A Hilbert-Based Approach to the ENF Extraction Problem

Andreas Triantafyllopoulos[a], Ioannis Krilis[b], Anastasios Foliadis[c] and Athanassios Skodras[d]
Electrical and Computer Engineering Department
University of Patras
Patras, Greece

[a]ece7937@upnet.gr [b]ece7818@upnet.gr [c]ece7960@upnet.gr [d]skodras@upatras.gr

***Abstract*— The estimation of location based on the time varying Electric Network Frequency (ENF) is a new emerging technology in Information Forensics. This requires the extraction of the ENF signal from multimedia recordings and a comparison with already known power grid signatures. In this paper, we focus on ENF signal extraction and statistical modelling of ENF signals. We introduce a novel technique based on instantaneous frequency estimation using the Hilbert transform, which shows promising results.**

I. Introduction

The Electric Network Frequency (ENF) analysis is used as a forensic science technique in order to estimate the location of an audio or video recording. The nominal value of ENF is either 50Hz or 60Hz, depending on the country of interest, with 60Hz used mostly in North America and Japan and 50Hz in most other countries. However, the ENF is not steady but fluctuates over time due to variations in the demand and supply of electric power. These variations present a generally consistent trend within the same grid [1]. ENF signals consist of changes in electric network frequency over time and can be extracted from recordings either directly from a power socket or using a portable audio recorder. These recordings are segmented and processed in order to produce an estimation of the frequency at each segment. It has been shown that signals extracted from video or audio recordings are similar to concurrently recorded clean power signals [2].

In this paper, (a) we implement two of the most commonly used techniques for ENF extraction from noisy, audio recordings, namely a Short-Time Fourier Transform (STFT) and a Spectrum Combining method [3], [4], and (b) we introduce a novel approach, based on the signal's Hilbert transform. Our Hilbert based approach differs from the other two methods in that it makes no implicit assumptions about underlying stationarity, whereas both of the other methods do process frames of the recorded signal under the assumption that the ENF signal is stationary within that frame. The Hilbert-based approach produces more accurate results even under lower signal-to-noise (SNR) conditions.

Applications using ENF signals include detection of tampering or modifications in a multimedia signal, time- and location-of-recording validation and region-of-recording identification [5]. As already mentioned, multimedia signals show similarities with power references from the same grid. Working under the assumption that variations of the ENF are consistent within the same grid and differ amongst distinct countries, statistical features can be extracted from ENF signals and a classification system to accurately identify their region-of-recording can be developed. This latter application is the one we focus mostly on, and our signal extraction techniques are qualitatively evaluated based on their performance in such a classification system.

Section II provides an overview of the approaches used to extract the ENF signals from our recordings and Section III details the classification method used and presents our results.

II. Extraction of ENF signals

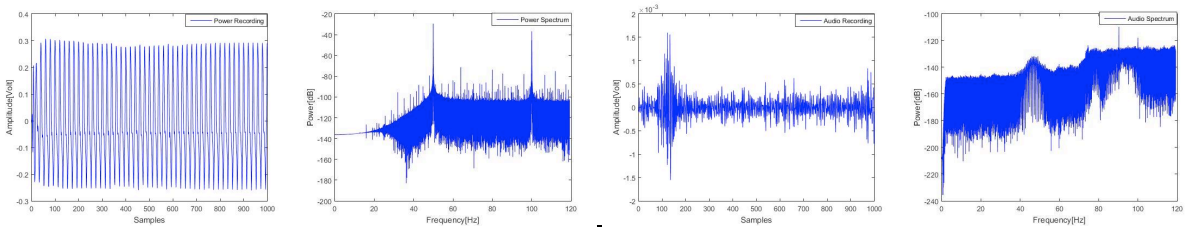
We worked on two different sets of recordings, namely some acquired using a sensing hardware and connected directly to a power socket, thus effectively measuring the power grid, and some acquired using a battery powered digital recorder. In the latter case the recorder is influenced by surrounding power sources and such recordings tend to be noisier, requiring different signal extraction approaches.

Fig. 1 shows two concurrent recordings, one from our own sensing hardware and one using a mobile phone as a recording device. Our initial observation is that audio recordings tend indeed to be noisier than the ones made with a power recorder, as is evident by both time and frequency domain plots. This justifies our choice of applying different ENF extraction methods for different noise conditions.

An ENF signal captured using a digital recorder, either battery powered or plugged into a power socket, can be modelled as:

$$x(n) = \sum_{k=0}^{N-1} a_k \cos(\omega n) + w(n) \qquad (1)$$

where $\omega = 2\pi f / f_s$, f is a random variable, denoting the instantaneous frequency of the power grid, f_s is the recorder's sampling frequency, N is the number of harmonics that fall under $f_s/2$ and can be captured by the recorder and α_k are the coefficients of each harmonic. These coefficients must also be considered random variables, as, in the case of a battery-powered recorder, they are directly linked to the electromagnetic interference created by the power grid, and, thus, dependent on the specific recorder's components and the overall physical structure of the underlying environment. Finally, $w(n)$ can be considered as Gaussian noise.



(a) (b) (c) (d)

Fig. 1. Concurrent (a) clean and (c) audio recordings and their respective spectra (b), (d).

The number N of harmonics, was chosen to be 9, because we considered the amount of power present in higher harmonics to be negligible. As a result, a sampling frequency of 1kHz was deemed sufficient. However, since most of the ENF signature's power is principally contained in the lower harmonics, one could also choose a different, lower sampling frequency.

In general, both f and α_k are non-stationary, as f depends on the load changes in the power grid as well as the available power supply mechanisms, and α_k can change over time, especially if the recorder is being moved around or there are moving objects in the surrounding environment. It is therefore desirable, to develop an ENF estimation method that can deal with the inherent non-stationarity of the data.

We proceed to describe four different methods for extracting ENF signals, a simple zero crossings algorithm that has been proven adequate for direct-from-power-socket recordings [5], an STFT approach used extensively in ENF related bibliography, a variation of the spectrum combining technique described in [4], and our own Hilbert based analysis.

A. Zero Crossings

Zero crossings is a method used in [5] that can produce a fast estimation of the frequency of a sinusoidal signal. The data are processed in frames of approximately 4s (4096 samples), where we consider the ENF to be stable. This approach was also used in [6], and could adequately capture the power grid signature. An estimation is produced for each frame, counting the number of zero crossings in that particular segment.

In our implementation we first estimate the central frequency of the signal and use passband filtering centered on this estimated frequency with a 10Hz margin. After that, the signal is divided into frames. For each frame, we find all subsequent samples that differ in sign, and calculate the time of crossing using linear interpolation. For two consecutive crossings, we compute the time difference between them and use it as an estimation of the signal's period. The ENF sample of that frame is the average of the computed frequencies across all crossings.

B. Short-Time Fourier Transform

We have seen from Fig. 1 that multimedia recordings are noisier than the respective recordings acquired directly from a power socket. As a result, the Zero Crossings method fails to extract the desired ENF signal. So, it is evident, that a different method of extraction needs to be applied in the case of audio recordings.

The most commonly used method to identify the fluctuations of a frequency varying signal is the Short-Time Fourier Transform (STFT). The signal is split into frames, and the Fourier Transform is computed for each one. Thus an ENF estimate can be produced for each frame. It is evident that, by computing the Fourier transform for each frame, we are making an implicit assumption that the ENF signal is stationary within its boundaries.

It is shown in [7], that in multimedia recordings, the ENF is also present in higher harmonics. In many recordings, it was observed that we could achieve a higher SNR in frequencies that differ from the nominal frequency. Therefore, we decided to focus our analysis on the frequency which presents the highest SNR, as we could obtain a better estimation of the ENF. To get an approximation of the SNR, we first compute the frequency with the highest power amplitude near the nominal harmonic frequencies (i.e. 50, 100, 150 Hz etc.) and consider this to be the actual harmonic of the power grid. We attribute the power of a 1Hz band around the computed harmonic to the ENF signal, and consider the rest of the 2Hz band around this harmonic to be noise.

Each frame is filtered using a passband filter centered around the estimated ENF of the previous frame, with a pass-band of 0.4Hz. The algorithm is initialized with the harmonic which presents the highest SNR. The STFT is applied to the filtered samples of the frame, and results in an estimated ENF value, which is the frequency with the highest power amplitude. Lastly, in order to increase the accuracy of the estimation, zero padding is applied in each frame, and the resulting frequency is computed through quadratic interpolation.

C. Spectrum Combining

In this section, we discuss a variation of the method used in [4] and produce an ENF estimation by combining base and harmonic spectral bands each with a corresponding weight. Our implementation differs from [4] in that we choose to use the chirp z-transform to approximate the maximum coefficient and the corresponding frequency.

The chirp z-transform is the z-transform of a signal along an arbitrary spinal contour. The contour used in each step of our algorithm is a segment of the unit circle, whose limits are specified by a range of 0.4 Hz around the frequency estimation computed in the previous step. This allows us to calculate the

coefficients inside the area of interest with increased accuracy. Our implementation consists of the following steps:

- Calculate the basic frequency of the recording.
- Calculate the SNR of each harmonic in the same manner as described in Section II.B., and use it as the combining weight. Only the two harmonics with the largest weights are taken into account for the next steps to decrease computation time.
- Segment the signal into frames of 4096 samples.
- Compute the chirp z-transform for each segment around the two harmonics with the largest weights, combine the resulting spectra, and use the frequency corresponding to the maximum value as the ENF estimation for that segment.

D. Hilbert Analysis

In this section, we describe an ENF estimation method based on the Hilbert transform or Hilbert transformer. For an arbitrary time-series $x(t)$, its Hilbert transform, $\hat{x}(t)$ is defined as

$$\hat{x}(t) = x(t) * \frac{1}{\pi t} = \frac{1}{\pi} PV \int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau} d\tau \qquad (2)$$

where * denotes the convolution and PV the Cauchy principal value. Using this notation we acquire the analytic signal

$$g(t) = x(t) + j\hat{x}(t) = r(t)\, e^{j\theta(t)} \qquad (3)$$

where $j = \sqrt{-1}$ and

$$r(t) = \sqrt{x^2(t) + \hat{x}^2(t)} \qquad (4)$$

$$\theta(t) = arctan(\frac{\hat{x}(t)}{x(t)}) \qquad (5)$$

This helps us define the instantaneous frequency of the signal as:

$$\omega(t) = \frac{d\theta(t)}{dt} \qquad (6)$$

The Hilbert transform is essentially defined as the convolution of $x(t)$ with $1/\pi t$; therefore, the local properties of $x(t)$ are emphasized. This quality makes the Hilbert transform ideal for dealing with non-stationary signals. We can thus proceed with our analysis without making any initial assumptions of either local or global stationarity.

The notion of instantaneous frequency is ambiguous, as we have mostly associated frequency with the Fourier transform, and specifically the ubiquitous presence of sinusoidal oscillations in our data. The instantaneous frequency defined in (6) is associated with each sample of the processed signal. From that end, the definition of instantaneous frequency seems to lack physical meaning, as we cannot define a period of oscillation using only a single sample. For it to make sense, it would have to be associated with a "monocomponent function", i.e. a function that at any given time has only a single component, and hence we could obtain a unique, well-defined frequency. This assumption places severe limitations in the form of the data we can manipulate. We can circumvent those by limiting our data to narrow band signals only. Even operating under this assumption though, we still obtain errors, for example negative frequencies, a problem also observed in [8].

However, we need not obtain an instantaneous frequency estimation for every sample of our recording. We can compute a single value for every frame of 4096 samples, where we consider the ENF to be stable, using the mean value of all calculated instantaneous frequencies which fall within the specified filter's range. In our application, we decided that using a band-pass filter with a 0.4Hz pass-band, would adequately filter out all unwanted frequency components. This is in accordance with the other methods previously discussed, but does not impose any stationarity restrictions on our data, as we combine the instantaneous frequency samples produced and, considering these results to be noisy estimations of the signal's true frequency, we adopt their mean value as an estimator of the ENF.

This approximation is expected to provide a good, albeit noisy, estimation of a signal's frequency. Since the global variance of ENF signal may exceed our specified band of 0.4Hz, it is impossible to filter the signal around the calculated central frequency, as this could perhaps filter out crucial spectral information. Instead, we choose to use an adaptive filtering scheme, processing each frame separately, and centering our filter on the frequency estimated for the previous frame. Again, as in the STFT method described in II.B, we initialize our algorithm using the harmonic which presents the highest SNR.

III. Experimental Data Analysis

A. Similarity between simultaneous recordings

The data used were collected with the sensing circuit developed for the purposes of the 2016 IEEE Signal Processing Competition[1]. The first measure used to evaluate our methods' performance was the degree of similarity between ENF signals extracted from concurrent recordings. One such metric is the correlation coefficient which quantifies the correlation and dependence between ENF signals. Using our recording device to establish the ground truth and an iPhone 4 as a battery-powered audio recorder to produce simultaneous recordings, we developed a database of clean and noisy ENF signals. Since the ENF signature is ubiquitous, we should be able to detect the same signal present in both recordings. We positioned the audio recorder in different environments, and different settings, to measure our methods' performance under various noise conditions.

We were able to produce better results using our Hilbert based method. Specifically, we obtained a mean value of

[1] http://www.icassp2016.org/SPCup.asp

correlation coefficient of 0.4975 compared to 0.3975 produced by the Spectrum Combining method and 0.2959 by STFT, with variances of 0.2359, 0.2210 and 0.2407, respectively. A number of 31 recordings were used, with total duration of approximately 20 hours. Excluding some recordings that showed a very low SNR, we obtained better results. Specifically, a mean score of 0.7627 for the Hilbert method, 0.5365 for Spectrum Combining and 0.4567 for STFT, with variances of 0.088, 0.1975 and 0.2179, respectively. This shows that, on average, our method can produce estimates of the ENF that are closer to the ground truth and is more resistant to noise.

Fig. 2 shows the results obtained by applying all methods to two simultaneous recordings made using our sensing hardware and an audio recorder. Our Hilbert-based method achieves a significantly higher correlation coefficient, contains fewer outliers and also presents a better visual correspondence with the ground truth. From the other two methods, STFT provides a better visual result, with fewer fluctuations and outliers, but the correlation coefficient does not differ significantly from Spectrum Combining. In general though, the latter method produced better results overall.

B. One Class Support Vector Machine (OC-SVM) Classification

Much of the recent work using the ENF criterion is focused on establishing region-of-recording using a supervised learning approach, without the need to compare the ENF signal from an audio recording to a simultaneously recorded clean ENF signal [9]. In simple words, the extracted ENF signal must be qualitatively similar to ENF signatures representative of that particular power grid. Following this line of reason, we developed another measure of performance for extraction techniques, which is the classification results produced by an OC-SVM classification system. We use recordings made directly from a power socket, using our hardware, to train an OC-SVM, and then the ENF signals extracted from audio recordings from the same grid for testing. Apart from the classification percentage, we also present the mean score for each configuration. This score results from our OC-SVM classifier and serves as a confidence measure for the classification result. Intuitively, the higher the resulting score for a particular example the more confident we are that it is classified correctly.

TABLE I. FEATURES USED IN OUR CLASSIFICATION MODEL

Index	Features
1	mean of ENF segment
2	log (variance) of ENF segment
3	log(range) of ENF segment
4	log (variance) of approximation after 9-level wavelet analysis
5-13	log(variance) of nine levels of detail signal computed through 9-level wavelet analysis
14-15	AR(2) model parameters α_1 and α_2
16	log(variance) of the innovation signal after AR(2) modeling.

The classification system used to measure the performance of our ENF extraction methods is based on the multiclass classification scheme developed in [9]. Our focus has been on developing an extraction method that would be able to distinguish between recordings made in different grids and that is why we evaluated these methods based on how well the produced estimates fit into a class, defined by ground truth ENF signatures.

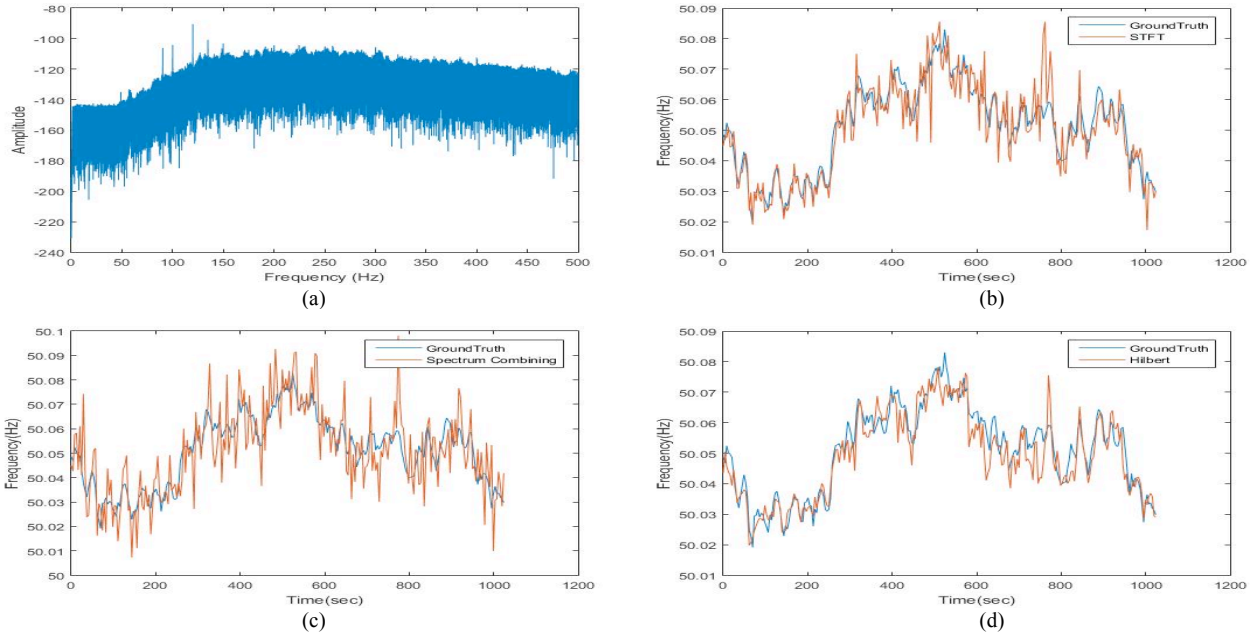


Fig. 2. (a) Recording under high SNR conditions. The corresponding correlation coefficients are 0.919 for STFT (b), 0.804 for Spectrum Combining (c) and 0.935 for Hilbert analysis (d).

TABLE II. RESULTS USING ALL RECORDINGS FOR TRAINING

Training	Testing	Quality metrics	Hilbert	Spectrum Combining	STFT	Noise
Power	Audio	Classification percentage (Mean Score)	71.41% (35.40)	42.86% (14.00)	44.19% (5.50)	No
			73.22% (52.40)	42.46% (28.88)	43.25% (4.56)	Yes
Audio	Power		99.90% (36.18)	84.70% (13.33)	61.79% (3.85)	No
			99.90% (52.37)	98.85% (16.34)	99.34% (21.37)	Yes

TABLE III. RESULTS USING ONLY NON-CONCURRENT RECORDINGS FOR TRAINING

Training	Testing	Quality metrics	Hilbert	Spectrum Combining	STFT	Noise
Power	Audio	Classification percentage (Mean Score)	71.41% (29.87)	51.12% (13.56)	46.80% (3.51)	No
			73.12% (37.98)	53.99% (27.43)	42.62% (10.44)	Yes
Audio	Power		99.67% (19.65)	72.34% (6.11)	8.56% (0.94)	No
			99.84% (26.44)	99.51% (16.89)	89.20% (7.53)	Yes

As a first step, we choose features that can quantitatively describe our grids. We use sets of 8-minute-long segments of ENF signals and the same features which have been used in [9], shown in Table 1. These features have proven to be good predictors of a segment's region-of-recording and accurately differentiate between recordings from different grids. We assume that they will also prove good predictors for our OC-SVM classification problem. To test this hypothesis, we first train our classifier using a subset of the power recordings, which we consider as ground-truth and certainly contain the ENF signature, and test the classifier using the rest of the power recordings. Using a 62% percentage of our clean signals for training and the rest 38% for testing, we obtained a 98% correct classification percentage with a mean score of 99.10, and therefore concluded that our assumption was valid.

In order to make our classifier independent of the segmentation method used, as it was possible to get segments that did not contain an ENF signature representative of the power grid, we chose to use overlapping segments both for training and testing. This approach has the added benefit of producing more training examples and creating a better defined class in the feature space.

We distinguished between two different arrangements for our classification system. In the first one, all of the available signals were used, both for the training and the testing set. In the second one, simultaneous recordings were isolated, a portion of them was reserved for the testing set and their counterparts were excluded from the training set. In general, we reserved the audio recordings which presented the highest SNR for the noisy dataset, and excluded their concurrent counterparts from the clean dataset. This arrangement is considered to be less-biased than the first one, as it makes certain that our system trains on and accurately identifies the power grid signature present in the ENF signals.

Additionally, two configurations were used for dealing with the recordings made by our sensing hardware. In the first one, we used the clean ENF signals extracted by applying the zero crossings method on the recordings. In the second configuration, the clean ENF signals were corrupted with Gaussian noise, a technique used in [9], where a noise-adaptation multi-class classification system was developed, and showed an increased performance on classification results. Again, this configuration was first tested by using only these corrupted signals both for training and testing an OC-SVM. The performance of our classifier did not change substantially, and yielded a 97.36% correct classification rate, with a mean score of 97.61.

Finally, all resulting scenarios were implemented in two versions, each time interchanging between audio and clean recordings for the training and testing set. This serves to test our extracted signals' ability to both fit in a created SVM class and form a well-defined class on their own.

Our results are presented in Tables II and III. The Hilbert-based technique performs consistently better in all possible testing scenarios, yielding both better classification results and confidence values. The Spectrum Combining (SC) and STFT methods show similar performance, with SC producing a higher confidence level in all its classification results.

Spectrum Combining shows an improved performance when evaluated on non-concurrent recordings. This improvement can be attributed to the fact that only the audio recordings with the highest SNR were used for the noisy database, thus acquiring ENF signals much closer to the ground truth, which are more likely to contain the power grid's signature. STFT is also severely limited when applied to classifying clean recordings with an OC-SVM trained on audio recordings, especially in the case of using only non-concurrent recordings. On the other hand, the Hilbert method

is more consistent in its results, which leads to the conclusion that it is, in general, more resilient to noise.

In all cases, the classifiers trained on audio recordings and tested with clean signals produced far better results. This is consistent with our interpretation of them as noisy versions of the correct signals, thus producing features with a higher variance that form a hyper-surface of larger volume in the feature space, which makes it easier to classify the clean ENF signals. Finally, corrupting the signals from clean recordings with Gaussian noise, has little impact on the classification percentages, when these recordings are used for training, but greatly improves the confidence values. Again, this can be attributed to the larger volume of the resulting hyper-surface in the feature space.

IV. Conclusions

The obtained results indicate that the Hilbert based approach is more suitable for ENF extraction, especially in applications which require identification of region-of-recording. Its main advantage is that it accounts for the non-stationarity of the acquired data, while both of the other methods used, assume a local stationarity within each particular frame. Moreover, our method performs better on average, even under relatively low SNRs, when the other methods fail to extract the ENF signal.

An interesting phenomenon, which was observed during the acquisition and analysis of the Greek power grid database, is that the ENF signature appears to slightly change over different months. Specifically, a portion of the recordings were made during early-to-mid winter and the rest during early spring. Presumably, this can be attributed to different load demands but may also be dependent on weather conditions and the way they impact performance of power generators and the power grid. This property can be further investigated by acquiring and comparing recordings made in different times of the year.

Future work in the ENF extraction problem should be directed towards identifying the ENF signature under any noise conditions. Of particular interest is performing a similar analysis to recorded calls made from mobile phones, an application of significant interest to Information Forensics.

ACKNOWLEDGMENTS

The authors would like to thank Georgios Roustas, Fanouria Athanasiou and Maria Papaioannou, co-members of the team which participated in the 2016 Signal Processing Cup. Our combined work, as members of that team, forms the basis of this paper.

REFERENCES

[1] M. H. Bollen and I. Gu, “Signal processing of power quality disturbances,” Hoboken, NH, USA: Wiley, 2006.

[2] C. Grigoras, “Applications of ENF analysis in forensic authentication of digital audio and video recordings,” J. Audio Engineering Society, vol. 57, no. 9, pp. 643-661, Sept. 2009.

[3] 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, 2009.

[4] 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, 2013.

[5] R. Garg, A. L. Varna, and M. Wu, “‘Seeing’ ENF: natural time stamp for digital video via optical sensing and signal processing,” Proc. 19th ACM Int. Conf. Multimed. - MM ’11, no. October, p. 23, 2011.

[6] M. Huijbregtse and Z. Geradts, “Using the ENF criterion for determining the time of recording of short digital audio recordings,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5718 LNCS, pp. 116–124, 2009.

[7] O. Ojowu, J. Karlsson, J. Li, and Y. Liu, “ENF extraction from digital recordings using adaptive techniques and frequency tracking,” IEEE Trans. Inf. Forensics Secur., vol. 7, no. 4, pp. 1330–1338, 2012.

[8] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,” Proc. R. Soc. Lond. A, vol. 454, no. 1971, pp. 903–995, 1998.

[9] A. Hajj-Ahmad, R. Garg, and M. Wu, “ENF-based region-of-recording identification for media signals,” IEEE Trans. Inf. Forensics Secur., vol. 10, no. 6, pp. 1125–1136, 2015.