ENF ANALYSIS ON RECAPTURED AUDIO RECORDINGS

Hui Su, Ravi Garg, Adi Hajj-Ahmad, and Min Wu
{hsu, ravig, adiha, minwu}@umd.edu
University of Maryland, College Park

ABSTRACT

Electric Network Frequency (ENF) based forensic analysis is a promising tool for timestamp authentication and forgery detection in such multimedia recordings as audios and videos. ENF signal is embedded in an audio recording due to electromagnetic interference from the power lines. The time of creation of a multimedia recording can be determined by comparing the ENF signal embedded in the recording with a reference ENF database collected from the power grid. In this paper, we conduct a study of the effect of recapturing of audio recordings on the ENF embedding. We demonstrate that recaptured audio recordings pick up two ENF signals: the content ENF signal which is inherited from the original audio recording; and the recapturing ENF signal which is embedded from the recapturing process. Conventional ENF signal extraction techniques on such recordings may fail when the two ENF signals are at the same nominal value. A decorrelation algorithm is proposed to extract the content ENF signal and the recapturing ENF signal. The experimental results show the effectiveness of the proposed method in the estimation of both the ENF signals.

Index Terms— Electric Network Frequency (ENF), Recaptured audio recordings, Frequency estimation.

1. INTRODUCTION

Today, an increasing amount of digital information is available in the form of audio, video, and other sensor recordings. These recordings are often stored with the metadata that describes such important information as the time-of-recording and the location of recording. With digital tools, it has become easy to modify this information. When media recordings are used in critical applications such as law enforcement scenarios, the authentication of such digital recordings including metadata is important. As time and location information in metadata is easy to be tampered, an emerging direction exploits the Electric Network Frequency (ENF) to perform authentication [1, 2, 3].

ENF is the supply frequency of power distribution networks in a power grid. The nominal value of the ENF is usually 60 Hz or 50 Hz, and it fluctuates slightly around its nominal value with time. The main trends in the fluctuations have been shown to be similar within the same power grid. Multimedia recordings created using devices plugged into the power mains or located near power sources can often pick up the ENF signals: due to electromagnetic interference or acoustic vibrations into audio [1] and due to invisible flicking in indoor lightings into videos [4]. It has been shown that the ENF signals extracted from the audio recordings exhibit a high correlation with the ENF extracted from the power mains signal at the corresponding time. By comparing these two ENF signals, the time of creation of the recording can be determined. ENF signal extracted from an audio has also been used to detect regions of tampering by analyzing its phase continuity [2]. Recently, improved methods for ENF signal estimation and ENF signal matching for media timestamping have also been proposed [5, 6, 7].

As ENF signal is embedded into multimedia recordings at the time of recording, several interesting questions arise about the ENF traces in recaptured audio recordings. If the recapturing of the recording is conducted in the region of the same nominal ENF as the original recording, the ENF traces due to the two recording processes may overlap with each other. How will such overlap affect the quality of the ENF signal extraction? ENF signals in recaptured audio recordings may contain two components: one is inherited from the original recording, referred to as the *content ENF* signal; and the other is embedded during recapturing process, referred to as the *recapturing ENF* signal. The content and the recapturing ENF signal may have different energies; signal with a higher energy is referred to as the *dominant ENF* and that with a lower energy as the *latent ENF*.

The question of ENF extraction in recaptured audio is also relevant with analyzing recordings of historical importance. For example, such historical recordings as NASA Apollo lunar mission audio recordings [8] and President Kennedy's White House conversations [9] were conducted in the analog era of 1960's. These recordings were recently digitized and made available online. Several interesting tasks can be done using such recordings. For example, multiple channels of NASA Apollo mission recordings can be used to create a time synchronized exhibit of the mission. As an ENF signal is time-varying, it can potentially be used to automatically align multiple audio recordings archived from such historical events. However, due to the digitization process, the recordings available online may also have been affected by the ENF signals corresponding to the time of digitization. No prior work has addressed the effect of recapturing of audio recordings on ENF signals.

As will be shown later in the paper, conventional ENF estimation techniques can only extract the dominant ENF signal. This observation motivates us to design algorithms to extract both the dominant and the latent ENF signals from recaptured recordings. As audio recapturing can also be used as an "anti-forensic" strategy by an adversary to alter the ENF traces to mislead a forensic examiner, developing techniques to extract multiple overlapping ENF signals may also complement the existing techniques to counter such anti-forensic operations [10].

In this work, we propose a decorrelation based algorithm to estimate both the dominant and the latent ENF from a recaptured audio. After estimating the dominant ENF using conventional ENF signal estimation techniques, a residual signal is computed by subtracting the estimated dominant ENF signal from the original signal. The latent ENF is then estimated from the residual signal.

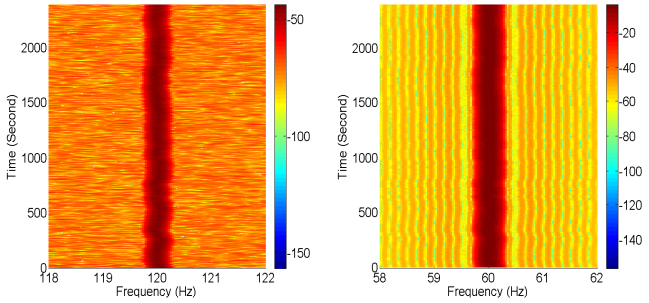
2. ENF SIGNALS IN RECAPTURED AUDIO RECORDINGS

ENF signal is generally present around its nominal value and the higher order harmonics in an audio recording. A simple way to visu-

alize its presence in audio is through spectrogram. The audio signal is divided into overlapping frames, and for each frame a high precision FFT is computed. In Fig.1 (a) and (b) of the spectrograms of an audio signal and the power mains signal recorded at the same time, we observe a strip of time-varying energy at 120 Hz and 60 Hz, respectively. This energy distribution corresponds to the ENF signal in these recordings. There are various ways to estimate the instantaneous frequency, and some comparisons are carried out recently in [7, 6]. We use the weighted energy method [4] as an example technique in this paper for its low complexity. The ENF signal is estimated by computing the dominant instantaneous frequency of each frame around the frequency of interest by:

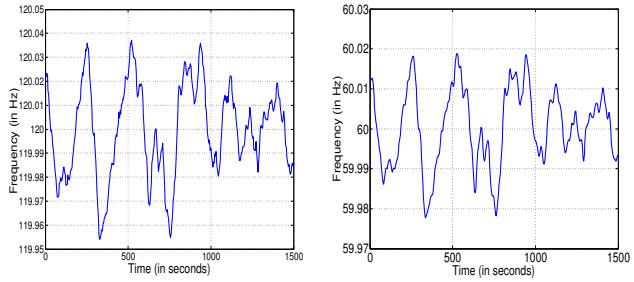
$$F(n) = \frac{\sum_{l=L_1}^{L_2} f(n,l)|s(n,l)|}{\sum_{l=L_1}^{L_2} |s(n,l)|}, \quad (1)$$

where f_s and N_{FFT} are the sampling frequency of the signal and the number of FFT points, respectively; $L_1 = \frac{(f_{ENF}-\Delta f)N_{FFT}}{f_s}$ and $L_2 = \frac{(f_{ENF}+\Delta f)N_{FFT}}{f_s}$; $f(n,l)$ and $s(n,l)$ are the frequency and energy in the l^{th} frequency bin of the n^{th} time frame, respectively.



(a) Spectrogram of an audio signal at 2^{nd} harmonic (120 Hz).

(b) Spectrogram of the corresponding power signal at 60 Hz.



(c) ENF signal estimated from the audio signal.

(d) ENF signal estimated from the power signal.

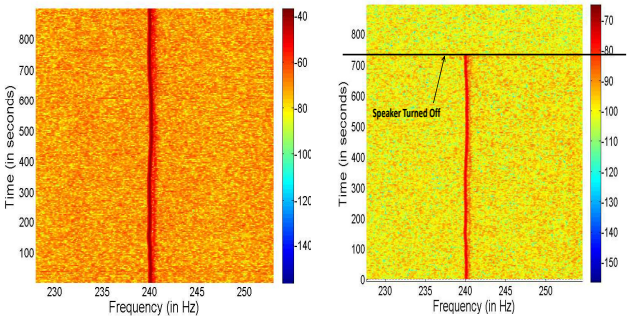
Fig. 1. Spectrograms and ENF estimates from audio and power signals recorded at the same time.

In Fig.1 (c) and (d), we plot the ENF signal estimated from the audio and the power recording. As can be seen from these figures, the audio ENF signal shows very similar fluctuations as the power ENF signal. If the ENF signal is continuously recorded from the power grid to form a reference database, the audio ENF signal can be compared with the database to determine the time of recording using the normalized cross-correlation coefficient [4].

As the aim of this work is to analyze the ENF signals present in a recaptured audio, we conduct the following experiment to test the robustness of ENF signals to a recapturing process. An audio is recorded in an office using a digital recorder. To simulate the conditions of recapturing, we play this recording on a stand-alone speaker in an acoustic anechoic chamber and re-record it using a digital recorder. Fig. 2(a) and (b) show the spectrograms of the original recording and the recaptured recording, respectively. From these figures, we observe that the ENF signal is present at the harmonic frequency of 240 Hz in the original recording and the recaptured recording. High correlation is observed between the ENF signals extracted from the original and recaptured recordings around 240 Hz. When we switch-off the replayed audio, the energy peak present at 240 Hz in the spectrogram of the recaptured audio recording disappears. This happens because no interference is present from power lines at 240 Hz in the acoustic chamber. In this example, the content ENF signals and the recapturing ENF signals are not interfering with each other.

As another example, we show in Fig 3 (a) the spectrogram of a historical recording from the President Kennedy's White House conversations, which are available at [9]. This recording was conducted in 1962 on an analog tape and digitized later. From this figure, we observe that two different ENF signals are present near 240 Hz, and one of them (present around 239 Hz) disappears well before the end of audio. After listening to the audio, we note that the original recording is turned off at this time. We conjecture that the 239 Hz signal is the content ENF signal and the 240 Hz signal is the recapturing ENF signal.

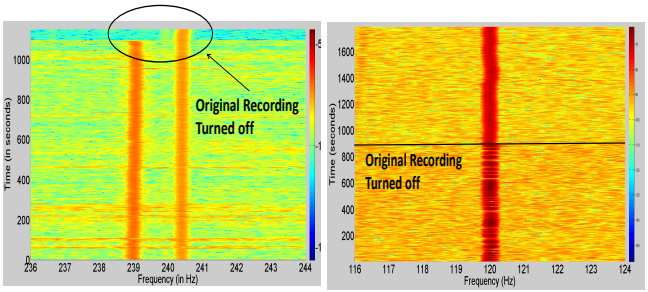
The two examples have demonstrated the case when the content ENF signal and the recapturing ENF signal are non-overlapping. From such recordings, both the ENF signal can be extracted easily by using suitable bandpass filters followed by conventional ENF estimation techniques around the frequency of interest. In less favorable cases, however, the content ENF signal and the recapturing ENF signal may overlap and interfere with each other. To illustrate this scenario, we conduct a recording in the acoustic chamber and recapture it in the same place. As the ENF signal in the same room is embedded from the electromagnetic influences of the same power sources, the content ENF and the recapturing ENF are overlapping at a frequency of 120 Hz. From the spectrogram of the recaptured audio shown in Fig. 3 (b), we observe that for the duration of the playback of the original audio on the speaker, the content ENF and the recapturing ENF overlap with each other and the energy distribution of the spectrogram appears noisy. After the original audio is switched-off, the ENF signal becomes cleaner as only the recapturing ENF is captured.



(a) Original Audio

(b) Recaptured Audio

Fig. 2. Spectrograms of the original and recaptured recordings.



(a) Digitized Kennedy conversation recording.

(b) Recaptured audio signal with overlapping ENF signals.

Fig. 3. Audio recording spectrograms.

3. ENF ESTIMATION FOR RECAPTURED AUDIO RECORDINGS

3.1. Synthetic Experiments

Conventional ENF signal estimation methods extract dominant frequencies in a narrowband around the frequency of interest (nominal ENF or its harmonics) of a given signal. As the content ENF signal and the recapturing ENF signal in recaptured recordings may be overlapping with each other at the same frequency range, the conventional methods fail to extract both the ENF signals. To demonstrate this, we generate two frequency sequences, $E_d(t)$ and $E_l(t)$, as following:

$$E_d(t) = 60 + N_d(t)$$
$$E_l(t) = 60 + N_l(t),$$

where $N_d(t)$ and $N_l(t)$ are drawn from i.i.d. Gaussian random processes of zero mean and variance 0.1. Using these two signals, we generate a time domain signal that varies according to the frequencies $E_d(t)$ and $E_l(t)$ as follows:

$$s(t) = \cos(2\pi \int_0^t E_d(\tau)d\tau) + \sqrt{\alpha}\cos(2\pi \int_0^t E_l(\tau)d\tau) + N(t), \quad (2)$$

where $N(t)$ is a Gaussian random process of zero mean and unity variance, and α is a constant with $0 \leq \alpha \leq 1$. We see from Eq. (2) that signal $s(t)$ consists of two sinusoids of different amplitudes with $E_d(t)$ and $E_l(t)$ being their instantaneous frequencies at time t. Based on this model of $s(t)$, $E_d(t)$ is the dominant ENF signal and $E_l(t)$ is the latent ENF signal, as the energy of sinusoid corresponding to $E_d(t)$ is greater than $E_l(t)$.

We use the expression for weighted energy frequency estimation given by Eq. (1) to extract the ENF signal from $s(t)$. We compute the normalized cross-correlation (NCC) between the estimated ENF signal and the ground truth frequency sequences $E_d(t)$ and $E_l(t)$, respectively. The experiment is repeated multiple times with different realizations of $N_d(t)$, $N_l(t)$, and $N(t)$. The mean and the variance of the NCC values obtained for different values of α is shown in Fig. 4. From this figure, we observe that when there is a significant difference between the energy of the dominant ENF and the latent ENF, the correlation between the extracted ENF and the dominant ENF is very high (0.6-0.7 range). However, as the energy of the latent ENF signal increases, this correlation value decreases and become very low (< 0.3 for α close to 1). Similar results were obtained for other frequency estimations methods such as the subspace based MUltiple SIgnal Classification (MUSIC) and Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT). Our preliminary results also show that these subspace-based approaches can only obtain reliable estimates when there is a sufficient margin between $E_d(t)$ and $E_l(t)$. This experiment on synthetic data verifies that the conventional ENF estimation techniques fail to extract the overlapped ENF signals, which is usually the case with recaptured audio recordings. In the following subsection, we describe a new algorithm to extract both the dominant and the content ENF from recaptured audio recordings.

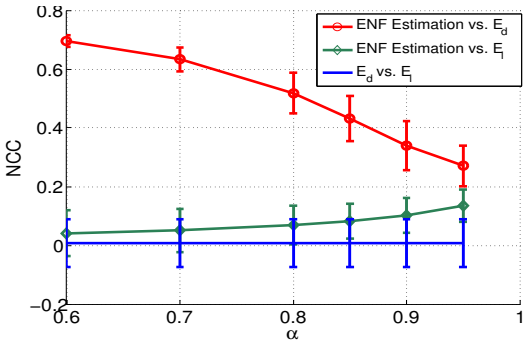


Fig. 4. The mean and the Variance of the NCC values.

3.2. Proposed Approach for ENF Signal Extraction

Our proposed algorithm to extract both the dominant and the latent ENF signals in a recaptured audio works in two stages: the dominant ENF is first estimated, followed by the estimation of the latent one. As discussed in Section 3.1, the dominant ENF, denoted by $E_d(t)$, in an audio s(t) can be estimated using conventional estimation techniques such as the weighted energy method. After estimating $E_d(t)$, we match it with the reference ENF database from the power grid to estimate the time of the dominant ENF signal embedding into the recording. We then subtract the power signal corresponding to the time of recording of the dominant ENF signal from the audio recording. As the magnitude of power measurements and the actual embedding in the audio may be different, the subtraction is done by estimating the appropriate scaling factor of the magnitude that makes the ENF signal of resulting audio signal $\hat{s}(t)$ maximally decorrelated with the ENF signal of the power recording corresponding to the time of dominant ENF embedding estimated before. More specifically, we have:

$$\hat{s}(t) = s(t) - \hat{a} \cdot P(t), \quad with$$
$$\hat{a} = \underset{a}{argmin} \left\{ corr(ENF(s(t) - aP(t)), ENF(P(t))) \right\}, \quad (3)$$

where $P(t)$ is the power measurement signal at time t and $\hat{a}$ is the estimated magnitude of the power. $ENF(\cdot)$ denotes the weighted frequency estimation function. As can be understood from the equation, the selection of $\hat{a}$ is essentially to search for the relative amplitude of the dominant ENF signal in the audio signal, with respect to the power signal. Ideally, after the decorrelation process, the resulting signal $\hat{s}(t)$ is free from the traces of $E_d(t)$. The ENF signal that is left in $\hat{s}(t)$ would come from the latent ENF signal, $E_l(t)$. So we estimate it using again the weighted frequency estimation approach on $\hat{s}(t)$.

To show the effectiveness of the proposed algorithm in extracting the dominant and the latent ENF, we conduct experiments on audio data. An audio recording was made in an acoustic anechoic chamber and recaptured later in the same place. The power measurements were also recorded during both the original recording and the recapturing process. The content ENF and the recapturing ENF signals are present around 120 Hz in this case. The ENF extracted directly from the recaptured audio signal shows similar fluctuations with the ENF signal estimated from the power signal at the time of the original recording (NCC 0.62), as can be seen from Fig. 5. So the content ENF signal is the dominant signal in this case. We then decorrelate the recaptured audio by subtracting the estimated dominant ENF signal as discussed previously. Since the ENF signal in the power measurement recording is centered around 60 Hz, we transfer it to 120 Hz by squaring the power signal and then feeding it into a bandpass filter with a narrow passband around 120 Hz. The processed power signal is used for decorrelation as in (3). The ENF signal estimated from the decorrelated audio signal shows high correlation with the ENF signal extracted from the power measurements at the time of recapturing (NCC 0.68). Both the content ENF and recapturing ENF are now successfully extracted from the recaptured audio recording using the proposed decorrelation method.

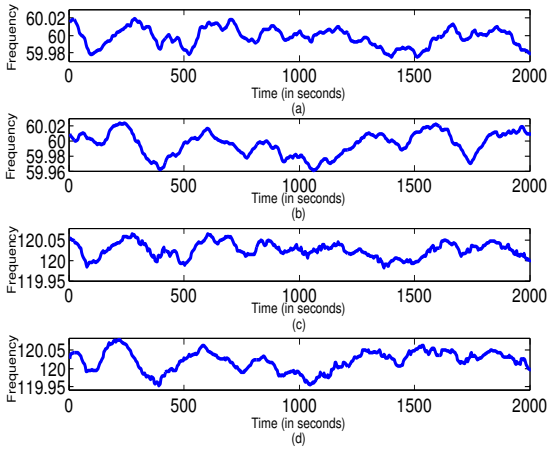


Fig. 5. ENF fluctuations as a function of time. (a) the ENF from the power measurement signal at the time of the original recording; (b) the ENF rom the power measurement signal at the time of recapturing; (c) the dominant ENF estimated from the recaptured audio recording; (d) the ENF estimated from the decorrelated audio signal.

4. APPLICATIONS OF RECAPTURE DETECTION

Under the assumption that the reference measurements from the power grids are available, the proposed decorrelation algorithm can be used for audio recapture detection, i.e., identify whether the given audio recording is original or a recaptured version. As discussed earlier, two ENF signals are embedded in a recaptured audio recording. The ENF signal estimated directly from a recaptured audio recording is the dominant ENF signal E_d. After decorrelation, the latent ENF signal E_l can be extracted from the residue audio signal. When comparing with the reference ENF database measured from power mains, these two ENF signals should match with different segments of the reference database, the time index of which are denoted as T_d and T_l, respectively.

If the recording is original, T_d and T_l are likely to be of similar values. In cases where T_d and T_l are very different, the peak correlation C, that is calculated between the ENF signal estimation from the decorrelated audio signal and the reference database, should be low since it is a false match. Under a hypothesis framework, the H_1 and H_0 cases and the decision rule can be formulated as follows:

$$\begin{cases} H_1: & \text{Test audio is recaptured.} \\ H_0: & \text{Test audio is original.} \end{cases}$$

$$\mathbb{1}(|T_d - T_l| > \delta) \times C \underset{H_0}{\overset{H_1}{\gtrless}} \tau$$

Here $\mathbb{1}(\cdot)$ is an indicator function, and τ is a decision threshold.

We conduct the following experiments to evaluate the proposed audio recapture detection scheme. Audio recordings were made in the acoustic chamber and a conference room. Some of these recordings were then recaptured in the acoustic chamber by playing on a speaker with variant volumes. The total test dataset includes 8.5 hours of original recordings and 16 hours of recaptured ones. The recordings are divided into short clips of 10, 20 and 30 minutes long, and each clip is considered a test sample. We evaluate the false alarm rate and detection rate with different values of τ to obtain the ROC curves as shown in Fig 6. The detection accuracy is higher with longer clips. Specifically, when considering audio clips of 30 minutes, 95% of the recaptured clips are correctly identified without any false alarms.

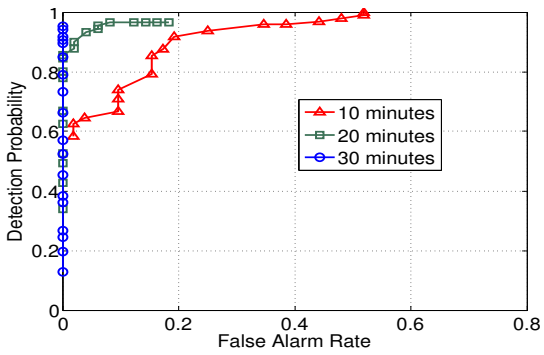


Fig. 6. ROC for audio recapture detection with different clip length.

5. CONCLUSIONS

In this work, we have considered the ENF signal analysis on recaptured audio recordings, which may contain two ENF signals: the content ENF signal that comes from the original recording and the recapturing ENF signal embedded during recapturing process. We have demonstrated such scenarios by conducting experiments on our recorded data and on a historical recording from President Kennedy White House conversations archive. We have also shown that for the cases where the content ENF signal and the recapturing ENF signal are overlapping, conventional ENF estimation techniques can extract only the dominant ENF signal as it has higher energy than the latent ENF signal. For recaptured recordings, we proposed a decorrelation based technique to estimate both the ENF signals. The proposed technique have been shown to successfully extract the dominant ENF and the latent ENF.

Acknowledgement The authors would like to thank Prof. Douglas W. Oard of University of Maryland for enlightening discussions that inspired this work.

6. REFERENCES

[1] C. Grigoras, "Applications of ENF criterion in forensics: Audio, video, computer and telecommunication analysis," *Foresnsic Science International*, vol. 167(2-3), pp. 136–145, April 2007.

[2] D. Rodriguez, J. Apolinario, and L. Biscainho, "Audio authenticity: Detecting ENF discontinuity with high precision phase analysis," *IEEE Transactions on Information Forensics and Security*, vol. 5(3), pp. 534–543, Septemper 2010.

[3] R. W. Sanders, "Digital authenticity using the electric network frequency," in *33rd AES International Conference on Audio Forensics, Theory and Practice*, June 2008.

[4] R. Garg, A. Varna, and M. Wu, "Seeing ENF: natural time stamp for digital video via optical sensing and signal processing," in *19th ACM International Conference on Multimedia*, Nov. 2011.

[5] R. Garg, A. L. Varna, and M. Wu, "Modeling and analysis of electric network frequency signal for timestamp verification," in *IEEE International Workshop on Information Forensics and Security*, Dec. 2012.

[6] A. Hajj-Ahmad, R. Garg, and M. Wu, "Instantaneous frequency estimation and localization for ENF signals," in *APSIPA Annual Summit and Conference*, Dec. 2012.

[7] O. Ojowu, J. Karlsson, J. Li, and Y. Liu, "ENF extraction from digital recordings using adaptive techniques and frequency tracking," *IEEE Transactions on Information Forensics and Security*, vol. 7(4), pp. 1330–1338, August 2012.

[8] "Apollo 11 onboard audio database," http://www.nasa.gov/mission_pages/apollo/40th/a11_audio_db.html.

[9] Miller Center at Univ. of Virginia, "Presidential recordings," http://millercenter.org/scripps/archive/presidentialrecordings/kennedy.

[10] W. H. Chuang, R. Garg, and M. Wu, "How secure are power network signature based time stamps?," in *ACM Conf. on Computer and Communication Security*, Oct. 2012.