are set at $119.95\,Hz$ and $120.05\,Hz$ and the filter order increased to 4801. The different specifications are due to the "noisy" nature of the audio recording. Next, the filtered signal is split into K overlapping frames. Each frame is obtained by applying a rectangular window of length L sec and is shifted by T sec from its immediate predecessor frame. Two choices for L, denoted as L_1 and L_2, are indicated in Table 1 along with T. In each frame, the power spectrum is estimated by various spectral analysis techniques.

A third dataset was created by concatenating the 10 recordings uttered by TIMIT female speaker ID `TEST\DR1\FAKS0` [6] to create an audio signal recorded in the US. The same 10 utterances were played back from the loudspeakers of a notebook connected to the power mains at Thessaloniki, Greece and recorded by various mobile phones during the collection of the MOBIPHONE database [13]. The European (EU) recording has duration 38.5677 sec. The utterance `SA2` "Don't ask me to carry an oily rag like that" in the European recording was replaced by the same utterance recorded in the US in a perceptually indistinguishable manner. By doing so, the European recording was altered by inserting a 3.6288 sec long utterance with nominal ENF of $60\,Hz$ starting from 5.3125 sec and ending at 8.9413 sec, while the remaining recording has a nominal ENF of $50\,Hz$. Let us refer to the third audio signal as mixed recording. Both the mixed recording and the European one have the same duration. The sampling frequency of all recordings is $16\,kHz$. Downsampling by a factor of 10 is found beneficial for the US and EU recordings. These recordings were filtered by a Finite Impulse Response (FIR) BP filter of order 4801 centered at the second harmonic of ENF with band-pass edges set at $119.95\,Hz$ and $120.05\,Hz$ as well as $99.95\,Hz$ and $100.05\,Hz$, respectively. On the contrary, the mixed recording was filtered by an FIR BP filter of order 1501 with band-pass edges set at $49.95\,Hz$ and $60.05\,Hz$.

3 SPECTRAL ESTIMATION METHODS FOR ENF EXTRACTION

Assuming stationarity within the frame, the simplest non-parametric method for estimating the ENF is the short-time Fourier Transform (STFT). That is, frame by frame, the periodogram of each frame is computed by squaring the magnitude of the STFT. Let $\hat{\phi}_r(\omega_q)$ be the periodogram of the $N{=}L\,F_s$ samples long rth frame, where $\omega_q{=}\frac{2\pi}{Q}\,q$, $q{=}0, 1, \ldots, Q-1$ are the frequency samples and F_s is the sampling frequency. Typically, $Q \geq N$, i.e., $Q{=}4\,N$. The frequency sample $\omega_{q_{\max}}$, which corresponds to the maximum periodogram value is extracted as a first ENF estimate. Next, a quadratic interpolation is employed, which fits a quadratic model to the logarithm of the estimated power spectrum about $\omega_{q_{\max}}$ [15, 17]. Hereafter, the aforementioned spectral estimation method is replaced by other non-parametric and parametric spectral analysis methods.

A refined periodogram method is the Welch method [18]. In this method, each frame is divided into overlapped segments and each segment is multiplied by a temporal window. Let $y_j(t)$ denote the jth segment. Adjacent segments overlap by 1000 samples and each segment has length of $M = \frac{N}{4} = \frac{LF_s}{4}$ samples. The Welch estimate of power spectral density (PSD) is given by $\hat{\phi}_w(\omega) = \frac{1}{S}\sum_{j=1}^{S}\hat{\phi}_j(\omega)$, where $S = 7$ and $\hat{\phi}_j(\omega)$ is the windowed periodogram corresponding to $y_j(t)$. Here, a rectangular window has been employed. The Welch method yields accurate ENF estimation without being affected by interferences, especially in the second harmonic of the ENF in both datasets.

The Welch estimator can be related to the Blackman-Tukey (BT) spectral estimator for suitable choices of the lag window and the auto-covariance estimate [18]. Accordingly, a natural choice for a refined periodogram is the Blackman-Tukey estimate given by $\hat{\phi}_{BT}(\omega) = \sum_{q=-(M-1)}^{M-1} w(q)\,\hat{r}(q)\,e^{-i\omega\,q}$, where $M = \frac{N}{2} = L\,\frac{F_s}{2}$ for the first and third harmonic and $M = N = L\,F_s$ for the second harmonic in both datasets. Another non-parametric method is the Daniell method [18], which yields the refined spectral estimate $\hat{\phi}_D(\omega_q) = \frac{1}{2J+1}\sum_{j=k-J}^{k+J}\hat{\phi}_p(\omega_j)$ for dense frequency samples $\omega_q =$

$\frac{2\pi}{Q}\ q,\ q = 0, 1, \ldots, Q-1$. Here, the values $J = 2$ and $Q = 4\,N = 4\,L\,F_s$ have been used.

The periodogram can be interpreted as a filter bank approach, which uses a band-pass filter whose impulse response vector is given by the standard Fourier transform vector $\boldsymbol{\alpha} = \left[1, e^{-i\omega}, \ldots, e^{-i\,(N-1)\,\omega}\right]^T$. Let $\hat{\mathbf{R}}$ be an estimate of the auto-covariance matrix

$$\hat{\mathbf{R}} = \frac{1}{N-m} \sum_{t=m+1}^{N} \begin{bmatrix} y(t) \\ \vdots \\ y(t-m) \end{bmatrix} \left[y^*(t), \ldots, y^*(t-m)\right] \tag{1}$$

The Capon method is a filter bank approach based on a data-dependent filter [18]: $\mathbf{h} = \frac{\hat{\mathbf{R}}^{-1}\,\mathbf{a}(\omega)}{\mathbf{a}^*(\omega)\,\hat{\mathbf{R}}^{-1}\,\mathbf{a}(\omega)}$, where $\mathbf{a}(\omega) = \left[1, e^{-i\omega}, \ldots, e^{-i\,m\,\omega}\right]^*$ and $[\cdot]^*$ denotes conjugate transposition. The Capon spectral estimate is given by:

$$\hat{\phi}(\omega) = \frac{m+1}{\mathbf{a}^*(\omega)\,\hat{\mathbf{R}}^{-1}\,\mathbf{a}(\omega)} \tag{2}$$

computed for dense frequency samples $\omega_q = \frac{2\pi}{Q}\ q, q = 0, 1, \ldots, Q-1$ with $Q = 300\,m$ and $m = 10$ for the second and third harmonics of Data 1 and Data 2. The first harmonic is computed with $m = 2$ and $Q = 5000\,m$ in each case.

The IAA is a non-parametric alternative to weighted Least Squares method as presented in [20]. Let $\mathbf{y}_N = \left[y(t), \ldots, y(t+N-1)\right]^T$ be the data vector. Let also $\mathbf{f}_N(\omega_q) = \left[1, e^{i\omega_q} \ldots, e^{i\omega_q(N-1)}\right]^T$ be the frequency vector, where $q = 0, 1, \ldots, Q-1$ and Q is the number of frequency samples taken as a multiple of N. Assume $\mathbf{F}_{N,Q} = \left[\mathbf{f}_N(\omega_0)| \ldots |\mathbf{f}_N(\omega_{Q-1})\right]$. The sample covariance matrix is given by $\mathbf{R}_N = \mathbf{F}_{N,Q}\mathbf{P}_Q\mathbf{F}^*_{N,Q}$, where $\mathbf{P}_Q$ is the diagonal matrix whose diagonal elements are obtained by the squared magnitude of the following estimate

$$x_q = \frac{\mathbf{f}^*_N(\omega_q)\,\mathbf{R}^{-1}_N\,\mathbf{y}_N}{\mathbf{f}^*_N(\omega_q)\,\mathbf{R}^{-1}_N\,\mathbf{f}_N(\omega_q)} \tag{3}$$

at the previous iteration, say $p_q = |x_q^-|^2$. Both $\mathbf{R}_N$ and x_q are calculated iteratively until practical convergence. For Data 1 and Data 2, $N = 2\,F_s$ and $Q = 2\,N$. Fast implementations of IAA, referred to as F-IAA, were proposed in [7, 8, 19], which exploit the Gohberg-Semencul factorization of the inverse covariance matrix, building on the Hermitian Toeplitz structure of the covariance matrix and resorting to fast Fourier transforms. Such fast implementations were applied here to reduce the extremely high computational requirements of IAA.

ENF estimation can be cast as a line spectrum estimation problem. Accordingly, one may choose a suitable parametric method for solving the just described problem, such as ESPRIT, as was done in [15]. In particular, one has to choose the size m of the biased $m \times m$ auto-covariance estimate $\hat{\mathbf{R}}$ and the number of frequency samples Q to be estimated. Let $\mathbf{I}_{m-1}$ denote the identity matrix of size $(m-1)\times(m-1)$. The frequencies $\left\{\omega_q\right\}_{q=1}^{Q}$ are estimated as $-\arg(\hat{v}_q)$, where $\left\{\hat{v}_q\right\}_{q=1}^{Q}$ are the eigenvalues of the estimated matrix $\hat{\boldsymbol{\phi}}$ [18]:

$$\hat{\boldsymbol{\phi}} = (\hat{\mathbf{S}}_1^*\hat{\mathbf{S}}_1)^{-1}\,\hat{\mathbf{S}}_1^*\hat{\mathbf{S}}_2 \tag{4}$$

$$\hat{\mathbf{S}}_1 = [\mathbf{I}_{m-1}|\mathbf{0}]\,\hat{\mathbf{S}} \tag{5}$$

$$\hat{\mathbf{S}}_2 = [\mathbf{0}|\mathbf{I}_{m-1}]\,\hat{\mathbf{S}} \tag{6}$$

and $\hat{\mathbf{S}}$ is the matrix having as columns the Q principal eigenvectors of $\hat{\mathbf{R}}$. Here, $m = 4$ and $Q = 2$.

The MUSIC algorithm [18] is another suitable method for ENF extraction. First, the biased $m \times m$ auto-covariance estimate $\hat{\mathbf{R}}$ is computed. Next, the so-called "pseudospectrum" is estimated:

$$P_{MU} = \frac{1}{\mathbf{a}^*(\omega)\,\hat{\mathbf{G}}\,\hat{\mathbf{G}}^*\,\mathbf{a}(\omega)} \tag{7}$$

P_{MU} reveals which sinusoidal components are present in the signal. $\hat{\mathbf{G}}$ denotes the matrix made from the eigenvectors of $\hat{\mathbf{R}}$ spanning the subspace of noise. The values $m = 4$ and $Q = 2$ for Data 1 and Data 2 are used.

Having extracted the ENF, a matching procedure has to be performed against the ground truth information in order to identify the recording time. The ground truth ENF time series is downsampled by a factor of 10. Using the notation introduced in [15], let $\mathbf{f} = [f_1, f_2, \ldots, f_K]^T$ be the extracted ENF signal, which comprises the ENF estimated at each second. Let also $\mathbf{g} = \left[g_1, g_2, \ldots, g_{\tilde{K}}\right]^T$ for $\tilde{K} > K$ be the reference ground truth ENF. In [2], the association is being done by minimizing the squared error between $\mathbf{f}$ and $\tilde{\mathbf{g}}(l) = [g_l, g_{l+1}, \ldots, g_{l+K-1}]^T$, i.e.,

$$l_{opt} = \underset{l=1}{\overset{\tilde{K}-K+1}{\operatorname{argmin}}} \|\mathbf{f} - \tilde{\mathbf{g}}(l)\|_2^2 \tag{8}$$

An alternative matching criterion, proposed in [12], is the correlation matching, i.e.,

$$l_{opt} = \underset{l=1}{\overset{\tilde{K}-K+1}{\operatorname{argmax}}} c(l) \tag{9}$$

where $c(l)$ is the sample correlation coefficient between $\mathbf{f}$ and $\tilde{\mathbf{g}}(l)$ defined as:

$$c(l) = \frac{\mathbf{f}^T\tilde{\mathbf{g}}(l)}{\|\mathbf{f}\|_2\,\|\tilde{\mathbf{g}}(l)\|_2} \tag{10}$$

To measure the accuracy of ENF extraction by various algorithms, one may employ the maximum correlation coefficient $c(l_{opt})$. Alternatively, one may employ the standard deviation of the error between the true ENF and the estimated one. The former figure of merit was found to be more accurate than the latter one [12].

Another method for matching the extracted ENF to the ground truth is the DTW. Let $d(k,l) = (f_k - g_l)^2$. To align the two time series $\mathbf{f}$ and $\mathbf{g}$, a warping path is found very efficiently using dynamic programming and enforcing proper constraints to evaluate the recurrence which defines the cumulative distance $\gamma(k,l)$, for $k = 2, \ldots, \tilde{K}$:

$$\gamma(k,l) = \begin{cases} d(k,l) + \gamma(k-1,l) & \text{for } l = 1 \\ d(k,l) + \min\big\{\gamma(k,l-1), & \\ \quad \gamma(k-1,l-1), \gamma(k-1,l)\big\} & \text{for } l = 2, \ldots, \tilde{K} \end{cases} \tag{11}$$

assuming that

$$\gamma(1,l) = \begin{cases} d(1,l) & \text{for } l = 1 \\ d(1,l) + \gamma(1,l-1) & \text{for } l = 2,\ldots,\tilde{K}. \end{cases} \quad (12)$$

Here, the modified version of the original DTW with novel optimizations introduced in [16] is used.

4 EXPERIMENTAL EVALUATION

First the ENF extraction methods are tested on Data 1 and Data 2 using the two choices of frame length L, namely L_1 as in [15] and L_2 as shown in Table 1. The first choice allows comparisons with the results disclosed in [15]. The second one is used for studying the behavior of ENF extraction methods, when longer frames are used. For the F-IAA implementation, a 2 *sec* frame length is used due to high time requirements that arise from repeatedly matrix inversions in the iterative process.

Table 2: Maximum correlation coefficient for various methods applied to Data 1 with frame length L_1

Algorithm	60 Hz	120 Hz	180 Hz
STFT	0.9886	0.985	0.9957
Welch	0.9983	0.985	0.9983
Blackman-Tukey	0.9924	0.985	0.9978
Daniell	0.9906	0.985	0.9977
Capon	0.9969	0.9909	0.9972
F-IAA	0.9571	0.9784	0.964
ESPRIT	0.9979	0.9913	0.9979
MUSIC	0.9979	0.9913	0.9979

For Data 1 using the frame length L_1, the accuracy between the ENF signal extracted by various spectral analysis methods and the ground truth, measured by frequency disturbance recorders with accuracy up to about $\simeq 0.0005\,Hz$ [15], is summarized in Tables 2 and 3. The maximum correlation coefficient is listed in Table 2, while the minimum standard deviation of error is gathered in Table 3. It is seen that the first harmonic and the third one is more accurately estimated than the second one. For the first and third harmonics, the Welch method yields the best performance with respect to both figures of merit. Compared to [15], the accuracy of the ESPRIT method applied to Data 1 is increased from 0.947 to 0.9913 for the second harmonic, which is the weakest. Similarly, the standard deviation of error is reduced from $6.57 \cdot 10^{-3}$ to $2.901 \cdot 10^{-3}$. The most effective method proposed in [15] is STFT (Tracking), which uses a discrete dynamic programming approach. The maximum correlation coefficient of this method applied to Data 1 is 0.9968 for the third harmonic and the standard deviation of error is $1.851 \cdot 10^{-3}$. With respect to both figures of merit, the methods discussed here outperform STFT (Tracking). Using DTW, the minimum cumulative distance between the aligned time series $\mathbf{f}$ and $\mathbf{g}$ is obtained when the Welch method is used, as shown in Table 4. Matching the extracted ENF to the ground truth with DTW is reliable even for the second harmonic, which is the weakest.

By employing the frame length L_2, one expects more fine spectral resolution at the cost of lower time resolution. This is evident in Table 5 for Data 1. Frame-by-frame, ENF extraction by employing STFT with frame length L_2 followed by quadratic interpolation

Table 3: Minimum standard deviation of error for methods applied to Data 1 with frame length L_1

Algorithm	60 Hz	120 Hz	180 Hz
STFT	$2.806 \cdot 10^{-3}$	$3.202 \cdot 10^{-3}$	$1.303 \cdot 10^{-3}$
Welch	$1.069 \cdot 10^{-3}$	$3.202 \cdot 10^{-3}$	$1.069 \cdot 10^{-3}$
Blackman-Tukey	$2.284 \cdot 10^{-3}$	$3.202 \cdot 10^{-3}$	$1.218 \cdot 10^{-3}$
Daniell	$2.542 \cdot 10^{-3}$	$3.41 \cdot 10^{-3}$	$1.245 \cdot 10^{-3}$
Capon	$1.445 \cdot 10^{-3}$	$2.659 \cdot 10^{-3}$	$1.395 \cdot 10^{-3}$
F-IAA	$1.491 \cdot 10^{-2}$	$9.263 \cdot 10^{-3}$	$8.523 \cdot 10^{-3}$
ESPRIT	$1.198 \cdot 10^{-3}$	$2.901 \cdot 10^{-3}$	$1.202 \cdot 10^{-3}$
MUSIC	$1.198 \cdot 10^{-3}$	$2.901 \cdot 10^{-3}$	$1.208 \cdot 10^{-3}$

Table 4: Cumulative distance between the aligned time series f and g found by DTW for various methods applied to Data 1 with frame length L_1

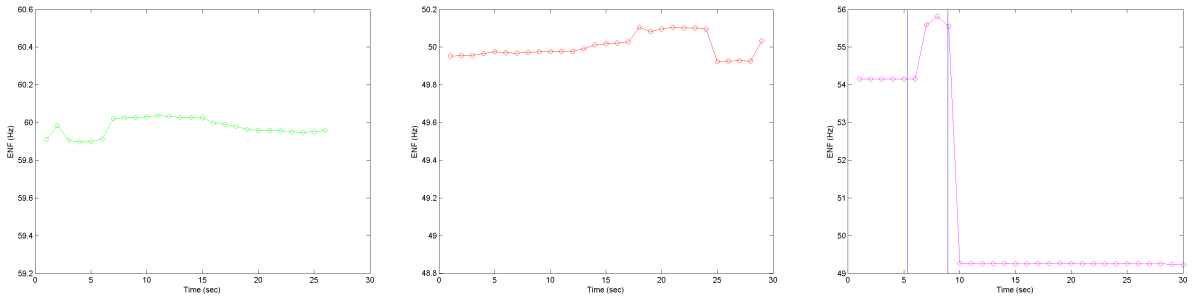
Algorithm	60 Hz	120 Hz	180 Hz
STFT	5.63429	6.05339	2.37785
Welch	1.98492	3.02637	1.98434
Blackman-Tukey	2.00499	6.7362	1.99956
Daniell	4.98134	6.33208	2.28147
Capon	2.2772	2.5221	2.2774
F-IAA	5.05359	4.11125	5.59216
ESPRIT	2.1139	2.4609	2.10688
MUSIC	2.11392	2.46089	2.11371

yields more accurate results with respect to both maximum correlation coefficient and minimum standard deviation of error than using L_1. Welch and Blackman-Tukey methods yield the best results in the second harmonic, although periodogram-based methods yield more accurate estimation in the first and third harmonics. The F-IAA accuracy for the second harmonic is 0.9784 and generally performs accurately although it employs a smaller frame length (2 *sec*).

Table 5: Maximum correlation coefficient for various methods applied to Data 1 with frame length L_2

Algorithm	60 Hz	120 Hz	180 Hz
STFT	0.9916	0.992	0.9964
Welch	0.9964	0.992	0.9965
Blackman-Tukey	0.9933	0.992	0.9967
Daniell	0.9926	0.9915	0.9967
Capon	0.9945	0.9913	0.9948
ESPRIT	0.9953	0.9916	0.9953
MUSIC	0.9953	0.9917	0.9953

Next, we proceed to ENF estimation from *Data* 2. In the speech recording, the first and third harmonics of the ENF are too weak [15]. Accordingly, we confine ourselves to the second harmonic ($120Hz$). Table 6 summarizes the findings for maximum correlation coefficient. It is seen that Capon method yields the most accurate results. The maximum correlation coefficient reported here is greater than 0.8446 reported in [15] for the same length L_1. Using DTW,



(a) ENF estimation in US recording **(b) ENF estimation in EU recording** **(c) ENF estimation in the mixed recording**

Figure 1: ENF estimation in authentic and altered recordings.

the minimum cumulative distance between the aligned time series $\mathbf{f}$ and $\mathbf{g}$ is obtained when the MUSIC and ESPRIT methods are employed. That is, DTW succeeds to align the extracted ENF time series to the ground truth, when strong interferences do exist, as is the case of Data 2. The same performance ordering of ENF extraction methods is observed, when a longer frame length L_2 is employed. Periodogram-based methods are ranked top with respect to the minimum standard deviation of error.

Table 6: Maximum correlation coefficient for various methods applied to Data 2 for both frame lengths.

Algorithm	$120\,Hz$ (L_1)	$120\,Hz$ (L_2)
STFT	0.9179	0.9328
Welch	0.9179	0.9328
Blackman-Tukey	0.9179	0.9328
Daniell	0.9176	0.9311
Capon	0.9351	0.9458
F-IAA	0.9182	-
ESPRIT	0.9318	0.9444
MUSIC	0.9318	0.9444

In Table 7, the computation time of various methods for ENF extraction applied to Data 1 with frame length L_1 is listed. The Capon method is the most time consuming with 188 sec, while its fast implementation resorting to the Gohberg-Semencul factorization of the inverse covariance matrix requires only 25 sec. STFT requires the least time. Increasing frame length to L_2, the order of the most time-consuming algorithms remains the same as for L_1.

To assess the performance limits of the ENF estimation methods, the various ENF estimation methods are applied to the 3rd dataset, which comprises of the US recording, the EU recording, and the mixed recording. The ENF in the US recording is expected to be found in $60\,Hz$, while the European one in $50\,Hz$. The ENF embedded in the mixed recording is expected to be found in $50\,Hz$ apart from a 4 sec part, where ENF is expected to exhibit an abnormal peak far from 50 Hz due to the alteration applied to the EU utterance as explained in Section 2. ENF fluctuations $\Delta f \geq 150\,mHz$ are considered to be abnormal [10]. The Daniell method employing a 9.5 sec long frame length is found to be able to estimate correctly

Table 7: Computation time (in sec) of various ENF estimation methods applied to Data 1 with frame length $L1$

Algorithm	$60\,Hz$	$120\,Hz$	$180\,Hz$
STFT	0.8836	0.9247	0.6713
Welch	7.8892	7.6732	7.4316
Blackman-Tukey	2.7167	2.5778	2.9981
Daniell	1.1282	1.9216	1.7166
Capon	188.2667	97.8226	97.2201
F-Capon	25.8116	6.5003	6.4371
ESPRIT	51.1890	51.7725	51.5155
MUSIC	51.0019	51.4458	51.4412

the ENF in the authentic US and EU recordings and to detect the alteration occurred in the mixed recording as can be seen in Figure 1. The vertical lines in Figure 1c indicate the starting and end time of the alteration.

5 CONCLUSION AND FUTURE WORK

Digital audio authentication requires high accuracy in ENF extraction in order to yield exact time/location estimation. Several frequency estimation methods have been tested on a frame-based approach by dividing the entire sequence into consecutive overlapping frames. It has been demonstrated by experiments that if the raw datasets are filtered by a properly designed band-pass filter, taking into account band-pass edges and filter order depending on the nature of the recordings, then either non-parametric or parametric techniques for spectral estimation provide an accurate estimation of the ENF. Certain challenges emerge. The experiments with the mixed recording have shown that the choice of the spectral estimation method is not trivial. Interferences may hinder ENF estimation. Exploiting the sparse nature of interferences in the formulation of ENF estimation could be a topic of future research.

REFERENCES

[1] D. Bykhovsky and A. Cohen. 2013. Electrical network frequency (ENF) maximum-likelihood estimation via a multitone harmonic model. *IEEE Trans. Information Forensics and Security* 8, 5 (May 2013), 744–753.

[2] A. J. Cooper. 2008. The electric network frequency (ENF) as an aid to authenticating forensic digital audio recordings. An automated approach. In *Proc. 33rd AES Int. Conf. Audio Forensics-Theory and Practice.*

[3] R. Garg, A. Hajj-Ahmad, and M. Wu. 2013. Geo-location estimation from Electrical Network Frequency signals. In *Proc. 2013 IEEE Int. Conf. Audio, Speech, and Signal Processing*. 2862–2866.

[4] R. Garg, A. L. Varna, and M. Wu. 2011. Seeing ENF: natural time stamp for digital video via optical sensing and signal processing. In *Proc. 19th ACM Int. Conf. Multimedia*. 23–32.

[5] R. Garg, A. L. Varna, and M. Wu. 2012. Modeling and analysis of electric network frequency signal for timestamp verification. In *Proc. 2012 IEEE Int. Workshop Information Forensics and Security*. 67–72.

[6] J. Garofolo. 1988. *Getting started with the DARPA TIMIT cd-rom: An acoustic phonetic continuous speech database.* Technical Report. National Inst. Standards and Technology (NIST).

[7] G. O. Glentis and A. Jakobsson. 2011. Efficient implementation of iterative adaptive approach spectral estimation techniques. *IEEE Trans. Signal Processing* 59, 9 (Sept. 2011), 4154–4167.

[8] G. O. Glentis and A. Jakobsson. 2011. Time-recursive IAA spectral estimation. *IEEE Signal Processing Letters* 18, 2 (Feb. 2011), 111–114.

[9] C. Grigoras. 2005. Digital audio recording analysis: the electric network frequency criterion. *Int. Journal Speech, Language, and the Law* 12, 1 (June 2005), 63–76.

[10] C. Grigoras. 2007. Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis. *Forensic Science Int.* 167, 2 (April 2007), 136 – 145.

[11] A. Hajj-Ahmad, R. Garg, and M. Wu. 2012. Instantaneous frequency estimation and localization for ENF signals. In *Proc. 2012 Asia-Pasific Signal and Information Processing Association Annual Summit and Conf.* 1–10.

[12] M. Huijbregtse and Z. Geradts. 2009. Using the ENF criterion for determining the time of recording of short digital audio recordings. In *Proc. Int. Workshop Computational Forensics*. 116–124.

[13] C. Kotropoulos and S. Samaras. 2014. Mobile phone brand and model identification using recorded speech signals. In *Proc. 19th Int. Conf. Digital Signal Processing*. Hong Kong, 586–591.

[14] D. P. Nicolalde-Rodriguez, J. A. Apolinario, and L. W. P. Biscainho. 2013. Audio authenticity based on the discontinuity of ENF higher harmonics. In *Proc. 21st European Signal Processing Conf.* 1–5.

[15] O. Ojowu, J. Karlsson, J. Li, and Y. Liu. 2012. ENF extraction from digital recordings using adaptive techniques and frequency tracking. *IEEE Trans. Information Forensics and Security* 7, 4 (Aug. 2012), 1330–1338.

[16] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh. 2012. Searching and mining trillions of time series subsequences under dynamic time warping. In *Proc. 18th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*. 262–270.

[17] J. O. Smith and X. Serra. 1987. *PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation.* CCRMA, Department of Music, Stanford University.

[18] P. Stoica and R. L. Moses. 2005. *Spectral Analysis of Signals.* Upper Saddle River, NJ: Pearson Prentice Hall.

[19] M. Xue, L. Xu, and J. Li. 2011. IAA spectral estimation: Fast implementations using the Gohberg-Semencul factorization. *IEEE Trans. Signal Processing* 59, 7 (July 2011), 3251–3261.

[20] T. Yardibi, J. Li, P. Stoica, M. Xue, and A. B. Baggeroer. 2010. Source localization and sensing: A nonparametric iterative adaptive approach based on weighted least squares. *IEEE Trans. Aerospace Electronic Systems* 46, 1 (Jan. 2010), 425–443.