I'm Listening to your Location! Inferring User Location with Acoustic Side Channels[*]

Youngbae Jeon
Korea University, South Korea
jyb9443@korea.ac.kr

Minchul Kim
Korea University, South Korea
k.minchul95@gmail.com

Hyunsoo Kim
Korea University, South Korea
aitch25@korea.ac.kr

Hyoungshick Kim
Sungkyunkwan University, South Korea
hyoung@skku.edu

Jun Ho Huh
Samsung Electronics, South Korea
junho.huh@samsung.com

Ji Won Yoon
Korea University, South Korea
jiwon_yoon@korea.ac.kr

ABSTRACT

Electrical network frequency (ENF) signals have common patterns that can be used as signatures for identifying recorded time and location of videos and sound. To enable cost-efficient, reliable and scalable location inference, we created a reference map of ENF signals representing hundreds of locations world wide – extracting real-world ENF signals from online multimedia streaming services (e.g., YouTube and Explore). Based on this reference map of ENF signals, we propose a novel side-channel attack that can identify the physical location of where a target video or sound was recorded or streamed from. Our attack does not require any expensive ENF signal receiver nor any software to be installed on a victim's device – all we need is the recorded video or sound files to perform the attack and they are collected from world wide web. The evaluation results show that our attack can infer the intra-grid location of the recorded audio files with an accuracy of 76% when those files are 5 minutes or longer. We also showed that our proposed attack works well even when video and audio data are processed within a certain distortion range with audio codecs used in real VoIP applications.

CCS CONCEPTS

• **Security and privacy → Privacy protections**;

KEYWORDS

Electrical network frequency, Location tracking, Side channel analysis

ACM Reference Format:
Youngbae Jeon, Minchul Kim, Hyunsoo Kim, Hyoungshick Kim, Jun Ho Huh, and Ji Won Yoon. 2018. I'm Listening to your Location! Inferring User Location with Acoustic Side Channels. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3178876.3186100

1 INTRODUCTION

With the increase in accessibility of high-speed Internet across the world, many VoIP applications that allow people to use voice and video chat online, such as Facebook messenger [15], Skype [2], and WhatsApp [28], have emerged over the years, and become popular. Also, many online streaming services, such as YouTube [55], Facebook Live [14], Twitter's Periscope [48], and Twitch [46], have also become popular.

Such VoIP applications or streaming services, however, may raise privacy concerns. As for VoIP applications, some users, e.g., those engaged in secretive meetings, anonymous reporting or those doing a secret chat in general, have to not only anonymize their identities but also their locations even when they do not perceive the location privacy threat because they are not intentionally sharing their locations. Several previous studies [3, 17] demonstrated that location information can reveal sensitive information about users. Therefore, some services already tried to anonymize or obfuscate a user's actual location. For example, Skype, which is one of the most widely used VoIP applications, recently updated its default application settings to use a proxy server to hide users' IP addresses [32].

Location privacy issues are also prevalent in streaming services. The safety of those broadcasting and hosting live shows at homes may be threatened because stalkers or potentially inappropriate fans could locate their victims, and make physical visits to the victims' private places. Hence, most streaming services might conceal not only content creators' (or broadcasters') IP addresses but also any other location-related information about them. Popular streaming services like Twitch already use an anonymity policy to hide users' network addresses for their privacy [47].

However, researchers have presented various ways of compromising location privacy. PowerSpy [38], for instance, is a technique that can infer a mobile phone's location with the only measurement of the aggregate power consumption of the phone. Furthermore, in another study on Android mobile phone [39], it can also be inferred only by using sensors like gyroscope, accelerometer and magnetometer without requiring any permission.

In this paper, we propose a novel side-channel attack for compromising user location based on a "*Location Inference using SignaTures generated from Electric Network frequencies*" (LISTEN) technique. Unlike previous work [38, 39, 53] that requires the installation of a specific malicious application on a victim's device, the LISTEN attack can be performed with popular VoIP applications or online

[*]This research was supported by the special research grant from Korea university. (grant number: K1711541)

WWW 2018, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5639-8/18/04.
https://doi.org/10.1145/3178876.3186100

streaming services that are already being used. In fact, the only piece needed to perform the attack is a target multimedia file.

To implement the LISTEN attack, an attacker collects electrical network frequency (ENF) signals transmitted from a victim's device via her microphone, and analyzes them to infer the victim's location. ENF is the supply frequency of electrical power in electricity distribution networks. In general, the ENF signals are mostly captured in a particular frequency, either 50Hz or 60Hz. Moreover, the patterns of fluctuations of ENF signals are very similar at time and space because those patterns are highly influenced by the difference between power supply and demand in the same power grid [21]. Since the fluctuations have spatial and temporal characteristics, they can be used as signatures to identify the victim's temporal location [5, 6, 20, 21, 27, 36, 40, 43].

The location identification techniques using the ENF signals have been intensively studied for several years. These researches allow us to figure out which power grid the ENF signals extracted from [25, 26], and also obtain the precise location information within the grid [18, 24]. However, the existing ENF processing techniques [18, 24] are not sufficient to implement the LISTEN attack. In general, they failed to infer geographical location information about a victim's place in real-time. Furthermore, it was not clear how the ENF signals should be well extracted from audio and/or video streaming data used in VoIP applications or streaming services, which is necessary for performing the LISTEN attack in a practical setting.

In our work, we present a novel approach which can handle these matters. We summarize our contributions as follows.

- We proposed a novel location privacy attack to infer a victim's location with the ENF signals extracted from the multimedia streaming data transmitted for VoIP applications or online streaming services. Our ENF signal collection method is much cheaper than existing approaches [7, 31, 33, 52] since we merely collect audio signals from online streaming services that contain the location information for the recorded multimedia streaming data without using any expensive hardware. Also, our attack does not assume any additional malicious application being installed on a victim's device besides VoIP applications or client applications for the target online streaming service.
- We evaluated the performance of the proposed attack in real-world environments. Both theoretical parameters and realistic environments for audio channels were used in the evaluation, showing that our approach provides an accuracy of 90% for inter-grid estimation with 40 minutes long audio, and 76% of intra-grid estimation with 5 minutes long audio.

The rest of the paper is organized as follows. In Section 2, we explain how ENF signals can be obtained from online multimedia stream data and used for location tracking. Section 3 describes the generic attack model, and Section 4 dives deep into the proposed LISTEN attack. Section 5 presents the attack evaluation results, and Section 6 discusses those results. Section 7 summarizes the previous studies related to our work. Our conclusions are in Section 8.

2 BACKGROUND

This section explains the processes involved in extracting ENF signals from multimedia data like audio and video files, and in constructing a ENF map for locations of interest.

2.1 Electrical network frequency (ENF)

ENF is the supply frequency of electrical power in electricity distribution networks. ENF signals are generally embedded in a particular frequency by a stabilizer of power supply systems [22]; either 50Hz or 60Hz frequency is used depending on geographic location. Europe and China use 50Hz for AC current, whereas the United States and Canada use 60Hz. In the real world, however, small fluctuations of ENF signals exist – this is because of the differences that exist between power being supplied and the demand for power at a given moment [21]. Such small variations that exist in ENF signals have been exploited in many application domains including abnormal event detection [7], electrical disturbances [22, 33, 52], and digital forensics [6, 21, 27, 36, 42]. To that end, many researchers have tried various ways to obtain accurate ENF signals.

One way to acquire ENF signals is to use specialized physical electrical devices such as a frequency disturbance recorder (FDR), which is a type of phasor measurement unit used in smart grids [56]. ENF signals can also be obtained from multimedia data such as audio and video files [21, 23, 33, 34, 44, 52] such as Figure 1. This figure demonstrates the spectrogram of an audio file that was recorded in Europe. This spectrogram is obtained using a short time frequency transform (STFT) technique to capture non-stationary ENF signals. As shown in this spectrogram, there exists a fluctuation around 100Hz[1].

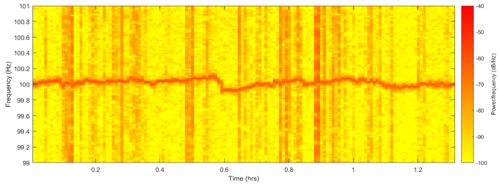

Figure 1: ENF spectrogram around 100Hz

2.2 Extracting ENF signals from multimedia

It is obviously useful if ENF patterns can be obtained from the multimedia data because it does not require any expensive physical devices. However, ENF patterns constructed from side-channels such as audio and video files often have much lower signal-to-noise ratio (SNR) than those directly acquired from an FDR device. Therefore, advanced signal processing techniques are needed to be applied to reduce or remove unwanted noise. The ENF signal extraction process mainly consists of the following four steps.

2.2.1 Decimating and framing. After multimedia data are decimated to 1kHz to save the storage space, we created frames of data sequences, where each frame overlaps with the half of the previous frame. Each frame contains 8192 samples, which comes

[1] ENF's base frequency is 50Hz but we plotted 100Hz because ENF patterns at 50Hz for multimedia data is less clear than those at 100Hz.

to about 8 seconds of decimated audio data in each frame. 4096 samples overlap with each other. This concept comes from the STFT technique.

2.2.2 Applying the quadratic interpolated fast Fourier transform (QIFFT) technique. The next step involves applying the QIFFT technique to each frame. It is necessary to improve the resolution of ENF signal estimation when frame sizes are small [24]. This step is designed to find the maximum value of ENF signals from a given frequency of each frame. A band pass filter is applied to truncate unnecessary frequency ranges from a given frequency domain to obtain the maximum value. We then apply the fast Fourier transformation (FFT) technique to each frame, identifying the index of highest frequency value – this is done by tracing the maximum value, moving from frame to frame. However, in this case, the maximum (peak) spectra value estimation is less precise than resolution estimation so, we apply the Quadratic Interpolation technique when the FFT process is complete [10, 24, 29]. That is, we can search for interpolated peaks on the composed spectra and links them using the QIFFT [1, 9] because the computation of the STFT is too heavy to extract signals quickly from hundreds of multimedia. The sampling rate should be infinite in order to obtain the perfect maximum value, but since it is impossible, we can obtain better estimation by approximating the signal to quadratic formula with using values which are nearby the maximum frequency value.

2.2.3 Enhancing ENF signals using Multi-tone harmonics. Additionally, we can find several horizontal lines in the spectrogram shown in Figure 2, demonstrating similar ENF signal oscillation patterns. Such signals are referred to as *harmonic signals*. More accurate ENF signals can be acquired by processing ENF signals at both base frequencies and harmonic frequencies[2] on the spectrogram [4, 10, 29]. The multi-tone harmonics method uses both a fundamental frequency and harmonic frequencies for exploring the peak position from the ENF spectrum. In this multi-tone harmonics method, the maximum-likelihood estimation technique is applied to the harmonic signals, using Cramer-Rao bound for frequency estimation error to show that the estimation accuracy of ENF signals can be improved by about $10 - 15\mu$Hz [4]. Therefore, to enhance the ENF patterns against the unwanted uncorrelated noise in the frequency domain, multi-tone spectra is obtained by summing all spectrogram at both fundamental and harmonic frequencies. The more harmonic signals we use, the better the accuracy of ENF signal estimation.

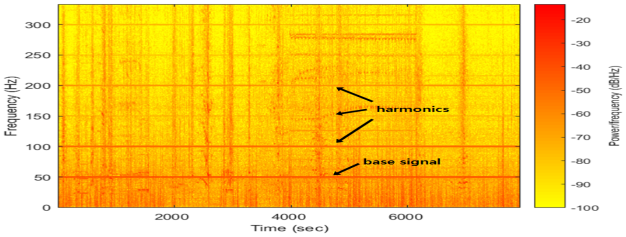

Figure 2: ENF signal at base and harmonic frequencies

2.2.4 Threshold dependent median filter (TDMF). After multi-tone estimation, we use the threshold dependent median filter (TDMF) on the final ENF signal. A median filter is a nonlinear filter that preserves the locality of signals being processed. Median filters, compared to linear mean filter, are more preferred way of reducing noise level. Even if we use both multi-tone estimation and median filter, we will not be able to identify maximum peaks (of ENF signals) if given ENF signals are weak and have relatively low spectra. Such weak ENF signals can be misleading, containing severely abnormal noise levels. To remove abnormal noises, we employ the threshold truncation approach – this approach is called the threshold dependent median filter (TDMF).

2.3 Construction of the ENF map

To construct a comprehensive map that can cover many application domains, it is necessary to collect and process ENF signals from a wide range of online streaming sources. However, building such a large ENF map would require a massive effort and budget. Specialized physical devices such as Frequency Disturbance Recorder (FDR) [52] that can capture ENF signals would need to be purchased, installed, and managed. Deploying and continuously monitoring such physical devices to cover all areas of interest is impractical and expensive.

Recently, an efficient scheme to construct a nation-wise ENF map has been introduced and it does not require purchase and installation of expensive physical devices [23]. We can automatically crawl worldwide ENF signals from online multimedia services such as “EarthCam [13]” and “Ustream [11],” significantly reducing costs, time, and amount of efforts are needed to create a map. However, additional signal processing techniques have to be applied because the ENF signals crawled from online sources are typically less clear.

3 THREAT MODEL

We assume that a service application is installed on the victim’s device equipped with a built-in or attached ENF capture device (e.g., AC microphone). The application has no permission to access GPS or any other location information (e.g., cellular base stations and WiFi APs). The installed application is just used for capturing ENF signals from the victim’s device and delivering the captured ENF signals to the attacker’s device via the Internet. In this environment, the attacker’s goal is to infer the victim’s geographic location by analyzing the received ENF signals. Such environments appear to be often made in many real-world situations. This is because ENF signals can be extracted from recorded audio and/or video signals when the recording device is *mains-powered* [21] which indicates the status of being connected to a stable electrical power grid. Note that *mains-powered* microphones are still popularly used in multimedia streaming services to improve the sound quality of recorded audio files. For instance, we found that about 36% of Twitch users use mains-powered microphones. Therefore, the attacker can collect the ENF signals generated from the victim’s device if the application can just record the audio and/or video signals at the victim’s device and access the recorded audio signals.

In practice, the victim often shares her own user-created contents with others through audio and video sharing sites (e.g., YouTube, Facebook Live, Twitter’s Periscope, and Twitch) by themselves.

[2] Harmonic ENF signals are captured at frequencies that are calculated by multiplying integer by base frequency [4].

In such situations, ENF signal embedded in audio and/or video signals can simply be downloaded by anyone including the attacker. Moreover, if the attacker communicates with the victim using a VoIP application, the attacker can naturally record the victim's audio and/or video signals and receive them without requiring any special permission on the victim's device.

We note that our attack scenarios are likely to apply even when network identifiers such as IP address are hidden from the attacker through an anonymous system (e.g., Tor network [37]) because the attacker does not require additional information from the victim, besides the transmitted recorded audio and/or video signals.

4 LISTEN ATTACK

Because ENF signals could be used as a spatio-temporal signature, the primary goal of the LISTEN attack is to identify the location of a victim's device with access to just ENF signals (side channel information). The attack mainly consists of three sequential processes: (1) **Construction of the ENF map using online streaming data on the Internet**, (2) **Extraction of reliable ENF signals from a target device**, and (3) **Location estimation**.

4.1 Construction of the ENF map using online streaming data on the Internet

The first step of LISTEN attack is to crawl and scrap audio streams from a few pre-selected online multimedia services. Audio and video streams from some multimedia services contain recording location information, including latitude and longitude information. We chose "EarthCam" [13], "Explore" [16], and "Skyline" [50] as the three online sources because both audio and video data were produced with devices which are mains-powered in Alternating current (AC).

The second step is to perform a series of signal processing techniques [23]: (1) checking whether scraped audio streams contain ENF signals, (2) extracting clear signals through noise reduction, (3) aligning incomplete and partial signals on a given time domain using signal alignment techniques, and (4) interpolating ENF signals in uncovered areas with collected neighboring ENF signals. We refer to collected ENF signals as *anchor nodes*, and use them as the sources for interpolation. This step allows us to infer precise locations from a victim's ENF signals by comparing them against interpolated ENF signals. The effectiveness of an interpolated ENF map can be explained theoretically from the fact that the ENF disturbance propagation speed is finite [18, 24] and was demonstrated using the Inverse Distance Weighted (IDW) interpolation technique [23, 35].

4.2 Extracting ENF signals from a victim when VoIP services are used.

After creating the ENF map, the next process is to set a target victim, and extract ENF signals from the victim's device or recorded voice. This process is similar to the way the ENF signals are collected in Section 2.2 but requires more sophisticated algorithms due to various communication systems and environments that hlneeds to be considered. For instance, the victim could be using a VoIP service that streams unreliable ENF signals as shown in Figure 3. Such signals could be distorted and carry a significant level of noise. Severely distorted ENF signals cannot be used for location estimation. Hence, the state of the audio channel need to be specified concretely based on the "frequency response," "time delay," "delay jitter," and "packet loss." These represent the quantified metrics for evaluating the quality of audio channels. In following paragraphs, we describe those metrics and the techniques how attackers mitigate those problems.

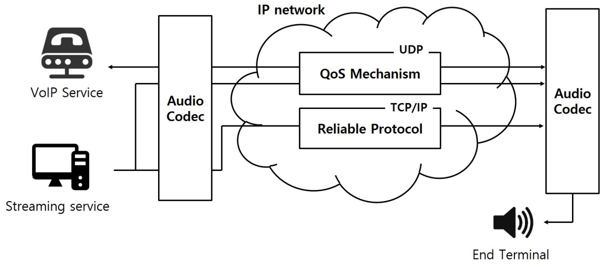

Figure 3: Architecture for audio streaming over IP network for VoIP applications and online streaming services. Reliable data for constructing a ENF map is collected from an online streaming service; sound recorded from a victim's device is received on an audio channel where packet loss may exist.

4.2.1 Frequency response. Audio recorded from the victim's device can be filtered or amplified when it passes through an audio channel. ENF signals cannot be constructed from such an audio file if (for some reason) the victim's ENF signals are deleted. As human audible frequency ranges from 20Hz to 20kHz, many audio codec standards include a band pass filter for better compression and higher quality given a limited data rate [30]. For example, in the case of Skype, the VoIP application uses its own codec called SILK [51]. The compression process of SILK uses a high pass filter for which the cut-off frequency is 70Hz [51]. Since the base frequency of ENF signal is 50 or 60Hz, SILK will filter it. To resolve this, we use the multi-tone estimation [45] shown in Section 2.2. We use harmonic signals with frequency of either 100Hz for the 1st frequency 50Hz or 120Hz for the 1st frequency 60Hz, or above.

4.2.2 Time delay and delay jitter. Since we compare the victim's ENF signals against the ENF map based on known locations (signatures), we need to know the exact time of ENF signal extraction. Hence, any time delay is integral and needs to be known. If a VoIP uses a signaling protocol that provides the exact time delay information, the recorded time can be obtained easily. However, there might be some cases that the exact time is hard to find. In such cases, we have to estimate the time delay by calculating normalized correlation coefficient of extracted ENF signals from a target node and those from anchor nodes. At the exact temporal alignment, the cross-correlation coefficient will have the highest value. This calculation must be performed approximately every eight seconds before ENF signals are framed. Here, each frame has 8192 samples.

Jitter, which is packet delay variation, is also one of the metrics of quality of audio channels. Jitter occurs when VoIP delay changes frequently: a sender transmits packets at a regular interval but a receiver receives packets irregularly. It is known that audio codecs in VoIPs or streaming services can reduce jitter [30]. Since this jitter reduction incurs its own time delay, aligning time against time delay is a only concern.

4.2.3 Packet loss. Packet loss is another important factor since ENF signals cannot be constructed with loss of information. If a service uses a reliable protocol, we can request for a 'packet resend' to a server when packet loss is detected. Otherwise, missing data cannot be restored. In particular, real-time voice chat services often use P2P protocols, which are unreliable channels and do not support packet resend. Let us consider a common case where a victim uses a laptop and Wi-Fi connection for voice chatting. As many streaming or VoIP services use UDP for a real-time service, packet loss can occur if Wi-Fi communication channel is unreliable.

According to the survey conducted in [49], packet loss rate for common VoIP users is about 2% or less. To deal with this packet loss problem, empty signals can be estimated by performing linear interpolation between the remaining ENF values in a given frequency domain.

4.3 Two step location estimation

The final process of LISTEN is to infer the victim's hidden location. This process consists of two steps: inter-grid estimation and intra-grid estimation. Since it is time consuming and difficult to estimate the exact location from the whole ENF map, we firstly apply inter-grid estimation to select candidate power grids. After choosing a certain power grid, LISTEN attack performs intra-grid estimation which infers the precise location in the selected power grid by matching the ENF map of the power grid and given victim's ENF signal. For more formal definition in a Bayesian framework, we need two random variables, l_A for identifying a specific power grid and l_B for locating a specific position. While l_A is a discrete random variable for $l_A \in \{1, 2, \cdots, G\}$ where G is the number of power grids, l_B is a pair of continuous random variables $l_B = (a, b)$ where a and b are longitude and latitude respectively. Denote by $\mathbf{M}$ the constructed ENF map using online multimedia streaming services. The goal of LISTEN attack is to calculate maximum a posterior (MAP) estimate by $l_B^* = \arg_{l_B} \max\ p(l_B|\mathbf{M}, \theta)$ where θ is a set of model parameters. However ENF patterns are highly correlated with hlthese corresponding power grid, so the posterior can be reformed by $p(l_B|\mathbf{M}, \theta) = \sum_{g=1}^{G} p(l_B, l_A = g|\mathbf{M}, \theta) = \sum_{g=1}^{G} p(l_B|l_A = g, \mathbf{M}, \theta) p(l_A = g|\mathbf{M}, \theta)$ and it can be reformed by

$$p(l_B|\mathbf{M}, \theta) \approx p(l_B|l_A = g^*, \mathbf{M}, \theta) \tag{1}$$

where $g^* = \arg_g \max\ p(l_A = g|\mathbf{M}, \theta)$. Here, note that Equation (1) can be derived because different l_Bs cannot have an identical l_A. Therefore, we have the following two step location estimation by

(1) $l_A^* = \arg_{l_A} \max\ p(l_A|\mathbf{M}, \theta)$ for Inter-grid estimation.
(2) $l_B^* = \arg_{l_B} \max\ p(l_B|l_A = l_A^*, \mathbf{M}, \theta)$ for Intrea-grid estimation.

4.3.1 Inter-grid estimation. Inter-grid estimation is about discovering which power grid collected ENF signals come from. Our assumption for Inter-grid estimation is that oscillation patterns of ENF signals might be similar to each other if two different ENF signals are collected from the same grid.

To localize ENF signals on multiple grids through classification, we applied the distance weighted k-nearest neighbor algorithm. After labeling the collected set of anchor nodes with location information, we determine the k-nearest neighbours with inversely proportioned weights. Here, k is selected based on the number of ENF signals collected to be used as anchor nodes.

4.3.2 Intra-grid estimation. Intra-grid estimation localizes points of the ENF signal captured inside a power grid. Intra-grid estimation is straightforward as every single cell of the ENF map has already been interpolated (see Sections 2.3 and 4.1).

To estimate an internal location from a given power grid, we calculate the Euclidean distance between a time-series sequence of interpolated signals in a single grid and the victim's ENF signal. Comparing to the method that uses the correlation coefficients for the signals [18, 24], Euclidean distance method is a more intuitive way of measuring the similarity of given signals, and takes much less computational time. However, this approach is still useful since it can visibly show an inferred location (see Figure 4). The color map represents the distances between interpolated ENF sequences and the victim's extracted ENF signal sequences. With the similarity measure, the red area of the color map denotes that the interpolated sequences are far away from the extracted sequence, and the yellow area means they are close and it is highly likely for the signal to be extracted from there.

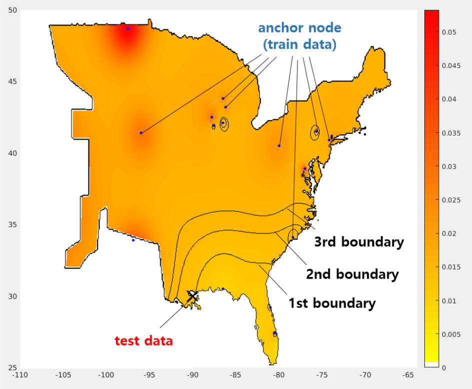

Figure 4: Euclidean distance between target ENF sequence and interpolated sequences in the Eastern power grid of the United States. Cross ('X') dot indicates where the signal actually collected. Red area means it is far from the target signal and yellow area means it is close.

To evaluate the accuracy of the location inference attack, the similarity color map for target region was divided into n parts of equal area where n is the number of ENF signal samples. The term "boundary" is used to separate and distinguish each area. The attack accuracy is defined as the probability that the actual location is included inside the s-th boundary where $s \in \{1, 2, \cdots, n\}$. As the order of boundary s increases, the attack accuracy also increases.

For example, in Figure 4, the cross mark ('X') indicates the location where an ENF signal was captured. If we select the ENF presence boundary by choosing the first out of n highest boundary probabilities ($s = 1$), resulting prediction could be wrong; if we set the boundary by choosing the second highest boundary probability ($s = 2$), resulting prediction is more likely to be right.

5 EVALUATION

This section presents the LISTEN attack performance evaluation results. We calculated the accuracy of inter- and intra-grid estimation (described in Section 4.3.1 and 4.3.2) using three different audio communication environments. In order to conduct this experiment, we first collected the audio stream from the online stream services. Then, we distorted the audio streams by passing them through a virtualized audio channel to mimic real-world communication. Therefore, experiments are categorized based on the following three conditions in the audio channel that were used to distort the stream data:

(1) **Raw audio streams (no distortion)**: This experiment uses raw audio streams directly obtained from online multimedia. That is, the communication channel is perfectly reliable so there is no error and distortion in the audio channel;
(2) **Skype+VPN**: This experiment uses audio streams that are distorted with Skype over a virtual private network (VPN). In this case, the stream data can be affected from unwanted influences such as packet loss, signal removal by filters, and time delay.
(3) **Torfone**: A VoIP application is used over a Tor network. Since Torfone uses its proprietary codec for real time communication, audio streams can often be distorted. Stream data can be affected from unwanted factors such as signal removal by filters and jitter from time delay although Tor network uses TCP protocol. That is, data loss in Torfone occurs through the Torfone's codec, not through Tor networks.

We describe those experimental setups in Section 5.1, and show inter-grid estimation performance and intra-grid estimation performance in Sections 5.3 and 5.4, respectively.

5.1 Experiment setups

5.1.1 PC and software specifications. We used two PCs each equipped with Intel(R) Xeon(R) CPU E5-2609 0 @ 2.40GHz, 64GB RAM, and Ubuntu 16.04.1 LTS (64-bit) operating system. We used Python as the programming language, and a Linux module called "ffmpeg" for scraping and decimating video and audio data from streaming services. MATLAB was used for data analyses.

5.1.2 Dataset used in virtualized audio channels. Virtualized audio channels were used for the three experiments to mimic real-world communications that contain noise. To construct virtualized audio channels, we crawled and scraped audio streaming data directly from online streaming services accessible through the Internet. Those online streaming services are listed in Table 1. We collected a total of 99 audio stream data from Earthcam, Explore and Skyline because their audio stream data contain the exact latitude and longitude information. To stably store and efficiently process the collected stream data, we decimated an hour-long wav extension file to 1,000Hz sound source streams, taking up about 10MB of disk space.

Table 1: Environmental factors of video streaming services. We used audio streams from Earthcam, Skyline and Explore which offer location information. They embed ENF signals with high presence rates.

Service	Categories	ENF presence rate(%)	the number of samples
Earthcam [13]	landscape	85.29	36
Skyline [50]	landscape	95.16	39
Explore [16]	nature	70.59	24

5.1.3 **Skype+VPN** *and* **Torfone**. To measure the effectiveness of the LISTEN attack performed on noisy audio channels, we considered two examples that use unreliable audio channels: **Skype+VPN** and **Torfone**. Environmental conditions for the two channels are shown in Table 2.

Table 2: Environmental condition of Skype and Torfone. Skype's SILK codec works as a high-pass filter whose cut-off frequency is 70Hz. Torfone supports various voice codecs including commonly used GSM.

Application	delay(ms)	codec	packet loss(%)
Skype+VPN	~400	SILK	1.23
Torfone	~2000	GSM	5

Skype, which is one of the most widely used VoIP services, uses peer-to-peer protocols to establish an Internet telephony network. Due to this peer-to-peer characteristic, Skype automatically (by default) reveals the participants' IP addresses to each other. For that reason, people who prefer using Skype anonymously often use location-concealing methods like VPN or Tor. However, since VPN or Tor usually slows down the connection speed, using VoIP over VPN would increase time delay as well. This experiment was designed to test whether ENF signals can be extracted and restored when there are both frequency filter being applied and some time delay.

The other channel we selected is Torfone. Torfone is a VoIP application that uses *onion* domains [3] as IDs, and connects users through Tor networks. Unlike Skype, Torfone offers several voice codec options for users. Torfone supports ADPCM, GSM, Codec2, and other common voice codecs. Among those options, we chose GSM for experiment after considering the popularity of the candidate codecs.

5.1.4 Virtualized audio channels to emulate VoIP client A (Caller) from a remote host. To run experiments reliably, all conditions except the channels that transfer data need to be kept consistent. However, running various channels at the same time would cause unintentional side effects such as increasing packet loss or time delay. In addition, it is difficult to run the experiments again for verification purposes. In such an experimental setup, it would be impossible to regenerate the exact same outside sound again, and there would be a risk of encountering white noises while mimicking

[3] Onion domain is the domain for onion network [19], which enables anonymous communication. This is also called The Onion Rounting (TOR) network.

the environmental conditions. Hence, this kind of experimental setup seems impractical due to this intractability of reproducing the exact same audio sounds in Skype+VPN and Torfone.

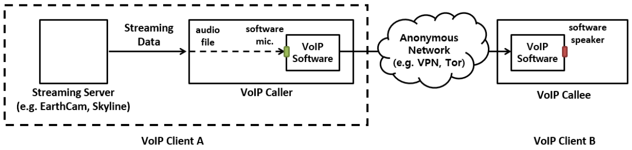

Figure 5: To emulate VoIP client A (Caller) from a remote host, we first obtained the audio data from streaming servers and inputted the audio file directly to VoIP software. To exclude unintended effects from physical devices, all the microphone and speaker operations were processed with virtualization technique. By splitting the experiment into two parts - gathering audio files and actually running VoIP S/W, we can increase the re-producibility without losing details.

To overcome this problem, we used an audio virtualization technique, which redirects sounds (from audio files) to a microphone of a computer. Considering that microphones and speakers work in a similar way, it is possible to redirect an output of a speaker (sound information) to an input of a microphone. Then the microphone would obtain the same sound input data as it would receive the speaker output without any sound loss and white noise. Thus, the caller would perform two subsequent steps. First he or she would receive audio files from multimedia streaming servers, and redirect those files to a VoIP software. Hence, the actual, final experiment design is as shown in Figure 5. This final experimental setup is more effective and efficient than simply putting a speaker next to a microphone, and physically replaying audio files. In such a setup, there are two key risks: (1) we cannot guarantee that sounds from a speaker are fully transferred to a microphone without data loss, (2) noise interference would be inevitable. However, with our audio virtualized environment, there is no risk of noises being added nor risk of losing original information since it inherently prevents physical environments to interfere after recording is done. By simple replaying audio files from the voice sending server, we could easily change the channel conditions without altering the sounds being transferred. All our experiments were conducted using the same condition except the VoIP channel.

5.2 Existence of ENF signals in audio streams

Given the experimental setups described above, we first checked the existence of ENF signals in online streaming data – even though streaming data for VoIP applications is transmitted over noisy audio channels (Skype on VPN, and Torfone on Tor network). This section presents Skype communication results and Torfone results.

People can typically hear 20Hz to 20,000Hz but this does not mean that every frequency in this range constitutes a human voice. Since there are certain frequency regions that mainly constitute daily-life sounds including human voice, many VoIP software apply special filters in sound data to provide better call quality.

To visualize the effect of the filter in the audio channel, Figure 6 plots 1D spectrum in the left sub-figures, and 2D spectrograms in the right sub-figures. Top, center, and bottom sub-figures represent (1) **raw audio streams (no distortion)**, (2) **Skype+VPN**, and (3) **Torfone**, respectively. There are two types of transparently coloured regions. Green regions are located at the base frequency range and purple region are located at the multiple harmonic frequencies respectively.

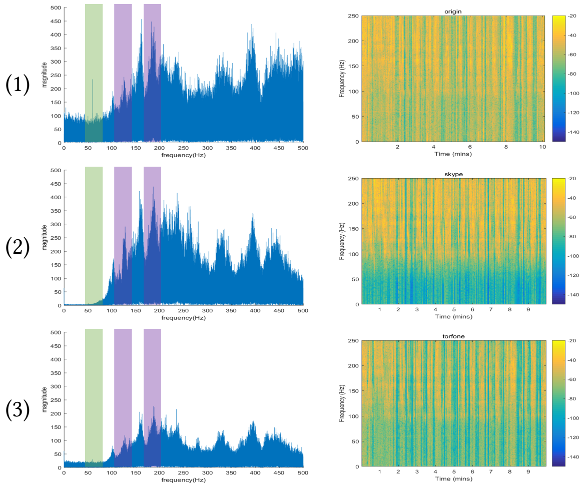

Figure 6: FFT results and spectrograms of the captured audio streaming data. The left figures plot FFT results of the raw audio streams (1), Skype (2), and Torfone (3), respectively; the right figures are their spectrograms. The base ENF signal at the fundamental frequency is not observed (green region) but the harmonics are still visible after passing Skype's (2) audio channel (purple region). Harmonics and base ENF signal are also visible in FFT results of Torfone (3).

As can be seen in Figure 6-(2), ENF signals with Skype are particularly filtered $\sim$ 70Hz frequency region. Since Skype filters out frequency areas lower than 70Hz, base frequency of ENF at 60Hz region are removed and suppressed. That is, LISTEN attack cannot be successful because of the absence of ENF signals at the base frequency. However, we can construct ENF signals by combining and extracting harmonics.

Meanwhile, in the case of Torfone, we can see that base ENF signal also remained around 60Hz as shown in Figure 6-(3) although packet loss exists in case of Torfone such as shown in 2D spectrogram. Since Torfone uses Tor-network for voice chatting, the frequency bandwidth of communication channel cannot digest the bandwidth of audio channel. Therefore, even though Tor-network uses TCP network, Torfone over Tor network frequently drops the lately arrived packets by force in order to provide real-time communication through its own codec.

5.3 Inter-grid estimation

Given a sample data set with annotated region IDs, we infer the power grid ID of a new sample of interest in the inter-grid estimation. For this experiment, we initially extracted 99 audio streams at the same time but only 68 audio streams are finally used because 31 audio streams which do not have ENF signals are removed. These streams were located in 7 power grids, which are Eastern and Western Interconnection of the United States, Central and Northern Power grid of Europe, Brazil, Peru and Cuba. Leave-one-out cross-validation is used to evaluate the inter-grid estimation. We partition the data into a training dataset with 67 streams and a testing dataset with 1 stream. We repeatedly run 68 different runs to obtain sound statistics. With this cross-validation, we measured the accuracy of the classification with varying the length of segment as shown in Figure 7. This figure shows that the accuracy of the experiment has the highest value at 90.77% when the segment length is 40 minutes. As the length of the victim's segment increases from 10 minutes to 40 minutes, the accuracy rates of the inter-grid estimation hardly raise.

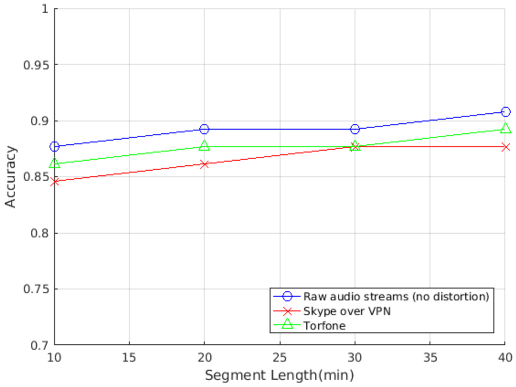

Figure 7: Accuracy vs Segment length (min) for inter-grid estimation with 40-minutes segment length.

In addition, as we mentioned in the previous section, Torfone has harsher environment than Skype over VPN so delay is longer and data loss rate is bigger. However, we find that the performance on using Skype over VPN is worse than one on using Torfone. This result indicates us that the more critical factor for constructing ENF signal against audio channels is the fact that fundamental (1st) ENF signal is filtered by high pass filter in audio codec of Skype [51].

5.4 Intra-grid estimation

Intra-grid evaluation was conducted based on 40-minute sound sources located in the eastern power grid of the United States among the audio streaming data collected from all over the world. 16 audio streams located in the eastern US power grid was collected from online multimedia services: 6 from Explorer, 9 from Skyline, and 1 from Earthcam. In this experiment, they are used to construct a reference ENF map so we name them anchor dataset. Given this reference map, we can infer the location of a new audio streaming data extracted target source through Skype and Torfone of our interest.

For the evaluation, we increase the order of the decision boundary from 1 to n and obtained accuracy is approximately 75% when the order (index) of decision boundary is bigger than 2. The accuracy almost does not increase for three subjects when the index of decision boundary is after 3 until 16. In the case of Eastern power grid of the United States, the area's size in the 3rd boundary is $V = 60,309 miles^2$ since the approximate total area of Eastern power grid is about $341,754 miles^2$. If we calculated an approximate radius R to represent the average distance in the area, we have $R \approx 138.55 mile$ since $V = \pi R^2$.

We also evaluated the accuracy conducted with the varying length of segments. The accuracy rate decreases as the segment length become shorter which is shown in Table 3. However, interesting observation is that the performance improvement in accuracy was not linear, and had step-like changes. We surmise that the tested ENF signals were categorized into three groups depending on geographic origins. 76.4% of samples (group 1) were clearly identified even with a short segment because their origins were located very far from each other. About 6% of samples (group 2) were uniquely distinguished when the segment length is between 30 and 35 minutes - this is because their uniqueness was obtained only with a longer segment. We failed to uniquely identify the remaining samples (group 3) even with a longer 40-minute segment. Much longer ENF signals would be needed to distinguish those samples.

Table 3: Accuracy vs Segment length (min) for intra-grid estimation.

Segment	Accuracy (%)		
length (min)	Raw data	Torfone	Skype + VPN
5	76.4	76.4	70.5
10	76.4	76.4	70.5
15	76.4	76.4	70.5
20	70.5	76.4	70.5
25	76.4	76.4	76.4
30	76.4	76.4	76.4
35	82.3	82.3	82.3
40	82.3	82.3	82.3

6 DISCUSSION

This section discusses defense mechanisms to mitigate the proposed LISTEN attack, and the attack's inherent limitations.

6.1 Mitigation techniques

We suggest three potential mitigation techniques to protect users of online multimedia streaming services and VoIP applications from the proposed (or similar) attacks.

6.1.1 Avoiding AC microphones. As mentioned in Section 3, the LISTEN attack works on users who are using input devices that capture ENF signals. In particular, the raw audio stream data being transmitted need to be produced through an AC microphone. Therefore, to preserve privacy, use of other types of microphones (e.g., a DC-powered microphone) can be recommended to reduce the risk of being exposed to ENF-based side-channel attacks.

6.1.2 Insertion of fake signals. In order to downgrade the performance of the LISTEN attack, one could add noise to target raw audio stream data. Noisy 50 or 60Hz signals can be inserted before transmitting stream data to attacker's device or uploading a recorded stream file to a streaming server. Added noise will make it more difficult to extract original ENF signals. This countermeasure needs to be designed carefully though as insertion of pure random noise might not be effective – pure random signals can be easily removed through a noise cancellation filter such as the *median* filter. A more ideal way of generating noise signals is to randomly choose *fake* signals from a collected set of real-world ENF signals. Such fake signals will be much harder to identify and filter.

6.1.3 Removal of ENF signal patterns. Another possible approach is to remove ENF signals from the raw audio stream data. Chuang et al. [8] presented several signal processing techniques to remove and modify ENF signals while guaranteeing high quality streaming of raw audio data. For example, we can use the band-stop filter to remove only ENF signals at the specific range of frequency band since the band-stop filter passes most frequencies of audio data unaltered but removes restricted small frequency region. The band-stop filtering techniques have been studied comprehensively in the field of signal processing [41].

6.2 LISTEN attack limitations

Although the LISTEN attack can be effectively used to identify recording places of content creators or physical location of VoIP users, the attack provides coarse-grained location information within a given a power grid (see Section 5.3). This degree of inferred detail might not be sufficient for applications that require more fine-grained location information. Also, the performance of the LISTEN attack could be degraded depending upon the segment length of given raw audio stream data. As mentioned in Section 3, the LISTEN attack requires the raw audio stream data to be produced by a device that is capable of capturing ENF signals; e.g., AC microphones.

6.3 Effectiveness of ENF map with a small number of ENF samples

Although our ENF map is validated by estimating cross-correlation between interpolated ENF signals and underlying ground-truth ENF signals (as shown in Section 4), the ENF map can become unstable when a small number of ENF samples are only used for constructing the map. In such environments, location cannot be accurately pinpointed. It is obvious that a more accurate map can be constructed and location can be identified with a higher accuracy as ENF sample size increases. That is, the attack (inferred area) accuracy would improve with the increase in collected dataset size. A possible way to increase the number of ENF samples is to combine the ENF samples collected from multimedia streaming services with physical ENF signals collected from GridEye/FNET system [22].

7 RELATED WORK

7.1 Inferring user location

In mobile phone, Narain et al. [39] presented the approach to infer the location and route of moving target with only gyroscope, accelerometer, and magnetometer information. They apply this information to already collected road information with their algorithm. In Android mobile phone, the permission should be approved by mobile phone user to install the application. While the GPS sensor permission is in critical level, other sensors which we previously mentioned does not need any additional permissions. Compared to our research, there are no applications that collect the information from these many sensors among the commonly used applications, that additional installations are needed because they are not provided by the server even if they are collected from common application. Similarly, Michalevsky [38] used power consumption only for inferring location, even if it is not considered as critical as sensor's information from [39]. The fundamental idea of this study is that power consumption depends on the location of the mobile device. They also gathered routes and power consumption information on road and applied those data to machine learning algorithm. Both studies were conducted on mobile device which is moving along the road, while our general attack target is motionless indoors.

7.2 Revealing anonymity in VoIP

In this section, we summarize other researches about tracking users' location through VoIP services and compare them with our results.

There are recent researches trying to extract useful information from audio data. By Wright et al. [54], most VoIP services use variable bitrates (VBR) audio codecs for encoding. In VBR codecs, vowels and consonants are usually encoded in packets of different lengths. Using this information, Wright et al. proposed way to find out which phrases were spoken from VoIP packet sizes. Using this result, Coskun and Memon proposed robust hashing scheme for VoIP packet to track VoIP calls. [12] They suggest hashing scheme which is able to pair original packet streams to distorted streams after delay, jitter, and packet drops. However, though its robustness can be applied in impairments-existent conditions, there are some limitations to apply their scheme to actually tacking VoIP callers. Since it is only able to check whether two packet streams store the same (or similar) data, it is needed to monitor all packets to identify pairs among all possible nodes and this complexity isn't reduced if we could control one endpoint.

8 CONCLUSION

Unlike existing location inference techniques [38, 39, 53] that require installation of a malicious application on a victim's device and an expensive ENF receiver, the proposed LISTEN attack can be performed with access to just the target video or audio file.

To demonstrate the effectiveness of the LISTEN attack, we experimented with the multimedia data collected from three online streaming services, Earthcam, Skyline, and Explore, as well as two VoIP applications, Skype, and Torfone. Our results show that the LISTEN attack can be highly effective in inferring the physical location of which a video or audio file was recorded. We achieved an accuracy of 76% which is a reasonable level when the multimedia source was 5 minutes or longer.

Our LISTEN attack is currently limited to the multimedia files recorded with mains-powered microphones. For generalization, we also plan to design ENF signals-based attacks for environments where mains-powered microphones are not used.

Although we positioned the findings as a way to perform an inference attack, our techniques could also be used to identify locations of criminals such as kidnappers, terrorists, or phishers who use multimedia to threaten and abuse people. Another future work is to extend our findings to develop such countermeasure technologies.

REFERENCES

[1] M. Abe and J. O. Smith III. 2004. Design criteria for simple sinusoidal parameter estimation based on quadratic i5nterpolation of FFT magnitude peaks. In *Audio Engineering Society Convention 117*. Audio Engineering Society.

[2] Skype and/or Microsoft. 2017. Skype | Free calls to friends and family. (2017). https://www.skype.com/

[3] C. Bettini, X. S. Wang, and S. Jajodia. 2005. Protecting Privacy Against Location-based Personal Identification. In *Proceedings of the Second VDLB International Conference on Secure Data Management*. Springer, Berlin, Heidelberg, 185–199.

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

[5] J. Chai, F. Liu, Z. Yuan, R. W. Conners, and Y. Liu. 2013. Source of ENF in Battery-Powered Digital Recordings. In *Audio Engineering Society Convention 135*. AES, Convention, 1–7. http://www.aes.org/e-lib/browse.cfm?elib=17055

[6] F.-C. Chang and H.-C. Huang. 2010. Electrical network frequency as a tool for audio concealment process. In *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2010 Sixth International Conference on*. IEEE, 175–178.

[7] L. Chen, P. Markham, C. Chen, and Y. Liu. 2011. Analysis of societal event impacts on the power system frequency using FNET measurements. In *Power and Energy Society General Meeting*. IEEE, Detroit, MI, USA, 1–8.

[8] W. H. Chuang, R. Garg, and M. Wu. 2013. Anti-Forensics and Countermeasures of Electrical Network Frequency Analysis. *IEEE Transactions on Information Forensics and Security* 8, 12 (2013), 2073–2088.

[9] A. J. Cooper. 2008. The electric network frequency (ENF) as an aid to authenticating forensic digital audio recordings–an automated approach. In *Audio Engineering Society Conference: 33rd International Conference: Audio Forensics-Theory and Practice*. Audio Engineering Society.

[10] A. J. Cooper. 2009. An automated approach to the Electric Network Frequency (ENF) criterion-Theory and practice. *International Journal of Speech, Language and the Law* 16, 2 (2009), 193–218.

[11] IBM Corporation. 2017. Live Streaming, Online Video & Hosting Services | Ustream. (2017). http://www.ustream.tv/

[12] B. Coskun and N. Memon. 2010. Tracking encrypted VoIP calls via robust hashing of network flows. In *International Conference on Acoustics, Speech and Signal Processing*. IEEE, Dallas, TX, USA, 1818–1821. https://doi.org/10.1109/ICASSP.2010.5495398

[13] Inc. EarthCam. 2017. EarthCam - Webcam Network. (2017). https://www.earthcam.com/

[14] Facebook. 2017. Facebook Live | Live Video Streaming. (2017). https://live.fb.com/

[15] Facebook. 2017. Messenger. (2017). https://www.facebook.com/messenger/

[16] The Annenberg Foundation. 2017. The largest live nature cam network on the planet World! (2017). http://explore.org/

[17] S. Gambs, M.-O. Killijian, and M. Nú nez Del Prado Cortez. 2014. De-anonymization Attack on Geolocated Data. *J. Comput. System Sci.* 80, 8 (2014), 1597–1614.

[18] R. Garg, A. Hajj-Ahmad, and M. Wu. 2013. Geo-location estimation from Electrical Network Frequency signals. In *ICASSP*. IEEE, Vancouver, BC, Canada, 2862–2866.

[19] D. Goldschlag, M. Reed, and P. Syverson. 1999. Onion routing. *Commun. ACM* 42, 2 (1999), 39–41.

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

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

[22] J. Guo, Y. Ye, Y. Zhang, Y. Lei, and Y. Liu. 2014. Events associated power system oscillations observation based on distribution-level phasor measurements. In *T & D Conference and Exposition*. IEEE, Chicago, IL, USA, 1–5. https://doi.org/10.1109/TDC.2014.6863463

[23] H. Kim and Y. Jeon and J. W. Yoon. 2017. Construction of a National Scale ENF Map using Online Multimedia Data. In *ACM International Conference on Information and Knowledge Management, (CIKM) 2017*. ACM, Singapore.

[24] A. Hajj-Ahmad, R. Garg, and M. Wu. 2012. Instantaneous frequency estimation and localization for ENF signals. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, Hollywood, CA, USA, 1–10.

[25] A. Hajj-Ahmad, R. Garg, and M. Wu. 2013. ENF based location classification of sensor recordings. In *2013 International Workshop on Information Forensics and Security, WIFS 2013, Guangzhou, China, November 18-21, 2013*. IEEE, Guangzhou, China, 138–143.

[26] A. Hajj-Ahmad, R. Garg, and M. Wu. 2015. ENF-based region-of-recording identification for media signals. *IEEE Transactions on Information Forensics and Security* 10, 6 (2015), 1125–1136.

[27] M. Huijbregtse and Z. Geradts. 2009. Using the ENF criterion for determining the time of recording of short digital audio recordings. In *International Workshop on Computational Forensics*. Springer, Berlin, Heidelberg, 116–124.

[28] WhatsApp Inc. 2017. WhatsApp. (2017). https://www.whatsapp.com/

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

[30] S. Karapantazis and F.-N. Pavlidou. 2009. VoIP: A comprehensive survey on a promising technology. *Computer Networks* 53, 12 (2009), 2050–2090.

[31] R. Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, Vol. 14. 1137–1145.

[32] Krebs on Security. [n. d.]. Skype Now Hides Your Internet Address. https://krebsonsecurity.com/2016/01/skype-now-hides-your-internet-address/. ([n. d.]). Online; January 2016.

[33] Y. Liu and Y. Ye. 2012. Monitoring power system disturbances based on distribution-level phasor measurements, In PES Innovative Smart Grid Technologies (ISGT). *Innovative Smart Grid Technologies* 00, 1–8.

[34] Y. Liu, Z. Yuan, P. N. Markham, R. W. Conners, and Y. Liu. 2011. Wide-area frequency as a criterion for digital audio recording authentication. In *Power and Energy Society General Meeting*. IEEE, Detroit, MI, USA, 1–7.

[35] G. Y. Lu and D. W. Wong. 2008. An adaptive inverse-distance weighting spatial interpolation technique. *Computers & Geosciences* 34, 9 (2008), 1044–1055.

[36] S. Mann, L. Cuccovillo, P. Aichroth, and C. Dittmar. 2013. Combining ENF Phase Discontinuity Checking and Temporal Pattern Matching for Audio Tampering Detection. In *GI-Jahrestagung (LNI)*, Matthias Horbach (Ed.), Vol. 220. GI, 2917=–2927.

[37] D. McCoy, K. Bauer, D. Grunwald, T. Kohno, and D. Sicker. 2008. Shining light in dark places: Understanding the Tor network. In *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, Berlin, Heidelberg, 63–76.

[38] Y. Michalevsky, A. Schulman, G. A. Veerapandian, D. Boneh, and G. Nakibly. 2015. PowerSpy: Location Tracking Using Mobile Device Power Analysis. In *USENIX Security*. USENIX Association, Washington, D.C., 785–800.

[39] S. Narain, T. D. Vo-Huu, K. Block, and G. Noubir. 2016. Inferring User Routes and Locations using Zero-Permission Mobile Sensors. In *Symposium on Security and Privacy (S&P)*. IEEE Computer Society, San Jose, CA, USA, 397–413.

[40] D. W. Oard, M. Wu, K. Kraus, A. Hajj-Ahmad, H. Su, and R. Garg. 2014. It's About Time: Projecting Temporal Metadata for Historically Significant Recordings. *iConference 2014 Proceedings* (2014).

[41] A. V. Oppenheim. 1999. *Discrete-time signal processing*. Pearson Education India.

[42] B. Porat. 1994. Digital Processing of Random Signals: Theory and Methods. (1994).

[43] H. Su, R. Garg, A. Hajj-Ahmad, and M. Wu. 2013. ENF analysis on recaptured audio recordings. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE Signal Processing Socieity, Vancouver, BC, Canada, 3018–3022.

[44] H. Su, A. Hajj-Ahmad, M. Wu, and D. W. Oard. 2014. Exploring the use of ENF for multimedia synchronization. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE Signal Processing Society, Florence, Italy, 4613–4617.

[45] J. Tabrikian, S. Dubnov, and Y. Dickalov. 2004. Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model. *IEEE Transactions on Speech and Audio Processing* 12, 1 (2004), 76–87.

[46] Inc. Twitch Interactive. 2017. Twitch. (2017). https://www.twitch.tv/

[47] Twitch Tips. [n. d.]. Protecting Yourself Online. https://twitchtips.com/protecting-yourself/. ([n. d.]). Online; 21 January 2016.

[48] Twitter. 2017. Watch LIVE. (2017). https://www.periscope.tv/

[49] T. Uhl. 2004. Quality of service in VoIP communication. *AEU-International Journal of Electronics and Communications* 58, 3 (2004), 178–182.

[50] S.r.l. VisioRay. 2017. SkylineWebcams | Live Cams around the World! (2017). https://www.skylinewebcams.com/

[51] K. Vos. 2011. SILK Speech Codec. (2011). https://tools.ietf.org/html/draft-vos-silk-02

[52] L. Wang, J. Burgett, J. Zuo, C. C. Xu, B. J. Billian, R. W. Conners, and Y. Liu. 2007. Frequency disturbance recorder design and developments. In *Power Engineering Society General Meeting*. IEEE, Tampa, FL, USA, 1–7.

[53] X. Wang, S. Chen, and S. Jajodia. 2005. Tracking anonymous peer-to-peer VoIP calls on the internet. In *Proceedings of the 12th ACM conference on Computer and communications security*. ACM, New York, NY, USA, 81–91.

[54] C. V. Wright, L. Ballard, S. E. Coull, F. Monrose, and G. M. Masson. 2008. Spot Me if You Can: Uncovering Spoken Phrases in Encrypted VoIP Conversations. In *Symposium on Security and Privacy (sp 2008)*. 35–49. https://doi.org/10.1109/SP.2008.21

[55] LLC YouTube. 2017. YouTube. (2017). https://www.youtube.com/

[56] Z. Zhong, C. Xu, B. J. Billian, L. Zhang, S. S. Tsai, R. W. Conners, V. A. Centeno, A. G. Phadke, and Y. Liu. 2005. Power system frequency monitoring network (FNET) implementation. *IEEE Transactions on Power Systems* 20, 4 (2005), 1914–1921.