Digital Investigation 22 (2017) S115–S126

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

DFRWS 2017 USA — Proceedings of the Seventeenth Annual DFRWS USA

Time-of-recording estimation for audio recordings

Lilei Zheng*, Ying Zhang, Chien Eao Lee, Vrizlynn L.L. Thing

Cyber Security Cluster, Institute for Infocomm Research, Singapore

Keywords:
Audio timestamp
Electrical network frequency
Pattern recognition
Sequence similarity
Large-scale search

A B S T R A C T

This work addresses the problem of ENF pattern matching in the task of time-of-recording estimation. Inspired by the principle of visual comparison, we propose a novel similarity criterion, the bitwise similarity, for measuring the similarity between two ENF signals. A search system is then developed to find the best matches for a given test ENF signal within a large searching scope on the reference ENF data. By empirical comparison to other popular similarity criteria, we demonstrate that the proposed method is more effective and efficient than the state-of-the-art. For example, compared with the recent DMA algorithm, our method achieves a relative error rate decrease of 86.86% (from 20.32% to 2.67%) and a speedup of 45× faster search response (41.0444 s versus 0.8973 s). Last but not least, we present a strategy of uniqueness examination to help human examiners to ensure high precision decisions, which makes our method practical in potential forensic use.

Introduction

The frequency of the electrical power grid – the electrical network frequency (ENF) – has been found as a unique fingerprint which is unintentionally embedded in audio (Hua et al., 2014) or video recordings (Garg et al., 2013). Centred at a nominal frequency of either 50 Hz (e.g., Singapore) or 60 Hz (e.g., United States), a real ENF signal contains random fluctuations over time around its nominal value and appears as a sequence of fluctuated frequency values. Moreover, these random fluctuations are consistent across different places within the same power grid (Grigoras, 2007). As a consequence, recordings captured in different places at the same time will have ENF fingerprints showing the same fluctuations. In order to verify if two recordings are captured at the same time, one solution is to compare their ENF fingerprints to see if they are visually matched to each other (Huijbregtse and Geradts, 2009).

A digital recording device can capture the ENF from the local power grid when it is directly mains-powered or placed near other mains-powered equipments (it has been verified that the low frequency signal can be captured by microphones in a short distance (Cooper, 2009). Specifically, an electrical transformer (Cooper, 2009) directly connected to power supply can be used to record and store pure ENF signals over a long period of time as a reference database. For recordings from other devices such as portable audio recorders and stationary surveillance systems, their ENF signals are compared to the reference and the best visual matches inform the time when these recordings were captured. This application is named as time-of-recording estimation (Huijbregtse and Geradts, 2009; Kantardjiev, 2011; Baksteen, 2015), which is of great potential use in multimedia forensics.

Given the ENF of an audio recording, visual comparison is only applicable to finding a match in a very short reference ENF sequence. For a large reference database, automatic comparison is needed and a searching routine is necessary to locate the best matches in the reference (Huijbregtse and Geradts, 2009). To simplify the interpretation through this paper, we call the ENF of the reference database as *reference ENF* and that of a single audio recording as *test ENF*. As we have mentioned, both the test ENF and the reference ENF are represented as sequences of fluctuated values. In the task of time-of-recording estimation, the reference ENF sequence is usually much longer than the test ENF. Classical searching algorithms include minimum mean squared error (MMSE) and maximum correlation coefficient (MCC) (Huijbregtse and Geradts, 2009; Kantardjiev, 2011; Baksteen, 2015) that compare a given test ENF to all possible reference ENF segments of the same length. The minimum or the maximum indicates the best match.

To ensure high accuracy of locating the best match, efforts have been made to two aspects: (1) pattern extraction of the test ENF in noise (Cooper, 2009; Garg et al., 2012; Chai et al., 2013; Bykhovsky and Cohen, 2013; Hajj-Ahmad et al., 2013) and (2) searching algorithm robust to noise (Hua et al., 2014). More attention has been

* Corresponding author.
E-mail address: zhengll@i2r.a-star.edu.sg (L. Zheng).

http://dx.doi.org/10.1016/j.diin.2017.06.001

paid to the former concerning audio signal processing than the latter part of pattern recognition. For example, using median filtering (Cooper, 2009; Hua et al., 2014), evaluating harmonic models (Bykhovsky and Cohen, 2013; Hajj-Ahmad et al., 2013; Chai et al., 2013) or developing autoregressive model (Garg et al., 2012) was shown to be effective in reducing the signal noise and improving estimation accuracy. In contrast, Hua et al. (2014) proposed a threshold based dynamic matching algorithm (DMA) to deal with the in-band noise and frequency resolution problem. The DMA method serves as a better substitute to the conventional MMSE searching algorithm. However, the DMA method strengthens the robustness of pattern recognition at the cost of more computation time. Hence its application was limited to audio timestamp verification where the reference ENF was within a small searching scope specified by the user (Hua et al., 2014). For ENF matching in a large reference database, searching efficiency is as essential as the matching accuracy (Kantardjiev, 2011).

In this paper, we propose a novel similarity measurement for evaluating the distance between two ENF sequences. Based on this similarity, we develop a fast search system to find the best matches from the long reference ENF sequence for a given test ENF. The contributions of this paper with respect to previous works are the following.

- We collect both power recordings containing reference ENF signals and audio recordings containing test ENF signals in Singapore. With these data, we establish a dataset for performance evaluation on the task of time-of-recording estimation. The test dataset consists of 187 practical audio recordings, which is much more sufficient than existing works.
- We invent the bitwise similarity (bSim) to compare a pair of ENF sequences. The bSim is inspired by the human visual comparison criterion that directly measures the proportion of local matching between two ENF sequences. Experimental results show that the bSim criterion, especially its binarization process, plays an important role in bringing up fast and accurate ENF matching.
- We build a search system that significantly surpasses previous time estimation systems in terms of both estimation accuracy and computational efficiency.
- We consider a Top-*n* retrieval strategy as *uniqueness examination* to assist human examiners to confirm the estimated time. This makes the proposed method practical in applications of forensics concerns.

Labelled ENF signals in Singapore

The entire city of Singapore is covered by a single large power grid that is operated by the SP PowerAssets company and regulated by the government agency, Energy Market Authority (Energy Market Authority, 2016; Singapore Power Group, 2016). As one of the most reliable electricity grid in the world, the island-wide ENF serves as good timestamp within audio recordings in Singapore. For the sake of building a practical system to estimate recording time for audio recordings, we establish the first dataset of labelled ENF signals in Singapore (LESS). This dataset is composed of two subsets, one is the subset of reference ENF captured in power recordings and the other is the subset of test ENF captured in audio recordings.

Reference ENF from power recordings

The frequency of the Singapore power grid is maintained around 50 Hz, with an allowed deviation of ±0.2 Hz (Energy Market Authority, 2016). Clean ENF data can be captured by digital recorders directly connected to power supply. We use a in-house sound card with a sampling frequency of 400 Hz to produce power recordings since 3 September, 2013. Hence up to now we have a large collection of more than 3 years' reference ENF data. Each noiseless power recording lasts for 1 h, and we directly apply time frequency analysis – short-time Fourier transform (STFT) plus quadratic interpolation (Hua et al., 2016) – to extract the ENF signal. Moreover, all the reference ENF data have automatically annotated recording times accurate to 2 s. This allows us to provide time estimation for a test audio if a successful ENF matching is found between its ENF and the reference ENF.

Test ENF from audio recordings

The ENF extraction procedure for an audio recording is slightly different from that of a power recording. The usual sampling frequency of an audio recording is much higher, e.g., 44.1 kHz for music or 8 kHz for speech. Pre-processing procedures such as downsampling and bandpass filtering are therefore needed before applying STFT to extract the ENF signal. The configuration and parameter setting of our ENF extraction procedure is the same with that in (Hua et al., 2014).

The quality of test ENF is usually lower than that of reference ENF. This is because audio recorders are not necessarily connected to any power supply but may be battery-powered. A portable audio recorder captures sound in relatively noisier and more complex environment than that of the power recordings, so a clean ENF signal in an audio recording is not guaranteed. In fact, a portable recorder can capture the valid ENF signal only when it is located near some mains-powered equipments. This condition is most likely to be satisfied in a smart city such as Singapore that is covered by a dense electricity grid and electrical equipments (e.g., household appliances, street lights).

From 30 June, 2016 to 24 August, 2016, we collected 187 audio recordings in different areas of Singapore. As mobile phones are the most common and portable recording device nowadays, all the audio recordings were taken by mobiles including iPhone (the most popular device in 2016), Android phone and Windows phone. Like the power recordings, the recoding time of our audio recordings is automatically noted by the mobile apps.

The distribution of the 187 recordings with respect to their recording locations is given in Table 1. The numbers of recordings are in accordance with the mobile owners' daily active areas. For example, the top three active areas are office, foodcourt and home, respectively. Table 2 shows the distribution with respect to the audio length. Most of the audio recordings are between 20 and 40 min.

Although the size of our test audio dataset is significantly larger than the data sizes in previous studies (Hua et al., 2014; Cooper, 2009; Garg et al., 2012; Huijbregtse and Geradts, 2009), it is still limited for in-depth investigation, e.g., there are too few recordings at outdoor locations such as gym, park and walkway. This prevents us from studying the influence of outdoor environment conditions. Thus a plan of collecting more audio recordings is expected in our future work.

Table 1
Distribution of audio recordings in the LESS dataset with respect to their recording locations.

Location	Bus station	Casino	Cinema	Foodcourt	Gym	Home	Mall	Office	Park	Sports centre	Swimming pool	Walkway	Total
No. of recordings	2	1	6	**42**	3	**30**	18	**61**	1	3	14	6	187

Distribution of audio recordings in the LESS dataset with respect to their recording locations. The top three active areas are highlighted in bold.

Table 2
Distribution of audio recordings in the LESS dataset with respect to their audio lengths.

Length (min)	0–10	10–20	20–30	30–40	40–50	50–60	60–70	Total
No. of recordings	5	17	**62**	**62**	20	16	5	187

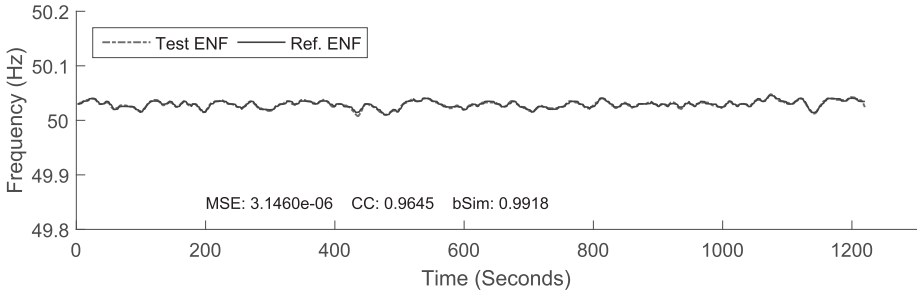
Distribution of audio recordings in the LESS dataset with respect to their audio lengths. The top two categories are highlighted in bold.

Fig. 1 illustrates three pairs of ENF signals extracted from our power recordings and audio recordings. Each pair of test ENF and reference ENF is at the same recording time. The first example, Fig. 1(a), shows a well matched pair that the test ENF (dashed red line – in the web version) lies on the reference ENF (solid blue line – in the web version). In the second example (Fig. 1(b)), a noisy test ENF is presented that no match can be observed between it and its corresponding reference. By manually examining the entire test set, we found that full matching or full mismatching seldom happens, and partial matching is the most common type. An example is given in Fig. 1(c), we can see that the major parts of the test and reference ENF well overlap with each other, but apparent mismatch exists in the last 3 min.

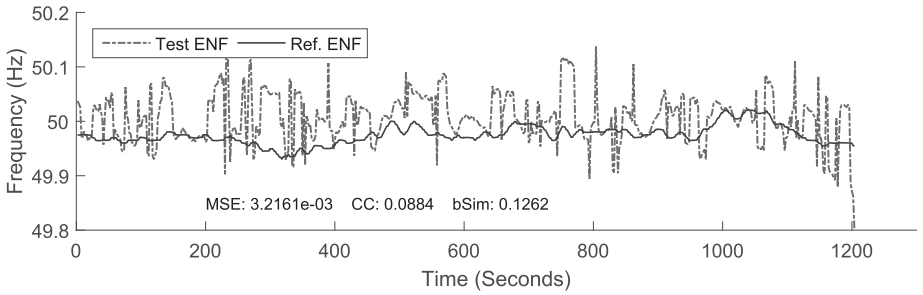
For the task of time-of-recording estimation, it is easy for the first case since a clean test ENF is conducive to locating perfect matching, and it is difficult for the second case due to heavy noise so that no reference segment matches the test ENF. In order to increase the accuracy of correct time estimation, more effort should be given to the third case by finding local matched fragments between the test and the reference.

Pairwise similarity between sequences

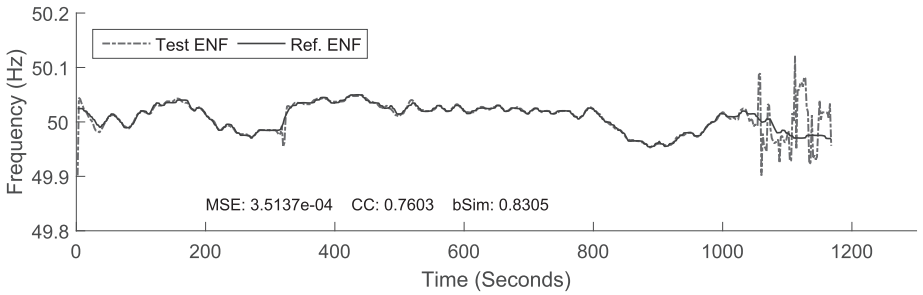
Visual comparison is the most natural way to determine if two sequences are similar to each other. This manual work need to be replaced by automatic comparison in order to increase the efficiency when given a long reference ENF sequence. In this section,



(a) ENF signals (20 minutes) @ 2016-07-06 12:08:30



(b) ENF signals (20 minutes) @ 2016-08-02 20:30:13



(c) ENF signals (20 minutes) @ 2016-08-20 13:00:38

Fig. 1. Test and reference ENF captured in Singapore at three different times. All the three pairs of signals last about 20 min from their starting time.

we first review the existing measurements of pairwise similarity for sequences, and then propose a new similarity measurement that better fits to the task of ENF matching.

Visual comparison

Visual comparison is the least efficient but the most effective way to evaluate ENF matching (Kajstura et al., 2005; Huijbregtse and Geradts, 2009; Baksteen, 2015). For the task of audio timestamp verification where the test ENF has a claimed recording time, a quick decision can be made by visually comparing the test ENF to the reference ENF at the same time. Like that shown in Fig. 1, it is easy to see whether the test ENF (dashed red line – in the web version) is matched to the reference ENF (solid blue line – in the web version): full matching in the first example, full mismatching in the second and partial matching in the last one.

However, when the test recording has no claimed date of origin, one needs to find a similar reference ENF segment from a large reference database for the test ENF, and the visual search is obviously not practicable. Moreover, visual comparison lacks numerical measurement of how similar the two sequences are. These disadvantages prevent visual comparison from being adopted in practical systems for time-of-recording estimation.

Mean squared error

The mean squared error (MSE) is a popular measure of the matching quality between two sequences (Hua et al., 2014; Cooper, 2009; Baksteen, 2015; Huijbregtse and Geradts, 2009; Kantardjiev, 2011). We use column vectors $\mathbf{t} = [t_1, t_2, \ldots, t_N]^T$ and $\mathbf{r} = [r_1, r_2, \ldots, r_N]^T$ to represent the sequences of test ENF and reference ENF, respectively. Here N is the length of the sequences and $(\cdot)^T$ denotes transpose. The MSE between them is given by

$$MSE(\mathbf{t}, \mathbf{r}) = \frac{1}{N}\|\mathbf{t} - \mathbf{r}\|^2 = \frac{1}{N}\sum_{i=1}^{N}(t_i - r_i)^2. \quad (1)$$

One can see that the value is always non-negative, and being closer to zero indicates better matching.

Correlation coefficient

An alternative criterion to the MSE is the Pearson's correlation coefficient (Hua et al., 2014; Baksteen, 2015; Huijbregtse and Geradts, 2009; Kantardjiev, 2011), which is formulated as

$$CC(\mathbf{t}, \mathbf{r}) = \frac{(\mathbf{t} - \bar{\mathbf{t}})(\mathbf{r} - \bar{\mathbf{r}})}{\|\mathbf{t} - \bar{\mathbf{t}}\|\|\mathbf{r} - \bar{\mathbf{r}}\|} = \frac{\sum_{i=1}^{N}(t_i - \bar{\mathbf{t}})(r_i - \bar{\mathbf{r}})}{\sqrt{\sum_{i=1}^{N}(t_i - \bar{\mathbf{t}})^2}\sqrt{\sum_{i=1}^{N}(r_i - \bar{\mathbf{r}})^2}}, \quad (2)$$

where $\bar{\mathbf{t}}$ is the arithmetic mean of the sequence $\mathbf{t}$. The value of the CC lies in the range [−1, 1], and a value of 1 indicates exact matching. Baksteen (2015) has proven that the CC is equal to zero mean MSE under the assumption that the standard deviation of the test ENF is equal to that of the reference ENF. However, this assumption is hard to meet for the purpose of searching in a large reference database because the reference sequence in comparison is probably different from the test sequence. Therefore, the MSE and the CC find their places with respect to different requirements: the CC performs better than the MSE in pattern matching for audio recordings shorter than 10 min (Huijbregtse and Geradts, 2009); but the CC has higher computation cost than the MSE (Cooper, 2009), making it less attractive for large-scale search.

Bitwise similarity

The above two criteria accumulate the local difference between each single pair of elements in the two vectors. For example, when there are spikes on the test ENF (see Fig. 1(b) or (c)), the large element-wise errors in the spike area contribute a lot to the final MSE and make the accumulated score increased significantly. In other words, besides the scope of the mismatching area, the MSE score also takes into account the scale of local mismatching. This is undesirable for measuring ENF matching because the scale of mismatching is out of control due to uncertain causes, such as the offset phenomenon noticed in (Kajstura et al., 2005; Huijbregtse and Geradts, 2009).

By making the ENF sequences to have zero means, the CC criterion reduces the influence caused by large local mismatching to a certain extent, so it was found as a better similarity measurement for robust ENF matching (Huijbregtse and Geradts, 2009). As mentioned above, this improvement is achieved at the expense of more computation time.

Inspired by the above criteria, especially by the visual comparison, we propose a novel similarity called *bitwise similarity* (bSim) to measure ENF matching. The idea can be expressed by

$$bSim(\mathbf{t}, \mathbf{r}) = \frac{1}{N}\sum_{i=1}^{N} s_i,\ s_i = \begin{cases} 1, & t_i \approx r_i \\ 0, & t_i \not\approx r_i \end{cases}, \quad (3)$$

where the assertion $t_i \approx r_i$ returns 1 if t_i is matched to r_i, otherwise it is 0. By binarizing the scale of local difference, the bSim criterion directly measures the proportion of local matching between two ENF sequences. Like the human visual system, the bSim criterion does not directly count the exact difference values but treats all large local mismatching the same, i.e., 0 as $t_i \approx r_i$ for all mismatched elements.

When our eyes can visually determine if t_i is matched to r_i or not, it is difficult for a machine to make such a decision without numerical measurement. The assertion $t_i \approx r_i$ is therefore realized by a binarization function $||t_i - r_i|| < \theta$, and Equation (3) is rewritten as,

$$bSim(\mathbf{t}, \mathbf{r}) = \frac{1}{N}\sum_{i=1}^{N} s_i,\ s_i = \begin{cases} 1, ||t_i - r_i|| < \theta \\ 0, ||t_i - r_i|| \geq \theta \end{cases}, \quad (4)$$

where θ serves as the threshold between matching and mismatching. After binarization, the difference between the two sequences becomes a single sequence of bits, where consecutive 1s indicate local matched fragments. This is why we name the proposed measurement as bitwise similarity.

Similar to the CC criterion, the bSim score falls within a closed interval [0, 1]. It reflects how much the test ENF is matched to the reference ENF. Fig. 1 shows the distance and similarity scores below the ENF curves, the threshold θ for bSim is set to 0.005. When the MSE score approaching 0 implies a full matching, the CC and bSim scores are better at revealing the proportion of local matching.

Moreover, the bSim score is efficiently computable. Comparing Equation (4) to Equation (1), we can see that for each pair (t_i, r_i), the MSE requires a squaring following a subtraction, the bSim takes the same subtraction and compares the result's absolute value to a threshold. Both of them consume less computation cost than the CC given in Equation (2). In summary, the proposed bSim takes advantages of both CC and MSE to achieve accurate quantization of the sequence correlation and fast computation.

Time-of-recording estimation

Large-scale search

Without a claimed recording time, the verification task between two sequences of the identical length (see Fig. 1) now becomes a searching task of finding the best matches for a test ENF from a long reference ENF. For example, given a test ENF of 10 min as shown in Fig. 2, we do not know the exact recording time but an approximate time range of 30 min. We have to compare the test ENF to every possible reference segment of 10 min within the range. This comparison procedure is known as sequence alignment, and a similarity matrix (Von Luxburg, 2007) (also called affinity matrix (Bengio et al., 2004) or distance matrix (Zheng et al., 2012; Hua et al., 2016)) is used to load the element-wise similarity or distance values.

Assuming that the lengths of the test ENF $\mathbf{t}$ and reference ENF $\mathbf{r}$ are N and M, respectively. We calculate the absolute error for every pair of elements between the two sequences and get a distance matrix $\mathbf{D}$ of size $N \times M$. Let d_{ij} denote the element of the matrix at the intersection of the i_{th} row and the j_{th} column, the value of d_{ij} is given by $d_{ij} = ||t_i - r_j||$.

Fig. 2(a) illustrates the distance matrix by a dotplot where large distance values are shown in dark and small distance values are represented by white dots. Since the test ENF is supposed to have a matched segment on the long reference ENF, one can discover the square region with a white diagonal from its left bottom to the right top. One may also notice a horizontal line across the whole dotplot near the top boundary, which is due to an unexpected spike near the end of the test ENF. At this stage, locating the position of the white diagonal will tell us the estimated recording time of the test recording. Existing methods (Huijbregtse and Geradts, 2009; Hua et al., 2014, 2016) directly search for the best ENF match in this distance matrix. However, we develop a simpler searching routine with respect to our proposed bitwise similarity.

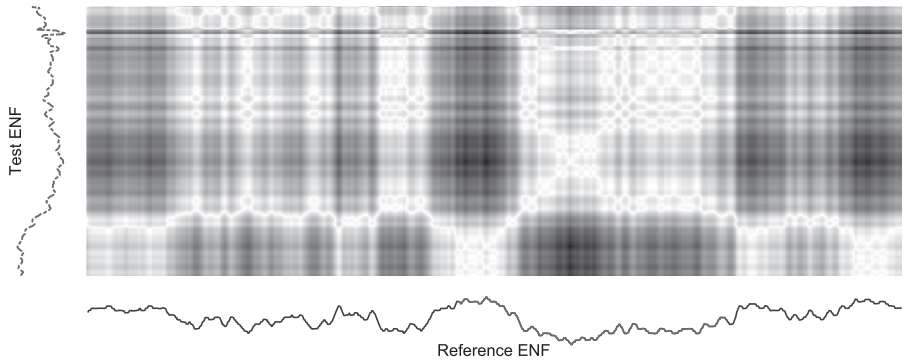
Binarized similarity matrix

According to the bSim defined in Section Bitwise Similarity, a pair of elements is considered as matched if their absolute error is smaller than a threshold θ, i.e., $||t_i - r_j|| < \theta$. As a result, the above distance matrix $\mathbf{D}$ can be transformed to a similarity matrix $\mathbf{S}$ by simply taking $s_{ij} = (d_{ij} < \theta)$. Fig. 2(b) plots the similarity matrix. Compared to the distance matrix in Fig. 2(a), the similarity matrix has higher color contrast, i.e., the white diagonal is more visually distinguishable from the background. In addition, the similarity matrix of binary values requires less machine memory space and enables faster computation than the distance matrix of floating-point numbers.

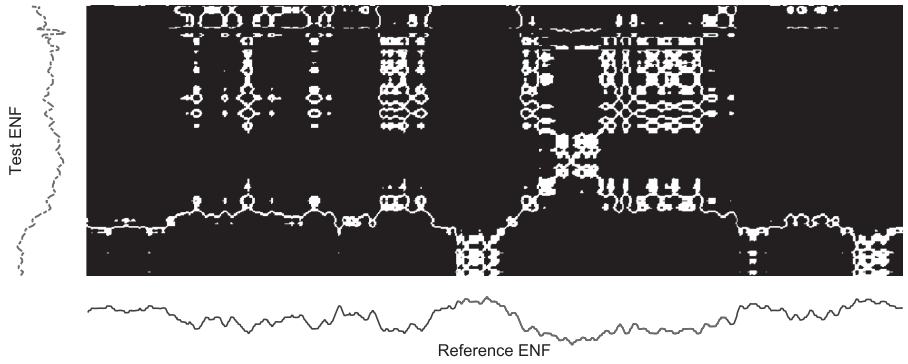
Among all possible segments on the reference ENF, we aim to find out the one having the maximum similarity with the test ENF. Let $\mathbf{r}^{(k)}$ denote the reference ENF segment of length N and with a start from r_k, the objective function is given by

$$\arg\max_k \; bSim\left(\mathbf{t}, \mathbf{r}^{(k)}\right) = \arg\max_k \; \sum_{i=1}^{N} \frac{s_{ij}}{N}, \; j = k + i - 1, \tag{5}$$

where s_{ij} is the $(i,j)_{th}$ element of the similarity matrix, and arg max denotes the argument of the maximum. The parameter k is an



(a) Before binarization



(b) After binarization

Fig. 2. Dotplots illustrating distance or similarity matrices between two ENF sequences. (a) Distance matrix before binarization and (b) similarity matrix after binarization. Dark points mean low similarity (respectively large distance) and light points mean high similarity (respectively small distance). Test ENF (10 min) @ 2016-08-23 23:20:52; reference ENF (30 min) @ 2016-08-23 23:10:00.

integer between 1 and $M-N+1$ to ensure that all the reference segments are of length N. This objective of finding the best k can be explicitly explained by Fig. 3 where the bSim scores between the test ENF and all reference segments are plotted as a curve, and the peak of the highest similarity value indicates the best matching.

Matched fragment localization

The argument of the maximum bSim tells us the estimated recording time of the test ENF. Like that in Fig. 1, we can draw the test ENF with its matched reference segment in Fig. 4. We also plot an indicator line, the binary curve of $\|t_i - r_i\| < \theta$ with respect to i, to show the local matched fragments by consecutive 1s. When the bSim score implies how well the test ENF is matched to the reference ENF, the local matched fragments inform where the exact matches are. This plays an important role in forensic proof that the audio content within the range of an exactly matched fragment is innocent from being tampered (Esquef et al., 2014; Hua et al., 2016).

We can see that there are three matched fragments in Fig. 4 where the first is the longest. And then we come to the problem that how to automatically locate these fragments. Instead of directly counting the 1s on the binary curve, we once again take advantages of our bitwise similarity setting and propose a fast solution using the XOR (i.e., exclusive disjunction) operation. Algorithm 1 summarizes the pseudo code for locating the fragments of consecutive 1s. The principle is that such a fragment always have bit value changed at its beginning (from 0 to 1) and end (from 1 to 0). Algorithm 1 returns a vector **p** containing the positions of value changes. A pair of adjacent changes indicates consecutive 1s or 0s, i.e., $[p_{2i-1}, p_{2i})$ denotes a fragment of consecutive 1s and $[p_{2i}, p_{2i+1})$ denotes a fragment of consecutive 0s.

Algorithm 1: Locating consecutive 1s on a binary sequence

input : $\boldsymbol{s}$ – a binary sequence of length N
output: $\boldsymbol{p}$ – an array containing pairs of indices

```
% initialization
s ← {0; s; 0};    % add 0 to the sequence head and tail, respectively
p←NULL;    % empty array
% successive checking: a bit value change implies a beginning or end of a fragment
for t = 1, 2, ..., N + 1 do
    if s_t ⊕ s_{t+1} then
        p ← {p; t};
return p;    % each pair [p_{2i-1}, p_{2i}) denotes a fragment of consecutive 1s
```

Uniqueness examination for ENF patterns

The most important assumption behind time-of-recording estimation is the uniqueness of ENF patterns, i.e., the ENF pattern at a certain time is different from all the patterns occurs at other times. Empirical experience in (Huijbregtse and Geradts, 2009; Baksteen, 2015; Hua et al., 2014; Cooper, 2009) found that ENF patterns at different times are usually unique but can be very similar in some cases. Therefore, examining the existence of uniqueness is necessary before asserting that the estimated time is correct.

A short test audio may have a higher probability of obtaining multiple reference matches than a long test audio. This is because the fluctuation band is as narrow as only 0.4 Hz (e.g., [49.8, 50.2] Hz in Singapore (Energy Market Authority, 2016)), short ENF patterns are usually more flat and it is easier to find similar patterns at other times. And the existence of similar patterns prevents us from identifying the correct recording time. Previous works (Huijbregtse and Geradts, 2009; Baksteen, 2015; Hua et al., 2014) suggested that a test recording of 2 min is too short and may result in the failure of finding the unique ENF matching, recordings with duration at least 10 min have sufficiently complicated ENF patterns.

In fact, ENF patterns are randomly generated by the power grid and similar patterns always exist in the reference database. If a test recording accidently captures one of the similar patterns, it will cause confusion to the following process of ENF pattern matching. From the perspective of probability theory, a long ENF pattern is a sequence of a few consecutive short ENF patterns so that its probability of being similar to other patterns goes down when its length increases. This makes the impression that long test recordings are easier to get correct time estimation than the short ones.

Intra-reference similarity matrix

The length of the longest similar patterns is determined by the specified similarity criterion and the searching scope of the reference. We propose to examine the intra-signal similarity matrix of the reference ENF to study this problem. As mentioned in Section Reference ENF from Power Recordings, an ENF vector of 1800 frames is extracted from each 1-h power recording (Hua et al., 2014). Thus the ENF signal in a single day is a sequence of length 43,200 (=1,800 × 24), and that of a month, 30 days, is a sequence of length larger than 1.29×10^6. Examining the huge intra-signal similarity matrix requires a fast searching algorithm for locating pattern matching. The proposed fragment localization algorithm (Algorithm 1) in the previous section meets our requirement.

Fig. 5 plots the intra-reference similarity matrix of 1 h (60 min) ENF data. The side length of this square matrix is 1800. The threshold θ for binarization is still set to 0.005 here. Different from the inter-signal similarity matrix between the test and reference (e.g., Fig. 2(b)), the intra-signal similarity matrix is apparently

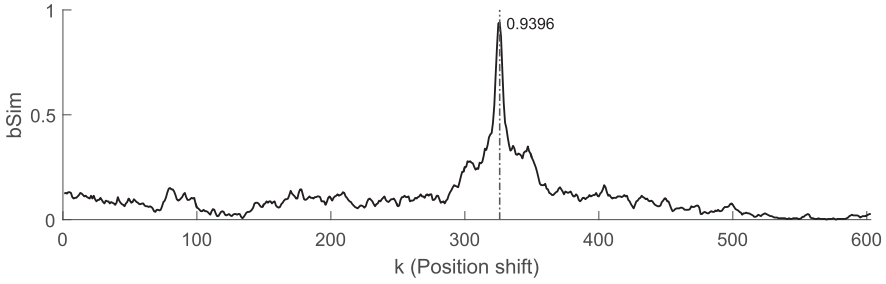


Fig. 3. Curve of the bSim for the test and reference ENF signals shown in Fig. 2. The maximum indicates the start position of the matched reference segment.

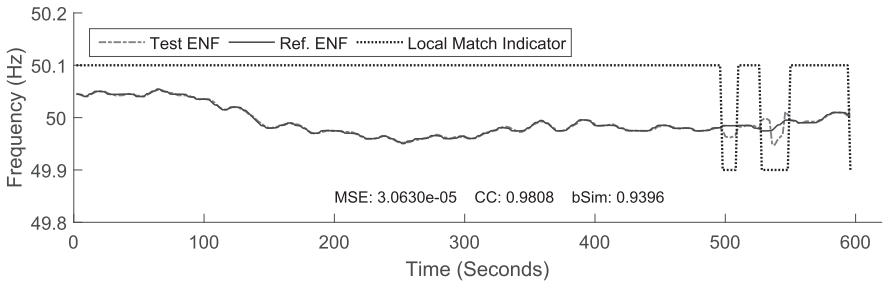
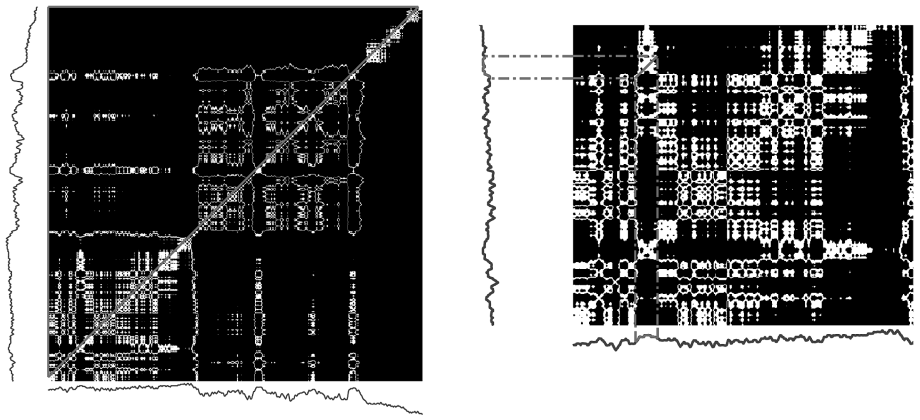


Fig. 4. ENF signals (10 min) @ 2016-08-23 23:20:52. The binary indicator line presents element-wise similarity values, where high values (consecutive 1s) denote local matched fragments.



(a) Intra-signal similarity matrix (b) Zoom in on the bottom left corner of (a)

Fig. 5. Intra-signal similarity matrix of the reference ENF (60 min) @ 2016-01-01 00:00:00. The longest matched fragment between different parts of this reference ENF is found of length 47 frames, i.e., 94 s.

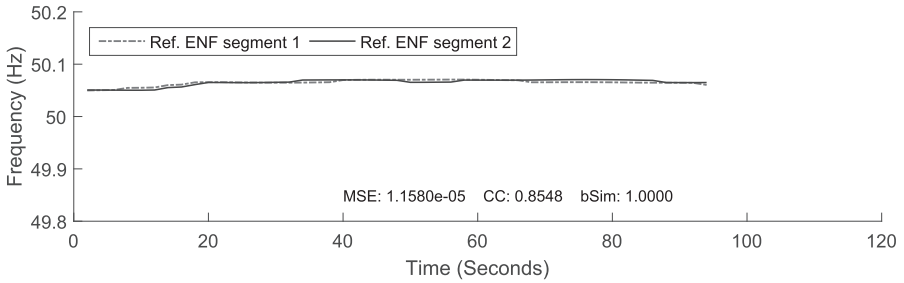


Fig. 6. Curves of the two matched reference segments in Fig. 5. Ref. ENF segment 1 (94 s) @ 2016-01-01 00:04:24; Ref. ENF segment 2 (94 s) @ 2016-01-01 00:17:24.

symmetric and its diagonal is an exact matching of the reference ENF itself. Since we aim at finding similar patterns at different times, the search area can be the upper triangle area excluding the diagonal area, e.g., within the red (in the web version) triangle in Fig. 5(a).

Specifically, in this area filled with binary values, Algorithm 1 can successively examine all the lines parallel to the diagonal and locate the longest fragment of consecutive 1s, which indicates the longest matched patterns at different positions (i.e., different recording times) of the reference ENF. Fig. 5(b) shows the positions of the longest matched patterns and Fig. 6 compares the two ENF segments. We can see good visual matching as well as a small MSE score and a large CC score. In other words, such a flat ENF pattern in Fig. 6 suggests two possible recording times within the specified hour, i.e., 2016-01-01 00:04:24 or 2016-01-01 00:17:24, which is an undesired situation for the purpose of time estimation.

We take the second reference ENF segment as an ideal test signal and put it into the searching process across the 1-h reference data. The curve of the bSim values for the test and reference ENF signals is shown in Fig. 7. The maxima indicate the possible start positions of the matched reference segments. The true recording time is marked with a vertical dashed line. However, it is difficult to determine the true recording time by the bSim curve only.

The length of the longest pattern in this example is of 94 s, which means a test recording of a shorter length is possible to be covered by such patterns and its recording time is uncertain.

The searching scope: reference length

The larger the searching scope, the longer the longest matched pattern. The reference length in the above example (Fig. 5) is of 1 h only. But the entire reference data in our LESS dataset is of more

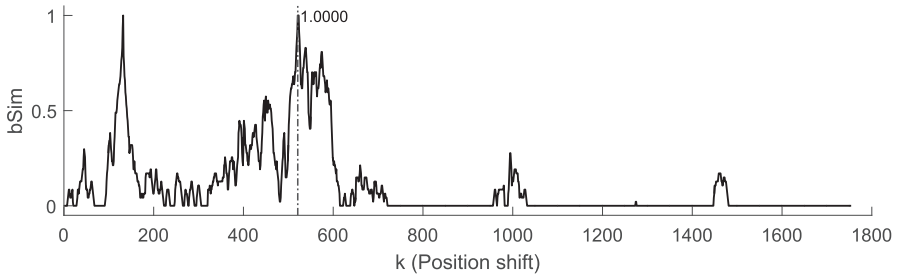


Fig. 7. Example of multiple peaks in the bSim curve. The two peaks indicate the possible recording times of the test signal, and the latter is the true one. The similarity threshold θ is set to 0.005.

Table 3
Length of the longest matched pattern found in the intra-signal similarity matrices of increasing reference lengths. The similarity threshold θ is set to 0.005.

Reference length (hour)	1	24 (1 day)	48 (2 days)	72 (3 days)	168 (1 week)	336 (2 weeks)	720 (1 month)
Pattern length (second)	94	298	298	298	310	310	436

than 26,280 h. Note that the time cost of searching in the upper triangle area increases quadratically with respect to the reference length. Hence the proposed Algorithm 1 running in linear time is particularly helpful for the sake of efficient searching.

Table 3 provides the lengths of the longest matched pattern with respect to reference ENF sequences of increasing lengths from 1 h to 1 month. In this experiment, all the searching scopes have the same start at 2016-01-01 00:00:00 and the similarity threshold θ is kept to 0.005. One can see that the longest matched pattern grows with respect to longer searching scope. For example, in the one-month reference ENF data, we have found a pattern matching as long as 436 s, i.e., more than 7 min.

The similarity threshold θ

The smaller the similarity threshold, the shorter the longest matched pattern. As mentioned before, the value of the threshold θ reflects the criterion of visual matching. Fig. 8 shows the effect of reducing the similarity threshold θ. With a smaller θ, the false peak in the bSim curve is indeed removed and the true recording time is obtained (compared to Fig. 7). Table 4 also explains that the length of the longest matched pattern increases with respect to larger similarity threshold.

Ideally, to reduce the probability of pattern matching in the reference itself, we can set the similarity threshold θ as close to zero as possible because the exact self-matching always exists for the reference data, i.e., the white diagonal in the intra-reference similarity matrix. However, this is impractical for pattern matching between the test and reference.

Table 4
Length of the longest matched pattern found in the intra-signal similarity matrix of the 3-day reference data with respect to different similarity thresholds.

Similarity threshold θ	0.0001	0.001	0.002	0.003	0.004	0.005	0.01
Pattern length (second)	38	102	114	114	142	298	882

Uniqueness examination

From the above analysis, we know that the two solutions of ensuring pattern uniqueness are to narrow the searching scope and adopt a tight matching criterion. However, neither of them can be satisfied in practical use. First, it is not guaranteed that the user can provide trusty information to narrow the searching scope. Second, exact matching only exists in intra-reference comparison but not in inter-signal comparison. For instance, visual matching between the test and the reference does not imply exact identical ENF values (see Fig. 1(a) by Zoom In). A too small θ will cause no ENF matching at all.

In short, the problem of similar patterns at different times is unavoidable in the task of ENF matching. Instead of making effort in improving the pattern uniqueness by smoothing the test ENF signals (Hua et al., 2014), we propose to retrieve the Top-n ($n \geq 2$) matched ENF patterns from the reference ENF to check if the top one pattern is unique. The idea is straightforward that if the top one matched reference segment has a significantly larger similarity value than the second best match, it implies that the captured ENF pattern is unique and the estimated time is reliable. When the top two retrieved segments have approximate similarity values, the

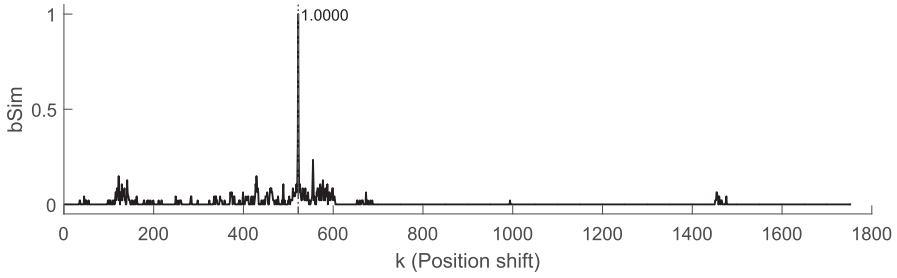


Fig. 8. With smaller similarity threshold θ, the false peak in the bSim curve (Fig. 7) is removed. Similarity threshold θ is set to 0.0001.

ENF criterion is hard to distinguish them and should seek help from other aspects, e.g., if the user can narrow the searching scope to exclude the false option. We name this strategy of Top-n retrieval as *uniqueness examination*. This strategy is simple but of important practical value to the task of time-of-recording estimation.

- It replaces the assumption of unique ENF pattern (Hua et al., 2014; Garg et al., 2013) by the process of examination. The assumption of uniqueness is doubted because pattern duplications always exist, even with very short searching scope and small similarity threshold. In Table 4, even with a θ as small as 0.0001, there is a matched pattern of length 38 s. The proposed examination detects the duplications and classifies non-unique patterns as an independent class of "unable to handle". This strategy is useful in filtering the exception that the test audio recording failed to capture the valid ENF pattern from the power grid, i.e., ensuring true positives.
- It makes the recording length as an indirect factor affecting the estimation result. Longer test recordings still have a larger probability to pass the uniqueness examination than the shorter ones, but this does not imply that a long ENF pattern is surely unique and a short one is certainly non-unique. For a test recording passed the uniqueness examination, no matter it is long or short, a trusty recording time can be given.
- It combines automatic search and numerical evaluation with visual comparison by showing the curves of top results. Visual comparison is more flexible and effective than numerical measurement in evaluating ENF matching (Kajstura et al., 2005; Huijbregtse and Geradts, 2009; Baksteen, 2015). Besides, for applications in forensic proofs, human experts should be involved to conduct the final judgement of the estimated time. Therefore, instead of delivering a single numerical similarity value for affirming the possible recording time, the Top-n comparisons are apparently more informative by showing the Top-n retrieved reference segments. Examples of the Top-n comparisons are referred to Section below.

Experiments and analysis

In the above section, we studied the phenomenon of ENF pattern duplication in the reference data and proposed to check pattern uniqueness for test ENF signals collected in daily life. In this section, we experiment with test audio recordings from the LESS dataset.

Experimental setup

For a given test ENF sequence, we search in the reference dataset for its optimal matches. The beginning time of a matched reference segment is regarded as a candidate of recording time for the test audio. Compared to its actual recording time, an estimation within a shift of 1 min is considered as correct, i.e., a tolerance window of 120 s with the actual time in the middle. This setting is in accordance with the human habit that people usually note time up to minutes.

We adopt the Top-n error as the main evaluation criterion. The Top-n error rate indicates the fraction of failure estimations that the target actual time is not in the Top-n retrieved results. In this work, we report the Top-1 and Top-3 errors as our experimental results, where the Top-1 error is always larger than the Top-3 error. Concerning the potential use in forensic applications, we also report the *precision* and *recall* (Xie et al., 2012) as additional performance evaluation. They are defined as

$$\begin{aligned} precision &= \frac{N_{cor}}{N_{ret}} \\ recall &= \frac{N_{cor}}{N_{test}} \end{aligned} \tag{6}$$

where N_{test} is the number of test samples, i.e., 187 for our experiment, N_{ret} is the number of estimations returned by the algorithm, and N_{cor} is the number of correct estimations. Without uniqueness examination, N_{ret} is equal to N_{test} so that precision and recall are actually the same, equal to one minus the Top-1 error. Forensic applications require high precision to ensure high confident judgement in practice. This is also one of the important reasons that the strategy of uniqueness examination is introduced.

Experimental results

The 187 test ENF sequences from the LESS dataset are compared to a long ENF reference first. Although all the test audio recordings were collected from 30 June, 2016 to 24 August, 2016, we select a large searching scope from 2, January, 2016 to 28 August, 2016 (240 days) in order to show the capability of large-scale searching. This plays an important role in practice when little information about the possible recording time is known. For the other determinant factor besides the searching scope, i.e., the similarity threshold θ, we tune it from 0.001 to 0.01.

Table 5 summarized the estimation results with respect to different similarity thresholds. One can see that the lowest Top-1 error of 3.21% (=6/187) is achieved by a proper threshold within the closed range [0.005, 0.007]. The error rates with θ equal to 0.001 and 0.010 are coincidentally the same but the respective error samples are actually different.

Comparison with the state-of-the-art

We name the proposed method as maximum bSim for short and compare it with three state-of-the-art approaches, the dynamic matching algorithm (DMA) (Hua et al., 2014), the minimum MSE and the maximum CC (Huijbregtse and Geradts, 2009; Kantardjiev, 2011; Baksteen, 2015). This comparison experiment is carried out with a narrow searching scope of 26 h because the DMA approach runs much more slowly than the others and thus is incapable of large-scale search. The setting of 26 h is according to the assumption of knowing the day of recording of each test audio and the consideration of overlapping search between adjacent days, i.e., 24 h in a day plus 2 h, each before and after the day. The similarity threshold θ is set to 0.005 according to the previous experiment.

Table 6 summarizes the comparison results in terms of the estimation error and average searching time. One can see that the proposed method succeeds in improving both effectiveness and efficiency over the state-of-the-art. Comparing the two baselines, the maximum CC obtains comparable error rate to the minimum MSE, with higher computation cost, i.e., from 1.4649 s to 1.9698 s. The recent prior art, DMA (Hua et al., 2014), reduces the estimation error by a significant relative reduction of 9.53% (from 22.46% to 20.32%), however, at the expense of a remarkable searching time increase. In contrast, the proposed minimum bSim not only

Table 5
Experimental results of time-of-recording estimation on the LESS dataset.

Threshold θ	0.001	0.003	0.005	0.007	0.009	0.010
Top-1 error	4.81%	3.74%	**3.21%**	**3.21%**	3.74%	4.81%
Top-3 error	4.28%	**3.21%**	**3.21%**	**3.21%**	**3.21%**	4.28%

Experimental results of time-of-recording estimation on the LESS dataset. The lowest error rates are highlighted in bold.

Table 6
Comparison of maximum bSim with other state-of-the-art methods on the LESS dataset. The average searching time is reported in seconds. Max.: maximum; min.: minimum.

Approaches	Min. MSE (baseline 1)	Max. CC (baseline 2)	DMA (Hua et al., 2014)	Max. bSim (this work)
Top-1 error	22.46%	22.46%	20.32%	**2.67%**
Searching time	1.4649	1.9698	41.0444	**0.8973**

Comparison of maximum bSim with other state-of-the-art methods on the LESS dataset. The average searching time is reported in seconds. Max.: maximum; min.: minimum. The lowest error rate and searching time are highlighted in bold.

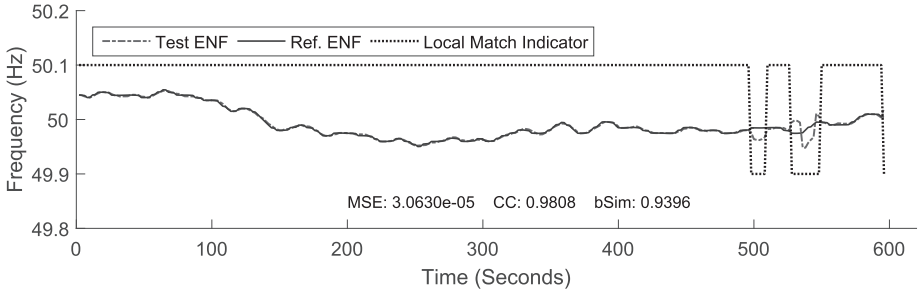
achieves the lowest estimation error rate (i.e., a relative reduction of 86.86% with respect to the DMA result), but also is the fastest approach in the ENF pattern search (i.e., even faster than the minimum MSE baseline). The benefits of the accurate and efficient performance come from the proposed process of binarization. Compared with DMA that adopts the median filter, binarizing the local similarity value at each frame is more useful for removing the influence of large noise.

Comparing the maximum bSim results in Tables 5 and 6, we can know that narrowing the searching scope from 240 days to 26 h also helps to ensure the pattern uniqueness and result in a smaller error rate (3.21% versus 2.67%). In addition, the searching time of the proposed method linearly increases with the length of searching scope.

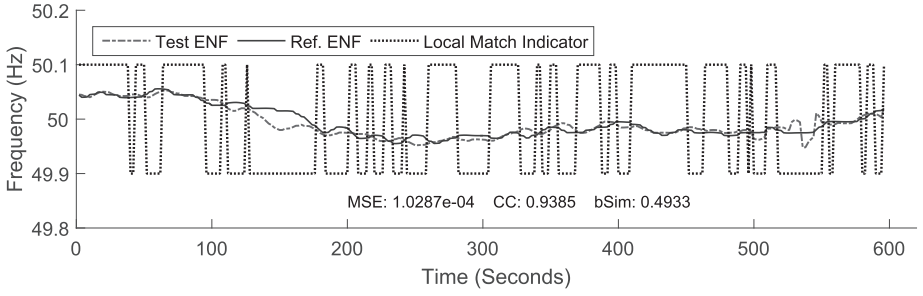
Importance of uniqueness examination

With these achievements, the proposed method is able to perform efficient and quick search for the time estimation task. However, since there is still an error rate of 3.21%, it is difficult to know the true positives, i.e., to select out the correct estimations. The precision, i.e., the positive predictive value, is as high as 96.79% but still not 100%. The top one retrieved result can not be simply trusted. Especially in practical application of forensics concerns, relying on such automatic-made decisions may cause moral and ethical issues (G. O. for Science).

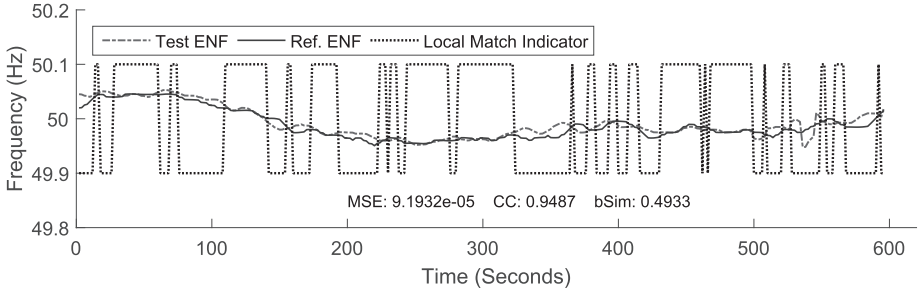
Therefore, instead of simply accepting the top one retrieved result, the uniqueness examination of Top-*n* retrieval is recommended. An example is given in Fig. 9, on the same test audio



(a) Reference ENF (10 minutes) @ 2016-08-23 23:20:54



(b) Reference ENF (10 minutes) @ 2016-04-13 23:49:05



(c) Reference ENF (10 minutes) @ 2016-04-22 18:46:24

Fig. 9. Uniqueness examination for the test ENF (10 min) @ 2016-08-23 23:20:52. The test ENF is compared to its Top-3 retrieved reference ENF segments within a large searching scope of 240 days.

recording of 10 min in Figs. 2–4. We compare the test ENF signal to its Top-3 most similar segments from the reference data, shown in Fig. 9(a) and (b) and 9(c), respectively. The criteria of the uniqueness examination can be summarized as following.

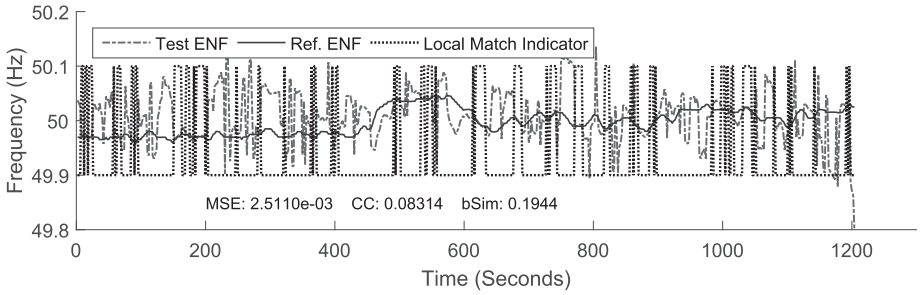
1) The Top-1 match presents better visual matching between the test and reference signals, i.e., having the least visible mismatching;
2) The Top-1 match has a significant lager bSim value than the other two, i.e., 0.9396 versus 0.4933;
3) The Top-1 match has the longest consecutively matched fragment, i.e., consecutive 1s on the indicator line.

With these evidences, a human examiner is able to determine that the start time of the Top-1 result is the most probable recording time of the test audio. Actually, this estimated time has only a small time shift of 2 s to the ground truth. A negative example in Fig. 10 illustrates that the top three matches having approximate similarity values indicates useless test ENF, i.e., the test audio did not capture the ENF pattern in the local power grid when the audio was recorded.

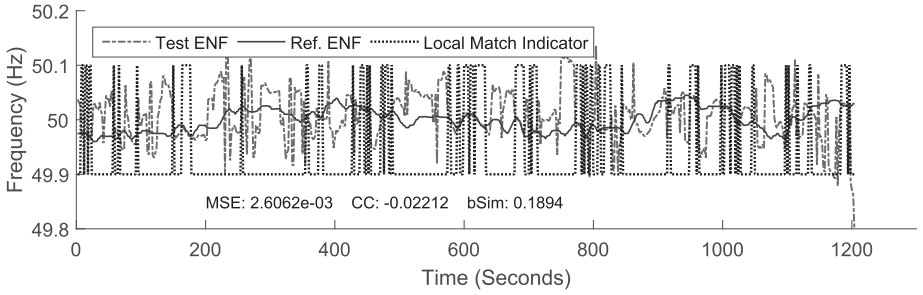
The procedure of uniqueness examination is conducted by human examiners to filter uncertain decisions and ensure that the passed decisions are correct, i.e., the precision is 100%. Formally, we denote the Top-3 retrieved similarity scores as $bSim_1$, $bSim_2$ and $bSim_3$, respectively. According to the second examination criterion that the Top-1 match should have a significant lager bSim value than the other two, we define the *significance gap* as

$$\begin{aligned} sg &= ||bSim_1 - bSim_2|| + ||bSim_1 - bSim_3|| \\ &= 2bSim_1 - bSim_2 - bSim_3. \end{aligned} \tag{7}$$

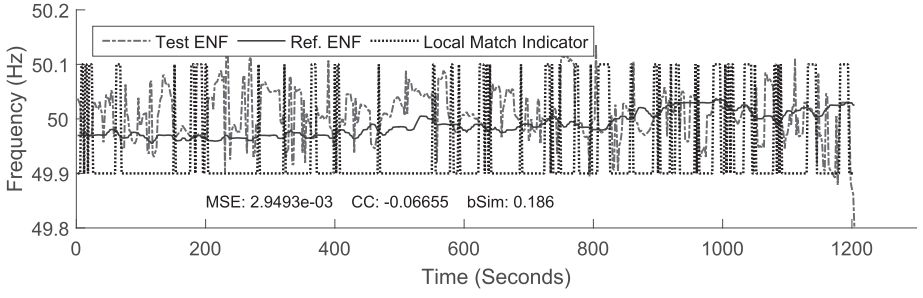
Specifically, a confident test result should have a large significance gap so that it can pass the examination, otherwise it will be marked as an uncertain decision and seek for other means to estimate the recording time. We tune the significance gap from 0 to 0.4 and find that the precision reaches 100% when the significance gap is larger or equal to 0.07, with the recall as high as 94.65% (see



(a) Reference ENF (20 minutes) @ 2016-08-09 12:19:08



(b) Reference ENF (20 minutes) @ 2016-05-01 09:36:35



(c) Reference ENF (20 minutes) @ 2016-04-07 11:35:42

Fig. 10. Uniqueness examination for the test ENF (20 min) @ 2016-08-02 20:30:13. The test ENF is compared to its Top-3 retrieved reference ENF segments within a large searching scope of 240 days. **The Top-1 match does not catch the valid ENF pattern at all**. Uniqueness examination evidences: (1) The Top-1 match presents inferior visual matching between the test and reference signals; (2) the top three matches have approximate bSim values; (3) none of the top three matches has long consecutively matched fragment on the indicator line. Actually, none of the three estimated times is the correct recording time of the test audio.

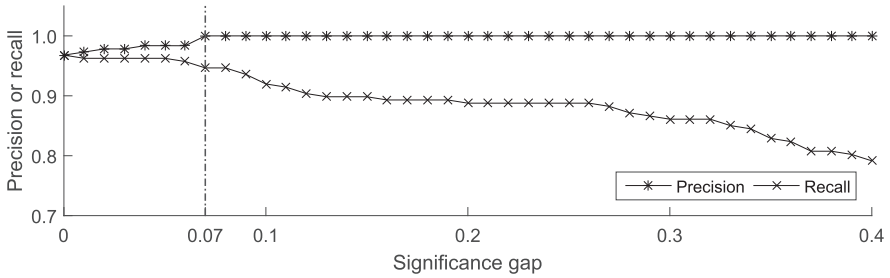


Fig. 11. With the significance gap set to 0.07, the precision reaches 100% and the recall remains as high as 94.65%.

Fig. 11). In other words, although the uniqueness examination is conducted by human examiners, the criteria of the examination strategy, e.g., by setting a large significance gap, are certain to eliminate human bias and output right decisions only.

Conclusions

In this paper, inspired by the principle of visual comparison, we proposed the bSim (bitwise similarity) for measuring ENF matching in the task of time-of-recording estimation. Through experimental evaluation, we demonstrated that the bSim is more effective and efficient than the two classical and general measurements, MSE and CC. The proposed method also goes beyond the state-of-the-art DMA algorithm by significant improvement, i.e., a relative error rate decrease of 86.86% (from 20.32% to 2.67%) and a speedup of 45× faster search response (41.0444 s versus 0.8973 s).

Moreover, although we have provided a much more precise solution for the task of time estimation, we pointed out the importance of human examination in forensics applications and proposed a novel examination strategy to check the uniqueness of targeted ENF pattern by visually comparing the Top-*n* retrieved results. This strategy provides details of pattern matching to human examiners and helps to filter the failures in ENF pattern collection, i.e., an audio recording may not capture the valid ENF pattern from the local electrical power grid (see Fig. 1(b)). The experimental validation was carried out on our own collection of ENF signals in Singapore, the LESS dataset in Section Labelled ENF Signals in Singapore. The proposed system was demonstrated to have addressed the problem of ENF pattern matching in the task of time-of-recording estimation, and future effort can be made to analyze the environment conditions when collecting audio recordings and then improve the collected test ENF quality.

Acknowledgement

This research work is supported by the Singapore Police Force, Ministry of Home Affairs, under the research grant CA/20150428/003.

References

Baksteen, T., 2015. The Electrical Network Frequency Criterion: Determining the Time and Location of Digital Recordings. Master's thesis. Delft University of Technology.

Bengio, Y., Paiement, J.-F., Vincent, P., Delalleau, O., Le Roux, N., Ouimet, M., 2004. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. Adv. Neural Inf. Process. Syst. 16, 177–184.

Bykhovsky, D., Cohen, A., 2013. Electrical network frequency (ENF) maximum-likelihood estimation via a multitone harmonic model. IEEE Trans. Inf. Forensics Secur. 8 (5), 744–753.

Chai, J., Liu, F., Yuan, Z., Conners, R.W., Liu, Y., 2013. Source of ENF in battery-powered digital recordings. In: Audio Engineering Society Convention 135. Audio Engineering Society.

Cooper, A.J., 2009. An automated approach to the electric network frequency (ENF) criterion: theory and practice. Int. J. Speech, Lang. Law 16 (2). http://103.4.94.107:8080/xmlui/bitstream/handle/123456789/15905/An%20automated%20approach%20(1).pdf?sequence=1.

Energy market authority of singapore, https://www.ema.gov.sg/System_Planning.aspx, [Accessed 3 November 2016].

Esquef, P.A.A., Apolinário, J.A., Biscainho, L.W., 2014. Edit detection in speech recordings via instantaneous electric network frequency variations. IEEE Trans. Inf. Forensics Secur. 9 (12), 2314–2326.

G. O. for Science, Artificial intelligence: opportunities and implications for the future of decision making.

Garg, R., Varna, A.L., Wu, M., 2012. Modeling and analysis of electric network frequency signal for timestamp verification. In: 2012 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, pp. 67–72.

Garg, R., Varna, A.L., Hajj-Ahmad, A., Wu, M., 2013. "Seeing" ENF: power-signature-based timestamp for digital multimedia via optical sensing and signal processing. IEEE Trans. Inf. Forensics Secur. 8 (9), 1417–1432.

Grigoras, C., 2007. Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis. Forensic Sci. Int. 167 (2), 136–145.

Hajj-Ahmad, A., Garg, R., Wu, M., 2013. Spectrum combining for ENF signal estimation. IEEE Signal Process. Lett. 20 (9), 885–888.

Hua, G., Goh, J., Thing, V.L.L., 2014. A dynamic matching algorithm for audio timestamp identification using the ENF criterion. IEEE Trans. Inf. Forensics Secur. 9 (7), 1045–1055.

Hua, G., Zhang, Y., Goh, J., Thing, V.L., 2016. Audio authentication by exploring the absolute-error-map of enf signals. IEEE Trans. Inf. Forensics Secur. 11 (5), 1003–1016.

Huijbregtse, M., Geradts, Z., 2009. Using the enf criterion for determining the time of recording of short digital audio recordings. In: International Workshop on Computational Forensics. Springer, pp. 116–124.

Kajstura, M., Trawinska, A., Hebenstreit, J., 2005. Application of the electrical network frequency (ENF) criterion: a case of a digital recording. Forensic Sci. Int. 155 (2), 165–171.

Kantardjiev, A., 2011. Determining the Recording Time of Digital Media by Using the Electric Network Frequency. Master's thesis. Uppsala University.

Singapore power group, http://www.singaporepower.com.sg, [Accessed 3 November 2016].

Von Luxburg, U., 2007. A tutorial on spectral clustering. Stat. Comput. 17 (4), 395–416.

Xie, L., Zheng, L., Liu, Z., Zhang, Y., 2012. Laplacian eigenmaps for automatic story segmentation of broadcast news. IEEE Trans. Audio, Speech, Lang. Process. 20 (1), 276–289.

Zheng, L., Leung, C.-C., Xie, L., Ma, B., Li, H., 2012. Acoustic texttiling for story segmentation of spoken documents. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5121–5124.