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

Anonymous author
Anonymous author
author@anonymous.com

Anonymous author
Anonymous author
author@anonymous.com

Anonymous author
Anonymous author
author@anonymous.com

***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 file to perform the attack. The evaluation results show that our attack can infer the 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 even when video and audio data are distorted through the use of anonymous networks like Tor.**

I. 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 [14], Skype [2], and WhatsApp [27], have emerged over the years, and become popular. Also, many online streaming services, such as YouTube [54], Facebook Live [13], Twitter's Periscope [47], and Twitch [45], 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], [16] 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 [30].

NDSS '18, 18-21 February 2018, San Diego, CA, USA

http://dx.doi.org/10.14722/ndss.2018.23xxx

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 [46].

However, researchers have presented various ways of compromising location privacy. PowerSpy [36], 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 [37], 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 [36], [37], [52] 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 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 [21], [38], [20], [34], [26], [42], [5], [6].

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 [24], [25], and also obtain the precise location information within the grid [17], [23]. However, the existing ENF processing techniques [17], [23] 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 [31], [51], [7], [29] 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 collected a very large amount of real-world ENF signals and constructed the first interpolated global ENF map to infer user location. Our novel technique enables efficient construction and continuous update of a global ENF map, which is an integral piece in facilitating wide coverage, and high accuracy and practicality in location estimation. The interpolated ENF map allows identification of previously undiscovered locations while previous studies use classification techniques to label user locations with reference to a fixed set of known locations.
- We evaluated the performance of the proposed attack in real-world environments without losing generosity. 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 when the portion of decision boundary area to total grid is 3 out of n, where n denotes the number of levels of a contour in a power grid.

The rest of the paper is organized as follows. In Section II, we explain how ENF signals can be obtained from online multimedia stream data and used for location tracking. Section III describes the generic attack model, and Section IV dives deep into the proposed LISTEN attack. Section V presents the attack evaluation results, and Section VI discusses those results. Related work is covered in Section VII and our conclusions are in Section VIII.

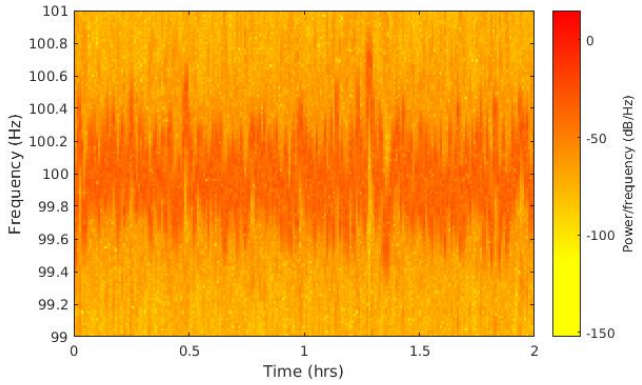
II. Background

This section explains the processes involved in extracting ENF signals from online multimedia streaming services, and in constructing a ENF map for locations of interest.

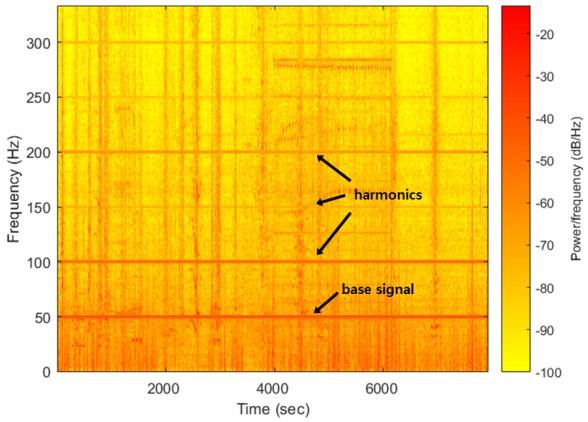
A. Electrical network frequency (ENF)

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], [51], [31], and digital forensics [21], [34], [41], [6], [26]. 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].



(a) ENF signal at base frequency



(b) ENF signal at base frequency and harmonic frequencies

Fig. 1. Spectrogram of an audio file with ENF signals. There are a few couples of horizontal lines in spectrogram, which is called harmonic signals of ENF.

ENF signals can also be obtained from side-channels such as audio and video files [31], [51], [32], [21], [43] such as Figure 1-(a). ENF patterns reconstructed from a side-channel often have much lower signal-to-noise ratio (SNR)

than those directly acquired from an FDR device. Therefore, signal processing techniques needed to be applied to reduce or remove unwanted noise when ENF signals are captured from side-channels like audio or video streams. Figure 1-(b) 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 horizontal line around frequency of 50Hz. Additionally, we can find several horizontal lines in the spectrogram shown in Figure 1-(b), 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[1] on the spectrogram [4], [25].

B. Extracting ENF signals from multimedia streaming data

To obtain reliable ENF signals, several signal processing algorithms need to be applied to audio and video files. The ENF signal extraction process mainly consists of the following four steps.

1) Decimating and framing: Our extraction method involves crawling and scraping audio and video files from online multimedia services in real-time. An efficient data storage mechanism is needed as the accumulated data sizes can quickly become large, and scrapped multimedia streaming data might not remain on streaming service servers. To save storage space, and improve the efficiency of our crawling program, we decimate collected audio signals to 1kHz before saving them.

Using decimated signals, we create frames of data sequences, where each frame overlaps with the half of the previous frame. Each frame contains 8192 samples, which comes to about 8 seconds of decimated audio data in each frame. 4096 samples overlap with each other. This concept comes from the STFT technique.

Note that it is important to choose the optimal number of samples to be included in one frame for extracting ENF signals. If the frame size is too small, the frequency resolution of each frame will also be too low to extract a meaningful frequency value. If the frame size is too large, we could end up with insufficient information being extracted from a given time as signals become blurry. In this case, ENF signal variations that exist in each frame will not be detected. This is called the "uncertainty principle," which is described in a research conducted by Cooper et al [10].

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 [23]. 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. If the sampling rate is denoted as f_s, and the frame size is denoted as n, then the unit value of each frequency resolution will be $\mathbf{f_s}/\mathbf{n}$. In our case, when $f_s = 1000$, and $n = 8192$, then the estimated ENF signal unit value would roughly be $122mHz$. Considering that the standard deviation of ENF signals is approximately $20mHz$ [10], this estimation is very imprecise. Hence, we apply the Quadratic Interpolation technique when the FFT process is complete [10], [23], [39]. 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 get better estimation by approximating the signal to quadratic formula with using values which are nearby the maximum frequency value.

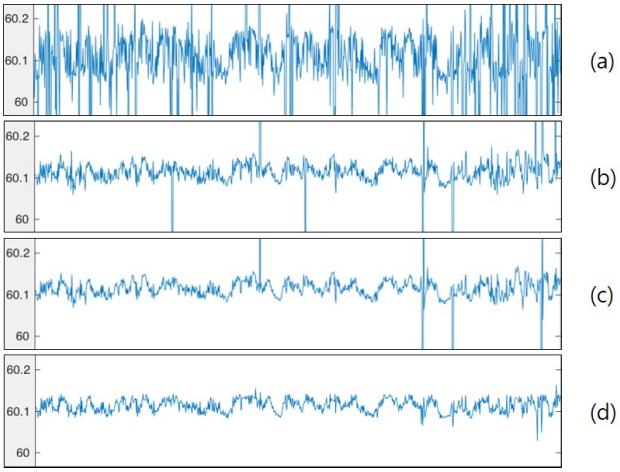


Fig. 2. Constructing ENF signals with multi-tone estimation. (a) using the base signal (60Hz); (b) using the base and from 2nd through 4th signals; (c) using the base and from 2nd through 6th signals; (d) using the base and from 2nd through 8th signals.

3) Multi-tone estimation: Initially, we would obtain ENF signals such as shown in (a) of Figure 2. As can be seen, initial ENF signals are highly corrupted due to unwanted noises. To remove or reduce noise level, we apply multi-tone harmonic signals to improve the quality of ENF signals [4], [10], [39]. 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. This process has been explained by Bykhovsky et al [4], showing that the estimation accuracy of ENF signals can be improved by about $10 - 15\mu$Hz [4]. Given a multimedia sound signal $x(t)$ in a time domain, we can transform the signal to $F(\omega) = \sum_{n=0}^{N-1} x(t)e^{-j\omega t}$ in the frequency domain. While ENF patterns exist around the $50/60Hz$ band of a fundamental frequency, similar ENF patterns are also present in harmonic frequency bands that are multiples of $50/60Hz$. Therefore, to enhance the ENF patterns against the unwanted uncorrelated noise in the frequency domain, multi-

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

tone spectra is obtained by summing all spectrogram at both fundamental and harmonic frequencies. Such improvements can be seen in Figure 2. The more harmonic signals we use, the better the accuracy of ENF signal estimation.

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).

III. 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. The threat model of LISTEN attacks is shown in Figure 3.

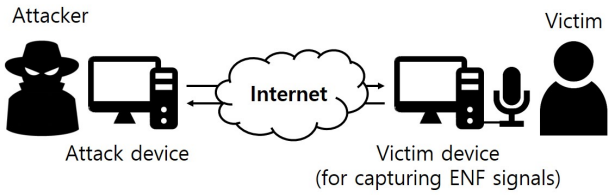


Fig. 3. Threat model of LISTEN attacks.

At first glance, our assumptions do not seem reasonable. However, 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. 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 [35]) because the attacker does not require additional information from the victim, besides the transmitted recorded audio and/or video signals.

IV. LISTEN Attack

ENF signal patterns that appear in the distribution network of a power plant are either 50Hz or 60Hz, depending on the geographic location. Moreover, those ENF signal patterns have temporal fluctuations based on the specific conditions of power distribution. Because of those properties, ENF signals could be used as a spatio-temporal signature for determining location and time.

The primary goal of the LISTEN attack is to identify the location of a victim device with access to just ENF signals (side channel information). The attack consists of three sequential processes as shown below.

1) **Construction of the ENF map for locations of interest.** First process is about constructing a real-time ENF map (this map is also referred to as "the ENF map" in the paper) by interpolating the unknown-ENF area that will be used as ENF sequences to compare against.
2) **Extraction of reliable ENF signals from a target device.** The second process involves extracting and processing ENF signals collected from a victim's device.
3) **Location estimation.** In the third process, the LISTEN attack attempts to identify the victim's location by efficiently comparing the signals from the ENF map against the victim's ENF signals.

The following sections describe those three processes and used algorithms in detail.

A. Construction of the ENF map for locations of interest

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. Ideally, we would need to cover all possible ENF ranges across the entire world. However, building such a large ENF map would require a massive effort and budget. Specialized physical devices such as (FDR) [51] 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.

In contrast, our approach does not require purchase and installation of expensive physical devices. Our automated

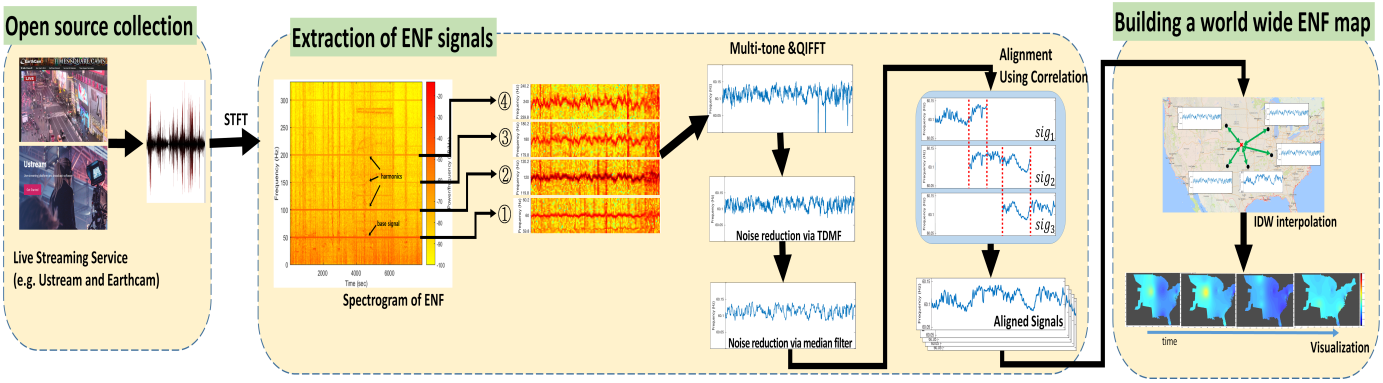


Fig. 4. A full procedure to construct the ENF map: the construction of the ENF map consists of three steps. We first collect audio and video streaming data from online. Then we extract ENF patterns from the audio signals and the patterns are refined using advanced signal processing filters such as multi-tone & QIFFT, TDMF and signal alignments. Finally, the ENF signals are used to interpolate the ENF patterns at unknown area.

programs crawl and scrap worldwide ENF signals from online multimedia services such as "EarthCam" and "Ustream," significantly reducing costs, time, and amount of efforts needed to create a map. However, more complicated signal processing techniques need to be applied (to improve signal quality) because ENF signals scraped from online sources are less clear.

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 [12], Explore [15], and Skyline [49] 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: (1) checking scraped audio streams contain ENF signals, (2) extracting clear signals through noise reduction, and (3) aligning incomplete and partial signals on a given time domain using signal alignment techniques. Those techniques are described in Section IV-D. Accurate ENF signals can be obtained after this step.

However, another problem is that we can only collect ENF signals from a fixed set of streaming source locations. We would miss signals from locations for which the selected services do not stream audio and video. One possible solution to this problem is signal interpolation. ENF signals from uncovered areas are reconstructed by interpolation with collected neighboring ENF signals. We refer to collected ENF signals as *anchor nodes* (this word is also referred to as "anchor node" in the paper), 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. More details are described in Section IV-C2.

The effectiveness of an interpolated ENF map can be explained theoretically from the fact that the ENF disturbance propagation speed is finite [23], [17], which means that all the inner grid ENF values will not be measured as the same value. Based on those characteristics, we can infer the location of a given ENF signal by going through the previously collected database of ENF signal signatures and known locations, and finding a signal that has the most similar pattern.

Before interpolation, we virtually divide the entire ENF map by 0.1 degree; as a result, each cell size becomes 0.1 by 0.1 degree. We then apply the Inverse Distance Weighted (IDW) interpolation technique to each cell using the values of sampled points [33]. The simple weighted interpolation value $\hat{y}$ can be expressed by $\hat{y} = \Sigma\lambda_i y_i$, where y_i represents evaluated values from anchor nodes, and λ_i represents weight of each point. The inverse distance weight implies that λ_i values are inversely proportional to distance values. Experimentally, we simulated the optimal order of 2 with cross-validation. Based on that λ can be expressed as $\lambda = d_i^{-p}/\Sigma d_i^{-p}$, where d_i is the Euclidean distance between two signals from victim and anchor nodes. The final, interpolated ENF map is shown in Figure 5.

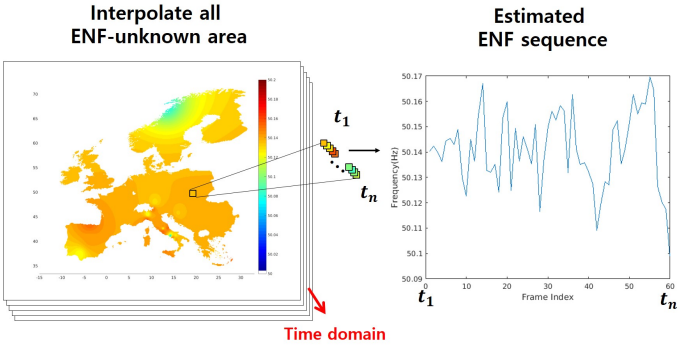


Fig. 5. Interpolated ENF signal sequence of the Europe continent. We can estimate any point in an ENF sequence using a series of interpolated matrix of ENF map.

Before building our LISTEN scheme, one of the most important prior steps is to validate the effectiveness of the interpolated ENF map because LISTEN cannot work without an accurate ENF map. Since ENF map is built through IDW interpolation, we performed to verify the availability of the IDW interpolation by optimizing the unknown p to minimize the error of IDW using 4 fold cross-validation and showing the similarity between the constructed ENF map and the ground-truth data collected from the FNET/GridEye server [22]. The data was randomly partitioned into into 4 sub-samples and labeled into two groups: a training set and a testing set. With

this partitioned dataset, 4 fold cross-validation was adopted by using ML (Maximum Likelihood) estimate in order to estimate the optimal p which is a critical model parameter of IDW interpolation. To prove the availability of the IDW interpolation, we also used a similarity metric to compare the constructed ENF map with the ground-truth data. The similarity was obtained by calculating the normalized cross correlation (NCC) between interpolated ENF signals and underlying ground-truth ENF signals collected from the FNET/GridEye server. In this comparison, the expectation and standard deviation of the cross-validation are 0.7 and 0.1, respectively, with 40 minutes long stream data.

B. Extracting ENF signals from a victim

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 IV-A but requires more sophisticated algorithms due to various communication systems and environments that need to be considered. For instance, the victim could be using a VoIP service that streams unreliable ENF signals as shown in Figure 6. Such signals could be distorted and carry a significant level of noise.

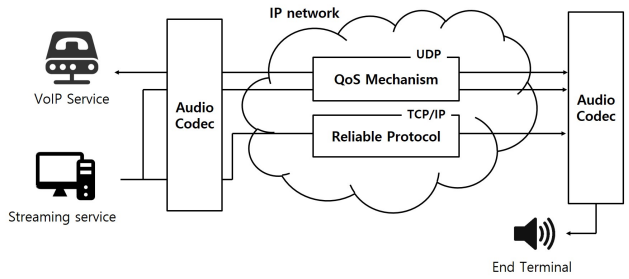


Fig. 6. Architecture for audio streaming over IP network for VoIP applications and online streaming services. Reliable data for constructing a ENF map is collected from an online streaming service; sound recorded from a victim's device is received on an audio channel where packet loss may exist.

Signal distortion could occur when an audio signal goes through an audio codec or QoS mechanism as shown in Figure 6. 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 about those metrics and the techniques how attackers mitigate those problem.

1) Frequency response: We first check frequency response, which includes a band pass filter. Audio recorded from the victim's device can be filtered or amplified when it passes through an audio channel. ENF signals cannot be reconstructed 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 [28]. For example, in the case of Skype, the VoIP application uses its own codec called SILK [50]. The compression process of SILK uses a high pass filter for which the cut-off frequency is 70Hz [50]. Since the base frequency of ENF signal is 50 or 60Hz, SILK will filter it. Only the remaining harmonic signals that pass the band pass filter will be made available.

To resolve this, we use the multi-tone estimation [44] shown in Section IV-A. Multi-tone estimation is enhancing the signals of our interest by combining multiple signals at fundamental frequency and harmonic frequencies as shown in Figure 2. We use harmonic signals with frequency of either 100Hz for the 1st frequency 50Hz or 120Hz for the 1st frequency 60Hz, or above.

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 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 $8,192$ 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 [28]. Since this jitter reduction incurs its own time delay, aligning time against time delay is a only concern.

3) Packet loss: Packet loss is another important factor since ENF signals cannot be reconstructed 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, 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. Another common case is the use of anonymous networks such as Torfone [18] where users try to conceal themselves. As anonymous networks do not have transparent latency and bandwidth, packet losses could occur when UDP-based services (or any other service that uses an unreliable protocol) are used over those networks such as Torfone.

According to the survey conducted in [48], 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.

With those mitigation previously mentioned, we restore the ENF signals which are distorted by passing through the audio channel such as Figure 7. The patterns seems almost similar, but reconstructed signal from Skype seems have more error

than that from Torfone. From this figure, we can infer that the lack of base ENF signal is more critical than the data loss.

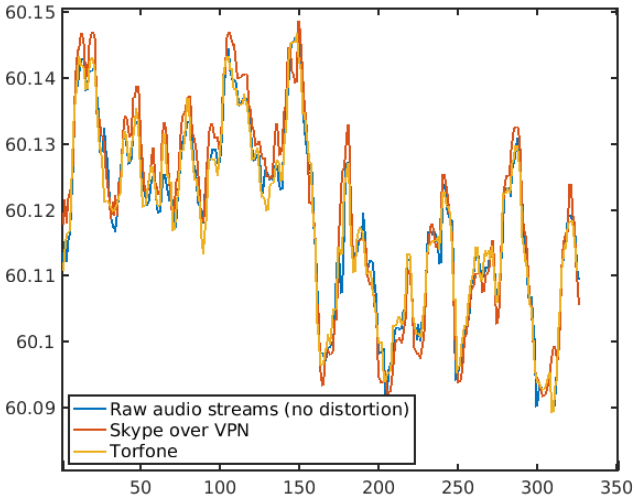


Fig. 7. Reconstructed ENF signals passed through the audio channel. (The figures are best viewed in colors)

C. Location estimating

The next process is to estimate the inter and intra-grid locations using the ENF signals processed through the techniques described above.

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 (DW **k**-NN) algorithm. After labeling the collected set of anchor nodes with location information, we determine the **k**-nearest neighbours with inversely proportioned weights. Since we are using **k**-nearest neighbours, other nodes will have an weight of 0. The expected labels can be denoted as follows: $arg\max(\Sigma(w_i/\Sigma w_i \times f(x_i)))$, where $f(x_i)$ is the label of node **i**, $w_i = d_i^{-2}$ is the assigned weight of node **i**, and d_i is the euclidean distance of signals between a classifying point and node **i**. Here, **k** is selected based on the number of ENF signals collected to be used as anchor nodes.

However, a distance weighted algorithm can be used to deal with the bias of sampling numbers of each grid. For instance, let assume that there are Eastern and Western power grid, and that the signal from target victim actually captured from the East. If there are not enough number of samples from the East, common **k**-NN algorithm might infer the location to the West even the Euclidean distances between a victim and the West anchor nodes are close. Therefore, we can resolve this problem by assigning the weights to 0 when the distances are too far.

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 Section IV-A).

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 correlation coefficient [17], [23], 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 8). The color map represents the distances

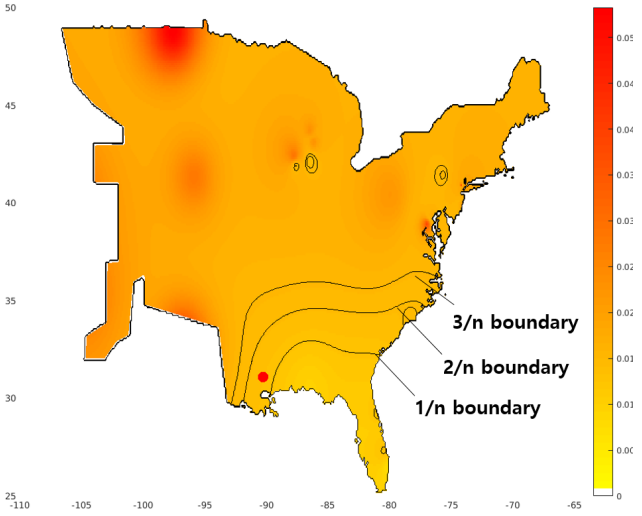


Fig. 8. Euclidean distance between target ENF sequence and interpolated sequences in the Eastern power grid of the United States. Red dot indicates where the signal actually collected. Red area means it is far from target signal and yellow area means it is close. The ENF presence decision boundaries divide total area equally with number of n.

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.

To evaluate the accuracy of the location inference attack, the target region was divided into n parts of equal area where n is the number of ENF signal samples. The term "decision boundary" is used to separate and distinguish each area. The attack accuracy is defined as the ratio between the number of correctly guessed (inferred) areas and the total number of areas.

For example, in Figure 8, the red point 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, resulting prediction could be wrong; if we set the boundary by choosing the second highest boundary probability, resulting prediction is more likely to be right. Since the degree of precision depends on n, we will discuss how n is chosen in Section V.

As the range of the decision boundary increases, the attack accuracy also increases. To set the right number of n, we determine the region decision boundaries by the number of anchor nodes in this paper so n is the number of anchor nodes in the same grid.

Since the electrical network characteristics and structures are not identical for all grids around the world, the propagation and convergence speeds could be different. Further, location estimation in intra-grid can be performed with many approximation techniques, losing high volumes of integral information. Despite those concerns, the effectiveness of intra-grid estimation has already been demonstrated with experimental results [17], [32], [23], [24], [25].

D. LISTEN attack optimization

Additionally, we developed more advanced signal processing methods for the LISTEN attack in order to extract more accurate ENF signals.

First signal processing technique is to check whether the scraped audio streams contain ENF signals of interest. It is evident that only some audio streams or files contain ENF signals in a known frequency domain. If any sound is not recorded through an AC microphone, we cannot capture ENF signals from it. Therefore, we need to check whether a collected audio file or stream includes ENF signals, and delete the file if it does not contain ENF signals.

There are two ways of getting rid of unreliable data. The first approach is to compare the ENF signals from the victim to all other reference signals from anchor nodes and check whether it is abnormally fluctuating. This is based on the assumption that ENF signals in an identical power grid will have minimal variations. This is effective since it uses the reliability of other signals. The other way to remove the unreliable data is to use self standard deviation. Without comparing against any other extracted signal, we can use the fact that a clean ENF signal has a standard deviation of ± 0.03Hz [10]. By filtering the ENF signals which have value higher than 0.03, we can have moderate quality of ENF signals.

After checking the existence of the ENF signals, we reduce noise and enhance the desired signals by using multi-tone harmonics analysis and QIFFT to improve the quality of ENF signals.

Finally, we need an additional step to build stable and longer ENF signals with signal alignments. Beyond the noise reduction, it is also important to obtain longer ENF signals since they can provide more accurate and stable analysis [7], [22], [31], [55], [17], [21]. However, most ENF signals directly obtained from multimedia data are too short to use because it is very difficult to directly extract a long ENF signal from a multimedia service. For instance, if the broadcaster of Ustream service offers content for a sufficiently long time, we can download and analyze the appropriate multimedia data for the ENF signal; however, if not, then we either cannot obtain the full signal or can obtain only incomplete signals. Therefore, to build stable and longer ENF signals, a signal alignment technique is introduced to align the multiple ENF signals from different multimedia broadcasts. After we collect many multimedia data streams in adjacent places and subsequent times with an overlap of the time range, we can construct a semi-complete ENF signal. To align the multiple partial ENF signals, the well-known normalized cross correlation (NCC) metrics is used since it provides the similarity of the overlapping signals.

V. Evaluation

This section presents the LISTEN attack performance evaluation results. We calculated the accuracy of inter- and intra-grid estimation 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.

We describe those experimental setups in Section V-A, and show inter-grid estimation performance and intra-grid estimation performance in Sections V-C and V-D, respectively.

A. Experiment setups

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.

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 I. 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 $1,000$Hz sound source streams, taking up about 10MB of disk space.

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 II.

Skype, which is one of the most widely used VoIP services, uses peer-to-peer protocols to establish an Internet telephony

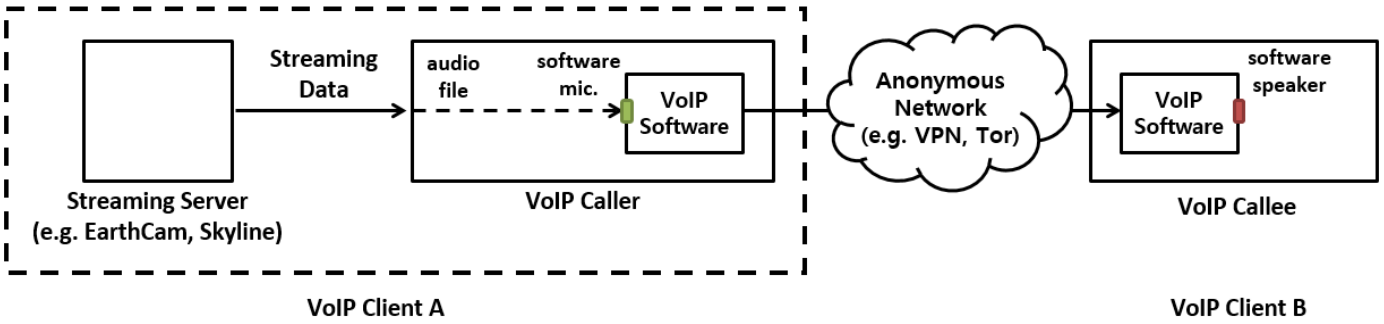


Fig. 9. To emulate VoIP client A (Caller) from a remote host, we first obtained the audio data from streaming servers and inputted the audio file directly to VoIP software. To exclude unintended effects from physical devices, all the microphone and speaker operations were processed with virtualization 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.

TABLE I. 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 [12]	landscape	85.29	36
Skyline [49]	landscape	95.16	39
Explore [15]	nature	70.59	24

TABLE II. 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

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 [2] as IDs, and connects users through Tor networks. Due to the privacy-preserving characteristics of Tor, Torfone provides stronger anonymity than a typical VPN service but could lose more packets in UDP connections (which most VoIP applications use). 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.

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 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.

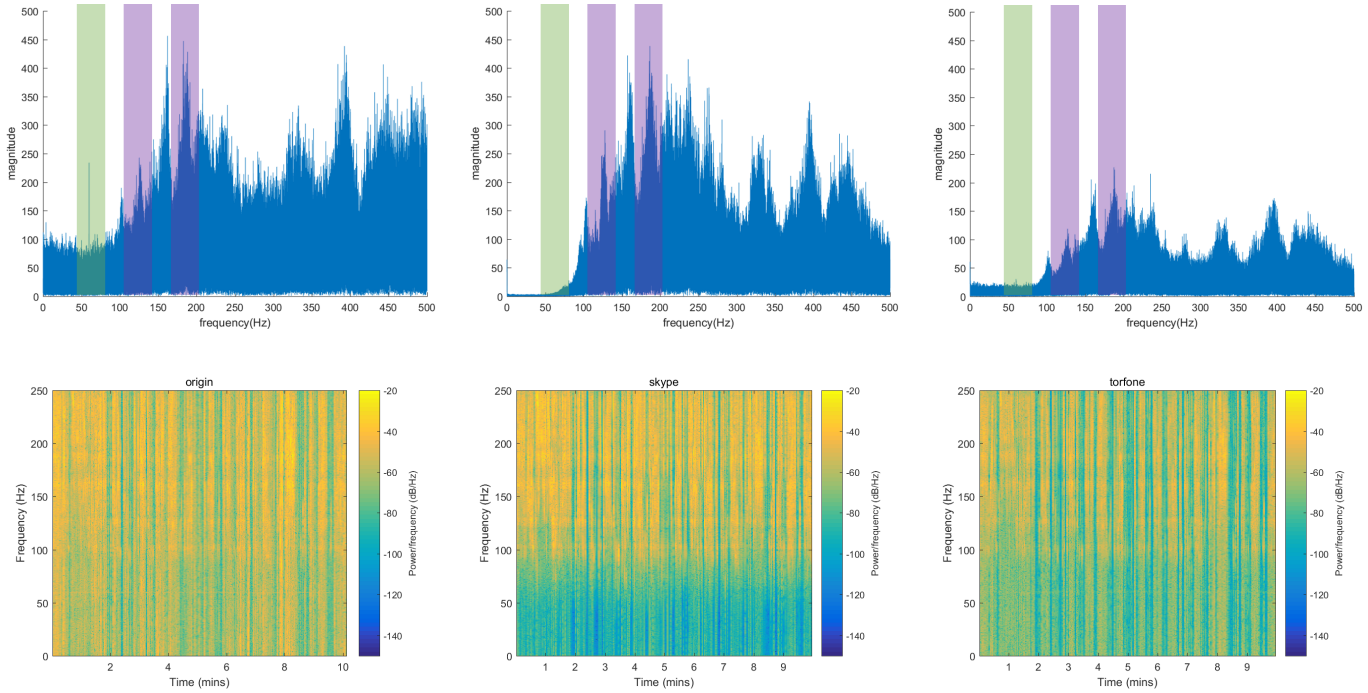
To overcome this problem, we used an audio virtualization technique, which redirects sounds (from audio files) to a microphone of a computer. Considering that microphones and speakers work in a similar way, it is possible to redirect an output of a speaker (sound information) to an input of a microphone. 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 9.

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 be 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. As shown in Figure 9, 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.

B. 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.

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



(a) Raw audio streams (no distortion) (b) Skype+VPN (c) Torfone

Fig. 10. FFT results and spectrograms of the captured audio streaming data. The upper figures plot FFT results of the raw audio streams (a), Skype (b), and Torfone (c), respectively; the lower 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 (b) audio channel (purple region). Harmonics and base ENF signal are also visible in FFT results of Torfone (c).

People can typically hear 20Hz to $20,000$Hz 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. Through this experimental step, we wanted to find out if ENF signals can be extracted and restored after voice passes a VoIP filter.

To visualize the effect of the filter in the audio channel, Figure 10 plots 1D spectrum in the top sub-figures, and 2D spectrograms in the bottom sub-figures. Left, center, and right 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 rage and purple region are located at the multiple harmonic frequencies respectively. The bottom sub-figures shows how the packets are lost during transmission. In the 2D spectrogram, the x-axis represents the temporal index and the y-axis is its corresponding frequencies.

As can be seen in Figure 10-(b), 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. Figure 10-(b) demonstrates that signal at the base frequency has been filtered while it passes through the audio channel of Skype but there are still the peak points in purple region, which are the rest of the harmonic signals and they are preserved by passing through above the cut-off frequency. This can be also shown in Figure 10-(e).

Meanwhile, in the case of Torfone, we can see that base ENF signal also remained around 60Hz as shown in Figure 10-(c). Furthermore, we can observe that packet loss exists in case of Torfone such as shown in Figure 10-(f). 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.

C. *Inter-grid estimation*

Inter-grid estimation is basically classification problem. Given know sample data set with annotated region IDs, we infer the power grid ID of a new sample of interest. For this experiment, we extracted 99 audio streams from the world at the same time, but we removed 31 audio streams which do not have ENF signals. Thus, in our experiments, 68 audio streams are finally obtained and used since they have ENF signals. 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

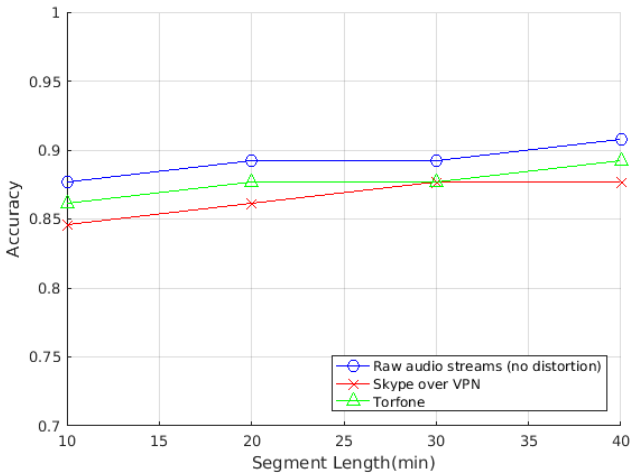


Fig. 11. Accuracy vs Segment length (min) for inter-grid estimation with 40-minutes segment length.

Figure 11. Figure 11 shows that the accuracy of the experiment has the most 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. This is because one of the key features for classifying power grids is the expectation value of ENF signals [25]. Although more information can be included as the segment length increases, the variation of the mean value is not evident for few minutes that we can classify the power grid with small size of ENF sequence.

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 [50].

D. 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 of our interest. Therefore, we extracted target source through Skype and Torfone.

For the evaluation, we set n as the number of anchor nodes in order to reflect the fact that more accurate location information can be estimated as the number of anchor nodes increases. Based on this, we obtained that accuracy is approximately 80 percent when the decision boundary is bigger than 3 as shown in Figure 12. The accuracy almost does not increases for three subjects when the index of decision boundary is after 4 until 16. Thus, we optimally take the presence decision boundary to 4. As we mentioned earlier, we also defined the presence of the decision boundary by the number of n in Figure 8.

We can also obtain accuracy with absolute measure of distance or area. In the case of Eastern power grid of the United States, the approximate total area of Eastern power grid is about $341,754 miles^2$. Considering our relative accuracy is 76% for 3 out of 17 decision boundary, the area of estimated location is $V = 60,309 miles^2$. Since the area of a circle is calculated by $V = \pi\ R^2$, we can calculate the approximate distance from the actual hidden location and the center of the estimated area by $R \approx\ 138.55 mile$.

We 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 III. Still, with the length of 5 minutes of ENF signals, we have the accuracy of 76% in our proposed attack. Furthermore, we can see that the signals passed through Skype is reconstructed better than the signal passed through Torfone. It means the effect by lack of base ENF signal is more critical to attacker than that by packet loss and delay.

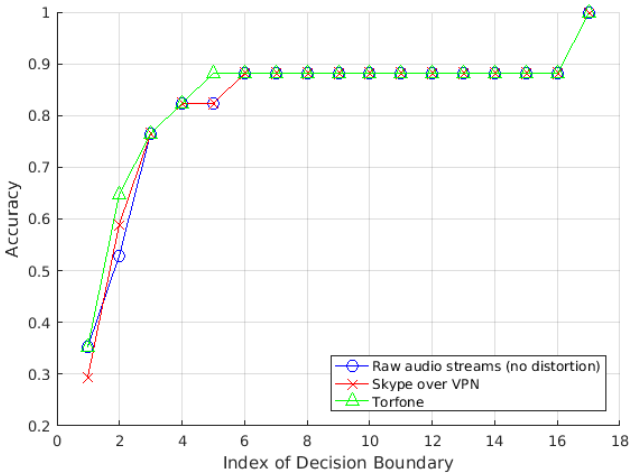


Fig. 12. Accuracy vs Decision Boundary for intra-grid estimation with 40-minutes segment length.

TABLE III. ACCURACY VS SEGMENT LENGTH (MIN) FOR INTRA-GRID ESTIMATION.

Segment length (min)	Accuracy (%)		
	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

VI. DISCUSSION

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

A. 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.

1) Avoiding AC microphones: As mentioned in Section III, the LISTEN attack works on users who are using input devices that capture ENF signals. In particular, the 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.

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 a *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.

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 [40].

B. 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 V-C). 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 III, 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.

C. 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 IV), 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].

VII. Related Work

Inferring user location with side-channel channel attacks is one of the hottest issues in the field of information security. The primary goal of our attack is to reveal the location of streamers and voice chatters with using ENF signals from their recorded sound. Many location estimation techniques using ENF signals are actively researched in recent years. First, we discuss about recent researches which have a different approach for side-channel attack. Then, we discuss about research of which target system is similar with ours, such as VoIP services (e.g., Skype). Finally, we discuss about recent works, that identifies location with ENF has been historically improved.

A. Inferring user location

In Narain et al [37], in mobile phone, they address the approach which can 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 permission. 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, in Michalevsky [36], they use only power consumption for inferring location, even it is not considered as critical as sensor's information from [37]. 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.

B. 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 [53], 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. [11] 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.

C. Location estimation with ENF signals

In this section, we introduce the recent researches for estimating the geo-spatial location with ENF signals from recorded audio for inter-grid and intra-grid.

To classify the ENF signals for inter-grid, some of the researches adopt machine learning algorithms, and some use correlation coefficient. One of good example for using machine learning algorithm is the research by Hajj et al [25]. They use statistical characteristics of ENF signals and auto-regressive model parameters for features, and apply those parameters to soft vector machine (SVM) classifier. In other ways, estimation can be performed with simply calculating the correlation coefficient value between sampled signals and target signals [23].

By the way, recent intra-grid location estimation researches used approximation techniques as we mentioned earlier. Those researches make their assumption and prove them with experimental results. One of the good techniques for intra-grid location identification is well introduced in research by Garg et al [17]. The basic idea of this work is that the correlation coefficient between two ENF signals might roughly inversely proportional to the distance between the points which those ENF signals come from. The first approach of them is to evaluate the distance of estimating point which is on a straight line. They calculate the distance with about 90% of accuracy and prove their assumption. Next is to localization with Half-plane intersection methods. Assume that two points are anchor nodes and we have to estimate the location of target ENF signals. With the principle of correlation coefficient and distance, we can briefly say that which node has higher correlation, which means it is nearer than other anchor nodes. By using those speculate, we can determine where the ENF signals has been captured between one of the planes, that divided by perpendicular bisector of two anchor nodes. They also evaluate their performance with variation boundary of parameters, which is correlation coefficient, while we use it for area.

VIII. Conclusion

Unlike existing location inference techniques [36], [37], [52] 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 results, however, need to be generalized with caution since the current ENF map only covers the Eastern power grid of the United States. As part of future work, we plan to expand the map to cover more locations, and evaluate the attack based on samples collected from areas uncovered in this paper.

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.

Acknowledgment

The authors would like to thank...

References

[1] M. Abe and J. O. Smith III, "Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of fft magnitude peaks," in *Audio Engineering Society Convention 117*. Audio Engineering Society, 2004.

[2] S. and/or Microsoft. (2017) Skype — free calls to friends and family. [Online]. Available: https://www.skype.com/

[3] C. Bettini, X. S. Wang, and S. Jajodia, "Protecting privacy against location-based personal identification," in *Proceedings of the Second VDLB International Conference on Secure Data Management*. Berlin, Heidelberg: Springer, 2005, pp. 185–199.

[4] D. Bykhovsky and A. Cohen, "Electrical network frequency (enf) maximum-likelihood estimation via a multitone harmonic model," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 5, pp. 744–753, 2013.

[5] J. Chai, F. Liu, Z. Yuan, R. W. Conners, and Y. Liu, "Source of enf in battery-powered digital recordings," in *Audio Engineering Society Convention 135*. Convention: AES, Oct 2013, pp. 1–7. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=17055

[6] F.-C. Chang and H.-C. Huang, "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, 2010, pp. 175–178.

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

[8] W. H. Chuang, R. Garg, and M. Wu, "Anti-forensics and countermeasures of electrical network frequency analysis," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 12, pp. 2073–2088, 2013.

[9] A. J. Cooper, "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, 2008.

[10] ——, "An automated approach to the electric network frequency (enf) criterion-theory and practice," *International Journal of Speech, Language and the Law*, vol. 16, no. 2, pp. 193–218, 2009.

[11] B. Coskun and N. Memon, "Tracking encrypted voip calls via robust hashing of network flows," in *International Conference on Acoustics, Speech and Signal Processing*. Dallas, TX, USA: IEEE, March 2010, pp. 1818–1821.

[12] I. EarthCam. (2017) Earthcam - webcam network. [Online]. Available: https://www.earthcam.com/

[13] Facebook. (2017) Facebook live — live video streaming. [Online]. Available: https://live.fb.com/

[14] ——. (2017) Messenger. [Online]. Available: https://www.facebook.com/messenger/

[15] T. A. Foundation. (2017) The largest live nature cam network on the planet world! [Online]. Available: http://explore.org/

[16] S. Gambs, M.-O. Killijian, and M. Núñez Del Prado Cortez, "De-anonymization Attack on Geolocated Data," *Journal of Computer and System Sciences*, vol. 80, no. 8, pp. 1597–1614, 2014.

[17] R. Garg, A. Hajj-Ahmad, and M. Wu, "Geo-location estimation from electrical network frequency signals," in *ICASSP*. Vancouver, BC, Canada: IEEE, 2013, pp. 2862–2866.

[18] V. Gegel. (2012) Tor fone - p2p secure and anonymous voip tool. [Online]. Available: http://torfone.org/

[19] D. Goldschlag, M. Reed, and P. Syverson, "Onion routing," *Communications of the ACM*, vol. 42, no. 2, pp. 39–41, 1999.

[20] C. Grigoras, "Digital audio recording analysis–the electric network frequency criterion," *International Journal of Speech Language and the Law*, vol. 12, no. 1, pp. 63–76, 2005.

[21] ——, "Applications of enf criterion in forensic audio, video, computer and telecommunication analysis," *Forensic Science International*, vol. 167, no. 2, pp. 136–145, 2007.

[22] J. Guo, Y. Ye, Y. Zhang, Y. Lei, and Y. Liu, "Events associated power system oscillations observation based on distribution-level phasor measurements," in *T & D Conference and Exposition*. Chicago, IL, USA: IEEE, April 2014, pp. 1–5.

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

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

[25] A. Hajj-Ahmad, R. Garg, and M. Wu, "Enf-based region-of-recording identification for media signals," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 6, pp. 1125–1136, 2015.

[26] M. Huijbregtse and Z. Geradts, "Using the enf criterion for determining the time of recording of short digital audio recordings," in *International Workshop on Computational Forensics*. Berlin, Heidelberg: Springer, 2009, pp. 116–124.

[27] W. Inc. (2017) Whatsapp. [Online]. Available: https://www.whatsapp.com/

[28] S. Karapantazis and F.-N. Pavlidou, "Voip: A comprehensive survey on a promising technology," *Computer Networks*, vol. 53, no. 12, pp. 2050–2090, 2009.

[29] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2, 1995, pp. 1137–1145.

[30] Krebs on Security, "Skype Now Hides Your Internet Address," https://krebsonsecurity.com/2016/01/skype-now-hides-your-internet-address/, online; January 2016.

[31] Y. Liu and Y. Ye, "Monitoring power system disturbances based on distribution-level phasor measurements," in *PES Innovative Smart Grid Technologies (ISGT)*, vol. 00. Los Alamitos, CA, USA: IEEE Computer Society, 2012, pp. 1–8.

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

[33] G. Y. Lu and D. W. Wong, "An adaptive inverse-distance weighting spatial interpolation technique," *Computers & Geosciences*, vol. 34, no. 9, pp. 1044–1055, 2008.

[34] S. Mann, L. Cuccovillo, P. Aichroth, and C. Dittmar, "Combining ENF Phase Discontinuity Checking and Temporal Pattern Matching for Audio Tampering Detection," in *GI-Jahrestagung*, ser. LNI, M. Horbach, Ed., vol. 220. GI, 2013, pp. 2917=–2927.

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

[36] Y. Michalevsky, A. Schulman, G. A. Veerapandian, D. Boneh, and G. Nakibly, "Powerspy: Location tracking using mobile device power analysis," in *USENIX Security*. Washington, D.C.: USENIX Association, 2015, pp. 785–800.

[37] S. Narain, T. D. Vo-Huu, K. Block, and G. Noubir, "Inferring user routes and locations using zero-permission mobile sensors," in *Symposium on Security and Privacy (S&P)*. San Jose, CA, USA: IEEE Computer Society, 2016, pp. 397–413.

[38] D. W. Oard, M. Wu, K. Kraus, A. Hajj-Ahmad, H. Su, and R. Garg, "It's about time: Projecting temporal metadata for historically significant recordings," *iConference 2014 Proceedings*, 2014.

[39] O. Ojowu Jr, 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, no. 4, pp. 1330–1338, 2012.

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

[41] B. Porat, "Digital processing of random signals: Theory and methods," 1994.

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

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

[44] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 76–87, 2004.

[45] I. Twitch Interactive. (2017) Twitch. [Online]. Available: https://www.twitch.tv/

[46] Twitch Tips, "Protecting Yourself Online," https://twitchtips.com/protecting-yourself/, online; 21 January 2016.

[47] Twitter. (2017) Watch live. [Online]. Available: https://www.periscope.tv/

[48] T. Uhl, "Quality of service in voip communication," *AEU-International Journal of Electronics and Communications*, vol. 58, no. 3, pp. 178–182, 2004.

[49] S. VisioRay. (2017) Skylinewebcams — live cams around the world! [Online]. Available: https://www.skylinewebcams.com/

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

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

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

[53] C. V. Wright, L. Ballard, S. E. Coull, F. Monrose, and G. M. Masson, "Spot me if you can: Uncovering spoken phrases in encrypted voip conversations," in *Symposium on Security and Privacy (sp 2008)*, May 2008, pp. 35–49.

[54] L. YouTube. (2017) Youtube. [Online]. Available: https://www.youtube.com/

[55] Y. Z. L. C. P. N. M. R. M. G. Y. L. Zhiyong Yuan, Tao Xia, "Inter-area oscillation analysis using wide area voltage angle measurements from fnet," *IEEE Power and Energy Society General Meeting*, 2010.

[56] Z. Zhong, C. Xu, B. J. Billian, L. Zhang, S.-J. S. Tsai, R. W. Conners, V. A. Centeno, A. G. Phadke, and Y. Liu, "Power system frequency monitoring network (fnet) implementation," *IEEE Transactions on Power Systems*, vol. 20, no. 4, pp. 1914–1921, 2005.