ENF Detection in Audio Recordings via Multi-Harmonic Combining

Han Liao, Guang Hua, *Member, IEEE*, and Haijian Zhang, *Member, IEEE*

***Abstract*—The detection of the electric network frequency (ENF) in digital recordings is an essential step before the subsequent ENF extraction and forensic analysis. In this letter, we extend the state-of-the-art single-tone time-frequency (TF) domain ENF detector to the multi-tone scenario and propose a multi-harmonic combining (MHC) method, exploiting ENF harmonic components for improved detection performance. To exclude the corrupted components interfering rather than contributing to ENF detection, the proposed detector first performs a pre-screening based on the estimated average subband signal-to-noise ratios (SNRs) to exclude interfering components. Then, with the selected harmonic candidates, a second screening process is applied based on the TF test statistics (TSs), i.e., the variances of observed subband traces. After that, the multi-harmonic components are combined to form the final TS, whose sign determines the final decision. The advantages of the proposed method are illustrated via both synthetic analysis and real-world experimental results using the ENF-WHU dataset.**

***Index Terms*—Electric network frequency (ENF), ENF detection, harmonic signal detection, multimedia forensics.**

I. Introduction

OVER the past decade, the electric network frequency (ENF) criterion has been extensively researched and engineered to authenticate digital media content including audio [1], video [2], as well as image [3]. Recent researches have focused on robust ENF processing [4], [5] and using deep learning techniques for forensic ENF analysis [6]. However, compared to the captured multimedia content, the capture of the ENF in multimedia files is not guaranteed because of the requirements of nearby electric activity and the weakness of the physical forms of ENF [7]. Therefore, prior to ENF-based forensic analysis, it is of high importance to ensure the existence of such a power signature in the questioned files.

Researchers and practitioners have extensively investigated the existence of the ENF in multimedia content. In particular, Chai *et al.* [8] studied the quality of the ENF captured in battery-powered devices, Hajj-Ahmad *et al.* [7] investigated the factors affecting ENF capture in audio recordings, while Vatansever *et al.* [9] addressed ENF detection in video recordings. For the detection of the ENF in the most commonly considered audio recordings, it is usually formulated as a binary hypothesis testing problem [10]–[12]. Noticeably, in [12] this problem has been comprehensively addressed considering both theoretical and practical detectors, and a time-frequency (TF) domain detector has been proposed and shown to be more effective than the likelihood-ratio test (LRT)-based ones.

It has been verified that the captured acoustic hum generated by surrounding electric activities can contain multiple harmonic components, and this fact has been utilized for multi-harmonic model based ENF estimation [10], [13]. However, multi-harmonic condition has not been considered for ENF detection. In this letter, to achieve improved ENF detection performance, we extend the state-of-the-art TF domain single-tone ENF detector [12] to the multi-tone scenario and propose a multi-harmonic combining (MHC) method. It first excludes corrupted harmonics based on an improved subband signal-to-noise ratio (SNR) estimator, followed by a sigmoid mapping function and the subsequent MHC mechanism. To ensure practical merits, only the length of testing recording is considered known in this letter. The proposed detector not only tells whether a questioned file contains reliable ENF traces, but also serves to indicate the overall ENF quality and number of available harmonic components.

II. The Proposed Detector

The flowchart of the proposed detector is depicted in Fig. 1. The testing audio signal is first preprocessed by downsampling and bandpass filtering with a comb filter. Then, the harmonic components are respectively processed before being fed into the MHC mechanism to generate the final test statistic (TS).

A. Harmonic Subband TF Detector

After downsampling and bandpass filtering, we model the multi-tone harmonic ENF detection problem as the following binary hypothesis testing

$$\begin{aligned} \mathcal{H}_0 &: x[n] = v[n], \\ \mathcal{H}_1 &: x[n] = s[n] + v[n], \end{aligned} \tag{1}$$

where $n \in \{0, 1, \ldots, N-1\}$, N is the length of signal, and $s[n]$ is the time-domain representation of the ENF signal and $v[n]$ is the bandpass filtered white Gaussian noise (WGN). For notation simplicity, the SNR is defined by the ratio of signal energy and noise energy before bandpass filtering which is not explicitly presented here. We now extend the single-tone model of $s[n]$ to

Manuscript received July 21, 2021; revised August 29, 2021; accepted August 30, 2021. Date of publication September 2, 2021; date of current version September 22, 2021. This work was supported by the National Natural Science Foundation of China under Grant 61802284. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Beibei Wang. *(Corresponding author: Guang Hua.)*

The authors are with the School of Electronic Information, Wuhan University, Wuhan 430072, China (e-mail: liaohan@whu.edu.cn; ghua@whu.edu.cn; haijian.zhang@whu.edu.cn).

Digital Object Identifier 10.1109/LSP.2021.3109773

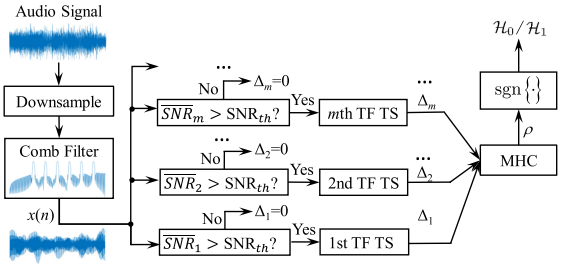

Fig. 1. The flowchart of the proposed detector.

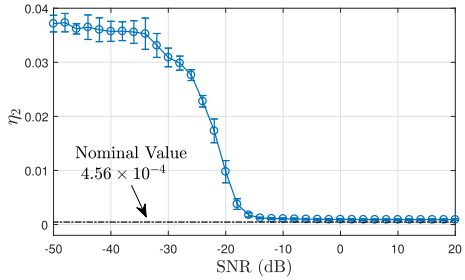

Fig. 2. Statistics of optimal η_m versus SNR, normalized to the 2nd harmonic.

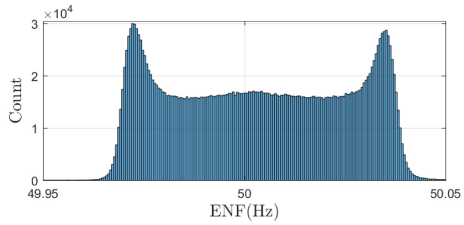

Fig. 3. Statistics of ENF reference signal in one week.

the multi-tone scenario, which is given by

$$s[n] = \sum_{m\in\mathcal{M}} A_m[n] \cos\left(2\pi T \sum_{i=0}^{n} m f[i] + \phi_m\right), \quad (2)$$

where $A_m[n] > 0$ and ϕ_m are the unknown time-varying amplitudes and initial phase of the mth harmonic, respectively, $f[n]$ is the ENF, T is the sampling interval, and $\mathcal{M}$ is the harmonic index set. Assuming frame-based processing and for each frame the underlying ENF is a constant [1], [12], [14], according to the linear relationships among harmonics, and based on the single-tone result in [12], the mth TF TS, denoted by T_m, is then given by

$$T_m = \frac{m^2}{L-1} \sum_{l=0}^{L-1} \left(f[l] - \frac{1}{L}\sum_{l=0}^{L-1} f[l]\right)^2, \quad (3)$$

where L is the number of frames, $l \in \{0, 1, \ldots, L-1\}$, and $f[l]$ is the estimate of the ENF at frame l. The subband detector determines $\mathcal{H}_1$ if $T_m < \eta_m$, where η_m is the TS threshold. The TS in (3) is in fact the estimate of in-band sample variance per subband, which is more effective than matched filter or LRT-based detectors, because it exploits the unique slowly-varying nature of the ENF [12]. Further, these detectors are constant false alarm rate (CFAR) ones since the noise statistics under $\mathcal{H}_0$ are known. Therefore, the theoretical subband thresholds could be obtained by extending the single-tone solution to the harmonic scale m, i.e.,

$$\begin{aligned}\eta_m &= E\{T_m; \mathcal{H}_0\} - \beta\sqrt{\mathrm{var}\{T_m; \mathcal{H}_0\}} \\ &\approx (B_m/6)^2 - \beta\sqrt{\frac{2(B_m/6)^4}{L-1}}, \end{aligned} \quad (4)$$

where β is a weight factor, B_m is the bandwidth of the mth passband of the comb filter. The above result is obtained by noting that the distributions of (3) under both hypotheses follow two gamma distributions, respectively, and the detailed derivations can be found in [12]. It can be seen that the theoretical thresholds are essentially determined by the corresponding passband width. When the passbands are linear scaled by the harmonic index depicted in Fig. 1, $B_m = mB_1$.

The above theoretical results are asymptotically optimal and are valid under the assumption of additive WGN. However, the two conditions are usually difficult to meet in real-world scenarios. Actually, the observation time in real-world situations, including the duration of the recording and the analysis frame size, are relatively short, and the noise is strong and non-stationary. Therefore, we empirically search for the optimal TS threshold η_m via Monte-Carlo experiments. Specifically, we synthesize ENF signals using the first-order autoregressive (AR) model [15], [16] with model coefficient set to 0.99, and the variance of the synthetic clean ENF is normalized to the measured value 4.56×10^{-4} Hz^2 in Central China grid. The lengths of synthetic data are set to $[20, 280]$ seconds, and we create pseudo random white noise with different SNR values within $[-50, 20]$ dB to generate noisy observations. The statistics of optimal η_m, evaluated at the normalized 2nd harmonic subband and averaged from 100 random realizations, are presented in Fig. 2, in which η_2 is optimal if it minimizes the detection error, and the vertical bars represent the plus and minus standard deviation of η_2 with respect to the data length. It can be seen from the very short standard deviation bars that the optimal thresholds are much less sensitive to the length of recording than to the SNR. The asymptotic property of the detector is observed from the decreasing standard deviations.

B. Multi-Harmonic Combining (MHC) Scheme

Before deriving the proposed the MHC detector, we analyze the SNR at each harmonic subband because when performing the short-time Fourier transform within a frame size, a frequency bound is set according to the statistical characteristics of the reference signal as shown in Fig. 3, in which the nominal ENF of Central China Grid is centered at 50 Hz. If the TF detector is used for ENF detection and the SNR is low, this bound may cause zero variance, which will lead to higher detection errors. Therefore, we propose to use the average SNR estimate at the mth harmonic, denoted by $\overline{\mathrm{SNR}}_m$, to choose the available harmonic components before further processing. Inspired by the local SNR (rough) estimator presented in [13] and [14], the estimated SNR of the lth frame at the mth harmonic is given

by

$$\mathrm{SNR}_{m,l} = 10\lg \frac{\|\mathrm{FFT}\left[s_{m,l}[n]\right]\|_2^2}{\|\mathrm{FFT}\left[x_{m,l}[n]\right]\|_2^2 - \|\mathrm{FFT}\left[s_{m,l}[n]\right]\|_2^2}, \quad (5)$$

where $\|\mathrm{FFT}[s_{m,l}[n]]\|_2^2$ is the desired power in the passband of $m \times [49.96, 50.04]$ Hz from Fig. 3, and $\|\mathrm{FFT}[x_{m,l}[n]]\|_2^2$ is the total power of the lth frame at the mth harmonic in $m \times [49.5, 50.5]$ Hz, which depends on the designed comb filter. Then the SNR at the mth harmonic is given by

$$\overline{\mathrm{SNR}_m} = \frac{1}{L}\sum_{l=0}^{L-1} \mathrm{SNR}_{m,l}. \quad (6)$$

Therefore, we will use (6) to screen the available harmonics, and the screen criterion will be introduced in section III. Then, we stick to the convention of the 2nd harmonic normalization and denote the normalized difference between the TF domain TS and threshold at the mth harmonic as

$$\Delta_m = 2(\eta_m - T_m)/m, \quad (7)$$

which indicates that the subband detectors determine $\mathcal{H}_1$ on the condition that $\Delta_m > 0$ and at the same time the $\overline{\mathrm{SNR}_m}$ screening process is passed, otherwise, we set $\Delta_m = 0$. One intuitive strategy for subband detection is to choose the minimal variance (MV) of all harmonics as the TS, however, the variation of ENF is tiny for short recordings (e.g., less than 2 mins as we demonstrate later), which leads to wrong decisions. To make a better use of the harmonic subbands, the following MHC detector is proposed.

From (7), the single-tone model based TF detector [12] could be equivalently written as $\Delta_2 > 0$. We extend (7) to multi-tone scenario and it is further mapped into a probability-like quantity via this simple and common function as follows,

$$\rho_m = \mathrm{sigmoid}\left\{\alpha \times \Delta_m\right\}. \quad (8)$$

The use of the sigmoid function here serves as a normalization process that turns the values of subband TSs into probability measurements, which can then be more effectively combined. Due to the tiny dynamic range of the ENF, Δ_m is usually small with a magnitude of 10^{-4} that will cause $\rho_m = 0.5$ for each harmonic, which is meaningless for the detector. Therefore, the coefficient α is used to proportionally amplify Δ_m to overcome this weakness, and so that α could be set in the range of 10^4. Based on the previous analysis, we have $\Delta_m \geq 0$, and consequently, regardless of whether ENF exists or not, $0.5 \leq \rho_m < 1$. If ENF does not exist, $\rho_m = 0.5$ for all harmonics, but if there are harmonics that satisfy both the variance and $\overline{\mathrm{SNR}_m}$ requirements, $0.5 < \rho_m < 1$.

Before the multi-harmonic combining is utilized to obtain the final TS ρ, a coefficient dependent on ρ_m is formed as

$$\omega_m = \begin{cases} 0, & \text{if } \rho_m = 0.5, \\ \frac{\rho_m}{\sum_{m\in\mathcal{M}} \rho_m}, & \text{if } \rho_m > 0.5. \end{cases} \quad (9)$$

Using (9), the proposed final TS is then given by

$$\rho = \sum_{m\in\mathcal{M}} \omega_m, \quad (10)$$

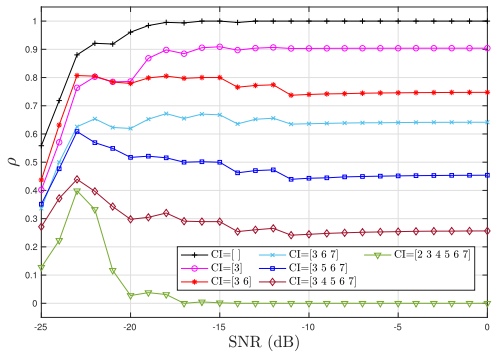

Fig. 4. Relationship between ρ and SNR using synthetic signals.

where $\rho \in [0, 1]$, and it is the TS for the MHC scheme. The detector will determine $\mathcal{H}_1$ if $\rho > 0$, otherwise $\mathcal{H}_0$. In addition, the value of (10) may also be used to determine the number of available harmonics for further ENF forensics when more than one harmonic component can be used. If $\mathcal{H}_1$ holds, then $\rho \in [\frac{[\rho_m]_{\min}\times r}{[\rho_m]_{\min}\times r+(R-r)\times 0.5}, \frac{[\rho_m]_{\max}\times r}{[\rho_m]_{\max}\times r+(R-r)\times 0.5}]$, where r and R are the number of available harmonics and harmonics need to be analyzed respectively, and $R = \frac{f_S 2}{50} - 2$, $r \leq R$. $[\rho_m]_{\min}$ and $[\rho_m]_{\max}$ could be 0.51 and 1 respectively to calculate the theoretical range of ρ under different r. Before using synthetic signals for further analysis, the SNR versus ρ is discussed firstly. Equation (2) is used to generate a 300-second synthetic signal, the SNR is $[-25, 0]$ dB, and we set $f_S = 800$ Hz, then the harmonic index is $\mathcal{M} = \{2, 3, \ldots, 7\}$, and the harmonic corrupted index (CI) is the subset of $\mathcal{M}$.

The Monte-Carlo experiments are shown in Fig. 4. It can be seen that when $\mathrm{SNR} < -20$ dB, it is impossible to accurately determine the number of available harmonics for ENF analysis by ρ, but it will make a more accurate decision if $\mathrm{SNR} \geq -20$ dB. The greater the value of ρ, the greater possibility that the ENF exists, and it means that ρ can reflect the quality of the ENF signal in an audio recording to a certain extent. In the following content, we evaluate the performance of the proposed MHC detector, and to ensure the practical merits of the proposed detector, we only consider the length of recording as the known parameter.

III. Empirical Optimization and Results

A. Synthetic Analysis

Since the TF subband detectors are CFAR detectors, thanks to the signal-independent TSs under $\mathcal{H}_0$, the per subband threshold η_m could be reliably selected. According to the statistical results shown in Fig. 2, we can set $\eta_2 = 0.01$ when $\mathrm{SNR} = -20$ dB, and the $\overline{\mathrm{SNR}_m}$ constraint is $\overline{\mathrm{SNR}_m} > -20$ dB. Detection accuracy (ACC) is used as the performance metric, which is given by $\mathrm{ACC} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$, where TP (True Positive) and TN (True Negative) mean the detector makes the right decision, and FP (False Positive) and FN (False Negative) mean the wrong decision. We compare the detection accuracy of the proposed ENF detector using synthetic data, and the simulation results are shown in Fig. 5.

We can observe from Fig. 5 that the detection accuracy of TF [12], LS-LRT [12], MV, and MHC detectors are close to

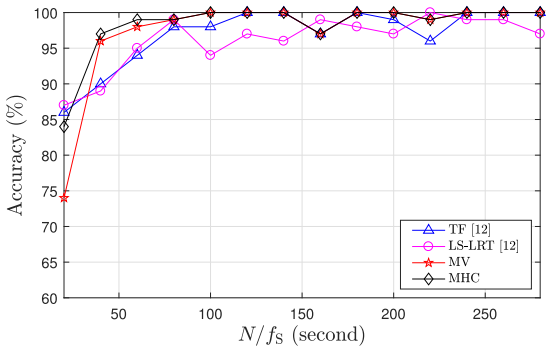

Fig. 5. Comparison of ENF detection accuracy using synthetic signals.

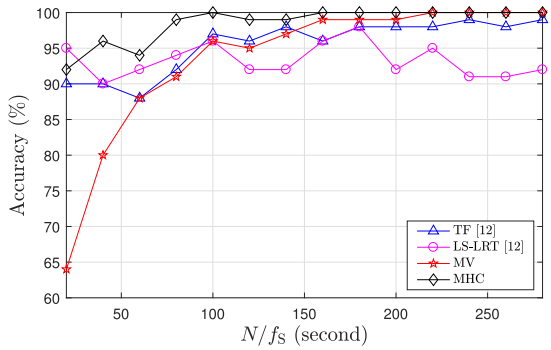

Fig. 6. Comparison of ENF detection accuracy using real-world recordings.

TABLE I
F1 SCORE FOR SYNTHETIC SIGNALS AND REAL-WORLD RECORDINGS

Scheme	Dur. (s)	Synthetic		Real-world	
		ACC (%)	F1 score	ACC (%)	F1 score
TF [12]	20	86	0.8600	90	0.9057
WMLE [13]		56	0.6901	63	0.7218
LS-LRT [12]		87	0.8506	95	0.9495
MV		74	0.7903	64	0.7353
MHC		84	0.8571	92	0.9259
TF [12]	40	90	0.8936	90	0.8958
WMLE [13]		62	0.7361	69	0.7559
LS-LRT [12]		89	0.8791	90	0.8980
MV		96	0.9608	80	0.8333
MHC		97	0.9703	96	0.9615
TF [12]	60	94	0.9455	88	0.8800
WMLE [13]		60	0.7222	62	0.7077
LS-LRT [12]		95	0.9550	92	0.9216
MV		98	0.9825	88	0.8929
MHC		99	0.9912	94	0.9434

or greater than 95% if the duration is longer than 60 seconds. When the signal length is 20 seconds, the accuracy of MHC and MV detectors are slightly lower than TF and LS-LRT. This is because when the duration is short, the ENF loses randomness and a bound is set according to Fig. 3, which may result in small or even constant variance which will raise FPs. For the MHC detector, with the use of the $\overline{\mathrm{SNR}_m}$ criterion to reduce FP, its accuracy is always better than MV. Meanwhile, when the duration is longer than 20 seconds, the advantages of MV and MHC detector are observed.

To further illustrate the influence of the data length on detection performance, we examine the F1 scores, and the results are shown in Table I. The performance advantages of the MHC detector is seen to be more significant for longer durations. Results of the weighted maximum-likelihood estimator (WMLE) [13] are also presented for comparison. Since the corrupted harmonic components have been used in te WMLE, it suffers from a clear performance drop. We also observe from the table that when the duration is 20 seconds, though the accuracy of LS-LRT detector is slightly higher than MHC, its F1 score is slightly lower.

B. Analysis Using Real-World Data

The recordings in ENF-WHU dataset are cropped at random locations to construct a larger dataset with duration values varying from 20 to 280 seconds, and for each duration, we randomly crop 50 clips for $\mathcal{H}_0$ and 50 for $\mathcal{H}_1$, respectively. For real-world recordings, the actual recording situations are more complicated than those in theoretical analysis. Therefore, combining the synthetic analysis of $\mathrm{SNR} = -20$ dB with the actual application scenario, we set more stringent constraints as $\overline{\mathrm{SNR}_m} > -15$ dB and $\eta_2 = 0.008$ to reduce FPs, and the results are provided in Fig. 6. In this figure, all the presented detectors are practical ones, and all of their detection accuracies can achieve 90% when the duration is longer than 80 seconds. It is further observed that the performance results of MHC detector is better than the competitors when the duration is over 20 seconds. To provide more insights, the ROC curves of the proposed MHC detector and three competitors are presented in Fig. 7 for 20-second recordings. Generally, from the results shown in Figs. 6, 7, and Table I, the advantages of the proposed method can be verified.

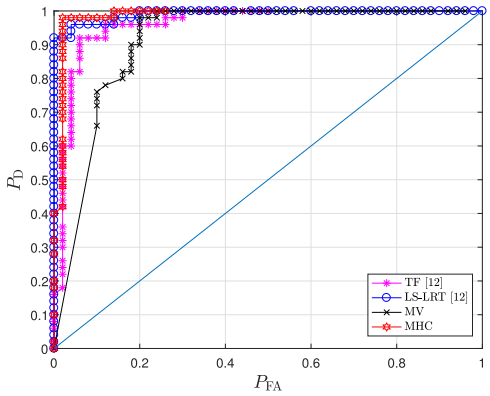

Fig. 7. ROC curves of detectors for 20-second recordings.

IV. CONCLUSION

A multi-tone model based ENF detection method is proposed in this letter. It first uses an averaged SNR criterion to screen available harmonic subbands, followed by the proposed MHC mechanism to generate the final detection TS function. The proposed multi-harmonic TS also serves as a measurement to determine the quality and number of usable ENF harmonics. Both synthetic and practical experiments have been carried out for performance evaluation, and it is found that the MHC detector could improve the detection accuracy over the previous single-tone based methods. This work ensures more reliable ENF detection in practical situations.

REFERENCES

[1] A. J. Cooper, "An automated approach to the electric network frequency (ENF) criterion-theory and practice," *Int. J. Speech Lang., Law*, vol. 16, no. 2, pp. 193–218, Apr. 2010.

[2] R. Garg, A. L. Varna, A. Hajj-Ahmad, and M. Wu, "'Seeing' ENF: Power signature based timestamp for digital multimedia via optical sensing and signal processing," *IEEE Trans. Inf. Forensics Secur.*, vol. 8, no. 9, pp. 1417–1432, Sep. 2013.

[3] C.-W. Wong, A. Hajj-Ahmad, and M. Wu, "Invisible geo-location signature in a single image," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 1987–1991.

[4] Q. Zhu, M. Chen, C.-W. Wong, and M. Wu, "Adaptive multi-trace carving for robust frequency tracking in forensic applications," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 1174–1189, 2021.

[5] G. Hua and H. Zhang, "ENF signal enhancement in audio recordings," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1868–1878, 2020.

[6] M. Mao, Z. Xiao, X. Kang, X. Li, and L. Xiao, "Electric network frequency based audio forensics using convolutional neural networks," in *Proc. Adv. Digit. Forensics XVI*, 2020, pp. 253–270.

[7] A. Hajj-Ahmad, C. Wong, S. Gambino, Q. Zhu, M. Yu, and M. Wu, "Factors affecting ENF capture in audio," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 2, pp. 277–288, Feb. 2019.

[8] J. Chai, F. Liu, Z. Yuan, R. W. Conners, and Y. Liu, "Source of ENF in battery-powered digital recordings," in *Proc. Audio Eng. Soc. Conv.* vol. 135, Oct. 2013, pp. 1–6.

[9] S. Vatansever, A. E. Dirik, and N. Memon, "Detecting the presence of ENF signal in digital videos: A superpixel-based approach," *IEEE Signal Process. Lett.*, vol. 24, no. 10, pp. 1463–1467, Oct. 2017.

[10] D. Bykhovsky and A. Cohen, "Electrical network frequency (ENF) maximum-likelihood estimation via a multitone harmonic model," *IEEE Trans. Inf. Forensics Secur.*, vol. 8, no. 5, pp. 744–753, May 2013.

[11] G. Hua, J. Goh, and V. L. L. Thing, "A dynamic matching algorithm for audio timestamp identification using the ENF criterion," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 7, pp. 1045–1055, Jul. 2014.

[12] G. Hua, H. Liao, Q. Wang, H. Zhang, and D. Ye, "Detection of electric network frequency in audio recordings-from theory to practical detectors," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 236–248, 2021.

[13] A. Hajj-Ahmad, R. Garg, and M. Wu, "Spectrum combining for ENF signal estimation," *IEEE Signal Process. Lett.*, vol. 20, no. 9, pp. 885–888, Sep. 2013.

[14] P. M. G. I. Reis, J. P. C. Lustosa da Costa, R. K. Miranda, and G. Del Galdo, "ESPRIT-Hilbert-based audio tampering detection with SVM classifier for forensic analysis via electrical network frequency," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 4, pp. 853–864, Apr. 2017.

[15] A. Hajj-Ahmad, R. Garg, and M. Wu, "Instantaneous frequency estimation and localization for ENF signals," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2012, pp. 1–10.

[16] R. Garg, A. L. Varna, and M. Wu, "Modeling and analysis of electric network frequency signal for timestamp verification," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, 2012, pp. 67–72.