VoIPLoc: Passive VoIP call provenance via acoustic side-channels

Shishir Nagaraja, Ryan Shah
{first.last}@strath.ac.uk
University of Strathclyde
UK

ABSTRACT

We propose VoIPLoc, a novel location fingerprinting technique and apply it to the VoIP call provenance problem. It exploits echo-location information embedded within VoIP audio to support fine-grained location inference. We found consistent statistical features induced by the echo-reflection characteristics of the location into recorded speech. These features are discernible within traces received at the VoIP destination, enabling location inference. We evaluated VoIPLoc by developing a dataset of audio traces received through VoIP channels over the Tor network. We show that recording locations can be fingerprinted and detected remotely with a low false-positive rate, even when a majority of the audio samples are unlabelled. Finally, we note that the technique is fully passive and thus undetectable, unlike prior art. VoIPLoc is robust to the impact of environmental noise and background sounds, as well as the impact of compressive codecs and network jitter. The technique is also highly scalable and offers several degrees of freedom terms of the fingerprintable space.

CCS CONCEPTS

• **Networks** → **Web protocol security**; **Network security**; • **Security and privacy** → **Pseudonymity, anonymity and untraceability**.

KEYWORDS

VoIP security, call provenance, source identification, location privacy, acoustic fingerprint

ACM Reference Format:
Shishir Nagaraja, Ryan Shah. 2021. VoIPLoc: Passive VoIP call provenance via acoustic side-channels. In *Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '21), June 28–July 2, 2021, Abu Dhabi, United Arab Emirates.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3448300.3467816

1 INTRODUCTION

Voice-over-IP (VoIP) protocols enable millions of individuals to communicate inexpensively regardless of geographic location. The rising popularity of VoIP over the years, especially during the Covid-19 pandemic, has also lead to its abuse by criminal gangs. Cyber-criminals exploit the relative anonymity of VoIP clients as the meta-data regarding caller location can be easily spoofed in VoIP infrastructure. This has lead to efforts to identify the origin of VoIP calls via coarse-grained techniques, such as those that leverage route-specific characteristics of call audio [5], among others. Such efforts will give a rough idea of where a caller is located down to a specific city. VoIPLoc enables fine-grained call location provenance by confirming whether or not a call was made from a specific room in a building, if that room has been previously fingerprinted.

WiSec '21, June 28–July 2, 2021, Abu Dhabi, United Arab Emirates

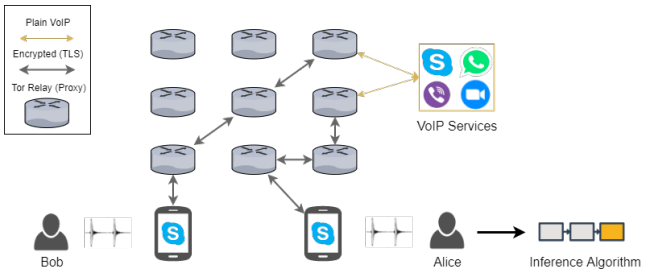


Figure 1: The VoIP-call Provenance Problem

For users of anonymous communications, VoIPLoc location fingerprinting is an attack of concern. Political dissidents coordinating a protest, journalists covering an event, or business persons making deals require location anonymity to ensure operational security. They often use VoIP applications over Tor to hide endpoint information. While Tor hides the source origin on the IP infrastructure, and mitigates coarse-grained attacks, side-channel attacks are a constant worry.

In this work, we demonstrate that VoIP channels inadvertantly advertise unique location fingerprints in the acoustic channel, as a fundamental property of recorded audio. This can be used to mount a sidechannel attack on Tor users using VoIP tools. As a concrete contribution, we developed an attack that exploits this vulnerability to establish call provenance, even over anonymous VoIP channels. The attacker's goal is to establish call provenance – *tracing the original location of an audio source*, as illustrated in Figure 1.

Our contributions are as follows. We describe and evaluate a fully-passive location fingerprinting technique (VoIPLoc) for anoynmous VoIP channels. We then demonstrate fine-grained tracking in call provenance. Finally, we demonstrate our technique's properties of uniqueness, time-invariance, and robustness to variance in network characteristics.

2 THREAT MODEL

The threat model is that the adversary has access to the audio stream, such as recorded speech, at a communication endpoint. For example, an intelligence agency analysing streaming audio from a dissident website, or tracing a dissident activist's location during a phone interview. From a communications perspective, the adversary is an insider engaged in a VoIP conversation with the victim. Such a threat model is reasonable, under the assumption that motivated adversaries would not restrict themselves to launching external attacks.

One of the criticisms of the insider threat model is that it is unrealistic. *Why would a victim hold a VoIP conversation with an attacker?* A

This is a peer-reviewed, accepted author manuscript of the following article: Nagaraja, Shishir ; Shah, Ryan. / VoIPLoc : Passive VoIP call provenance via acoustic side-channels. WiSec 2021 - Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks. New York : Association for Computing Machinery, Inc, 2021. pp. 323-334

common counter-argument to this position is the following stance: any motivated adversary will be an insider. As in most real-world situations, trust isn't binary. A salesperson for a firm dealing in radioactive materials, or surveillance equipment, might be contacted by attackers. Posing as prospective clients, they can gain information about the salesperson's clients after compromising their location privacy. Snowden's revelations famously revealed the tracking of sales personnel by tracking the victim's mobile phone [44]. As opposed to the macro-level location information provided by cell-tower localization, we report attacks that can carry out fine-grained indoor-location identification. Specifically, down to a specific room or corridor the victim made a voice call in.

VoIPloc can be also used to identify the location of attackers involved in running spam campaigns, remotely carrying out Covid-19 quarantine check-ins, or other types of cybercrime and fraud. These types of attacks engage victims via audio calls. Thus both legitimate and illegitimate actors should be concerned about their location privacy. The undetectability of fully passive location fingerprinting attacks poses a significant threat to online privacy of VoIP participants.

We have specifically chosen to study passive adversaries to demonstrate the power of the attacks. Even a passive attacker can confirm the source-location of an audio call. The passive attacker model studies the lower bound of attacker success, given highly restrictive circumstances for location identification. Speech codecs apply band-pass filters to compress audio, which excludes techniques that use inaudible frequency bands such as ultrasound [9] or LF/VLF band audio (eg. the buzz of mains frequency). This forces active attacks to use audible frequencies, lacking in stealth and thus showcasing the importance of passive attacks.

2.1 Requirements

Following the threat model, we now establish our requirements for practical VoIP call provenance techniques:

(1) *Stealth* – Provenance should be established via passive methods, in order to prevent detection by call participants. With an obvious stealth advantage, passive attacks are less observable than active attacks. Further, they are more robust to the compressive effects of codecs, which may remove ultrasound and infrasound components due to the deployment of aggressive band-pass filters.
(2) *Fine-grained tracking* – Provenance should be established down to the specific room used to make a call, thus affording the caller the smallest anonymity set.
(3) *Uniqueness* – A provenance fingerprint must be distinct.
(4) *Time invariance* – Provenance fingerprints should not rely on leveraging background sounds for fingerprinting. For example, relying on proximity to an external noise source that might be unavailabl can result in an unreliable fingerprint.
(5) *Robustness* – Provenance fingerprints should be robust to the presence of background noise, such as HVAC and fans. The source (i.e victim) should not be required to carry specialised hardware in order to reliably establish provenance.

3 ATTACK TECHNIQUE

The attack workflow consisting of decomposition and classification is shown in Figure 2. Location fingerprints are computed by extracting acoustic-reflection characteristics of the location, whilst

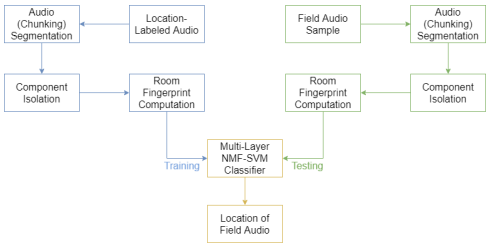


Figure 2: Attack workflow

filtering out all background sounds. A multi-layer (deep) classifier then maps the fingerprint to a location by comparing it to a database of previously pre-labeled audio samples.

3.1 Preprocessing step – Audio segmentation

The initial step in the attack is the collection of audio traces by the attacker, as an insider on a VoIP call. Thereafter, the attacker proceeds to segment audio traces into chunks, such that each chunk corresponds to a single *utterance* – the smallest unit of human speech with a brief silence at either end. We followed standard practice of Ishizuka et al. [22] towards segmentation, by passing the traces through a voice activity detector and a silence detector [48]. Successive bursts of spectral flatness, with a burst of spectral energy in between, is used to differentiate between voice activity and silence.

3.2 Multi-layer decomposition – Isolating location-specific signal components

The next step is to isolate the relevant signal component that is a function of the location from each chunk.

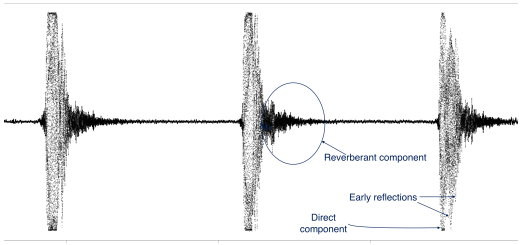


Figure 3: Components of a sound sample

Each audio chunk is made from three signal components (see Figure 3). First, *direct sound* is the sound transmitted in a direct path, from the speaker to the microphone with no reflections. Second, *reflections* follow direct sound. These are distinct reflected sounds that arrive at the VoIP sender's microphone along a predictable path. Third, the *reverberant component* is composed of higher order reflections, which are a combined function of all the room surfaces. To fingerprint a location, the reverberant component is the most relevant. This is because it is a stable function of acoustic information, diffused throughout the location. Isolating the reverberant component is non-trivial due to the time-overlapping nature of the three components.

As the speaker utters a sound, the direct-sound components overlaps with the early-reflections and reverberation components from previously spoken syllables. This is because the reflections from a previously spoken syllable are still above the noise floor when the current syllable is uttered. The challenge of isolating reverberant

component is therefore to decompose a given speech segment back into the reverberant component, early reflections, and direct-sound. Since we are only interested in the reverberant component (which is a function of the room rather than the speaker), it is critical that any direct-sound components and early reflections are removed from the reverberant component. This is a technical challenge that a fine-grained fingerprinting technique using human voice must solve.

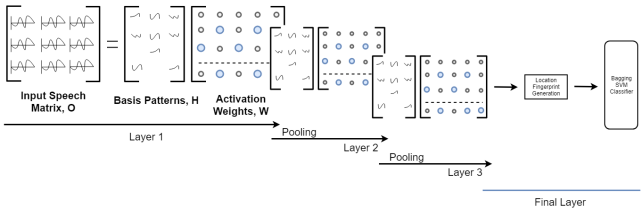


Figure 4: Multi-layer Inference Algorithm

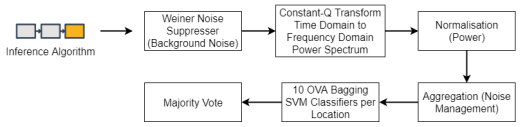


Figure 5: Fingerprint computation after Inference

To isolate overlapping components and suppress background noise sources, VoIPLoc uses a compressive decomposition technique (Figure 4), which partitions recorded audio into direct sound, reflections, and the reverberant component. Partitioning leverages the observation that the shape and form of reflections are, by definition, very similar to direct sound. As such, early reflections can be constructed using the same direct sound and the addition of appropriate location-specific transformation. Partitioning has three stages: (a)*K layers of nested NMF decomposition with pooling*; (b) *Fingerprint generation function*; and (c) the *non-linear classifier.*

3.2.1 Decomposition. The decomposition approach in this section was originally developed by Nagaraja et al. [41] for binary analysis. We extend this to the non-binary case and significantly augment it.

In preparation for decomposition, the attacker loads audio traces into the audio matrix O, where each row correponds to one chunk. Each chunk is split into t time intervals. Within each time interval the attacker computes the average signal amplitude, thus yeilding a vector of signal amplitudes (in dB).

First, input traces are expressed as a linear combination of direct sound and location-specific transformations (H), and mixing weights (W) as: $O_{ij} = \sum_{k=1}^{K} H_{ik} W_{kj}$. The key intuition here is to represent speech segments using the sparsest and fewest number of speech sources. Maximising sparsity is the basis for compressing a signal containing the direct sound and their reflections back into a handful of direct-sound patterns and the intensities of reflection. Within each layer, this is governed by the *optimisation* function which is formally stated as: $\min||O^1 - H^1 W^1||$ such that $W^1_{ij} \geq 0, H^1_{ij} \geq 0$, i.e. W^1 and H^1 are non-negative, hence the name Non-Negative Matrix factorisation, where $||\odot||$ is the Frobenius norm. *Optimisation* is carried out by starting out with randomly initialised positive-valued matrices W^1 and H^1, and updating them iteratively using (standard practice) multiplicative update rules [29]: $H_{ir} = \frac{H_{ir} \sum_r \frac{O_{ij}}{(HW)_{ij}} H_{rj}}{\sum_l H_{jr}}$ and $W_{rj} = W_{rj} \sum_i H_{ir} \frac{O_{ij}}{(HW)_{ij}}$. By deploying multiple layers of decomposition we can further promote sparsity. This motivates the use of Deep Neural Networks (DNN) [27]. In each layer, sound components are partitioned into sub-components, sub-sub-components, and so on.

Second, we use pooling to build robustness to the position and movement of the speaker. Thus each decomposition layer is followed by max-pooling, a moving-window function which takes the rows of the weight matrix W^k as input, and replaces a subset of the row by the maximum value of the subset. This approach was first suggested by Boureau et al. [6]. The pooling function for row i of weight matrix W in layer k as $F(W^k_{i*}) = \max W^k_{ip} | j-c \leq p \leq j+c, 0 \leq j \leq N$, where N is the number of columns of the input audio-traffic matrix O^1 and c is a constant. Accordingly, $F(W^k) = (F(W^k_{1*}, ..., W^k_{M*})$ is the pooling function for matrix W^k. In our evaluation, we used a value of $c = 20$, which corresponds to a moving window of 20Hz. Upon convergence, the columns of H^1, with normalised weight $\sum W^X_{i*} / \sum W^X$ greater than 0.9, contain direct-sound signals, while the rest correspond to the reverberant component, where $W^X = H^2 ... H^K W^K$.

The final step is to isolate the reverberant component. We regenerate the sound signal using all but the patterns corresponding to direct sound in H^1. The columns of H corresponding to direct sound are set to zero and the signal O_r is regenerated $O' = H^1 W^X$. Non-zero rows of O' contain the reverberant component.

An important system parameter is the choice of number of columns of H, i.e. the number of basis patterns. This should roughly be set to constant times the number of possible sound sources that are simultaneously active in a location. For instance, if victim is a single VoIP caller sitting alone in a sound-proof room, then the number of sources should be at least 5 (one for the speaker, and a few for the reverberant signal and noise), there is no upper limit. We set the number heuristically at 100 in all our experiments.

3.3 Fingerprint computation

Fingerprint computation (Figure 5) consists of several steps: noise suppression, signal aggregation and classification.

3.3.1 Noise suppression. The reverberant component computed thus far contains transformation noise (from the reflections) and environmental noise, both of which must be removed. We used a Wiener noise suppressor [45], to remove the influence of background noise on the location fingerprint. This approach uses harmonic regeneration noise reduction (HRNR), to refine the signal-to-noise ratio before applying spectral gain to preserve speech harmonics.

3.3.2 Reflection measurement. The key idea underlying location fingerprint computation is to compute the signal power in each frequency band, within the reverberant component normalised by the corresponding signal power within the direct sound components i.e we measure signal attenuation as a function of room geometry.

The standard tool for frequency analysis is the Fourier transform. For digital signals, the textbook approach to ascertain signal power by frequency band is to apply the Discrete Fourier Transform; often using the Fast Fourier Transform (FFT) algorithm. However, FFT is unsuitable for our purpose. The average minimum (fundamental) frequency for human speech varies from 80 to 260 Hertz; 85 to 180 Hertz for Basal and Tenor voices and from 165 to 255 Hertz for Contral to Soprano voices. Therefore, using FFT the frequency resolution

would be insufficient. FFT with 512 temporal samples recorded at a sampling rate of 44.1 KHz, has resolution of 86.1 Hz between two FFT samples. This is not sufficient for low frequencies found in human voice. For instance, the distance between two adjacent vocal tones could be as low as 8 Hz to 16 Hz. The frequency resolution can be improved by using a higher number of FFT samples. For instance, with 8192 temporal samples, the resolution will be improved to 5.4Hz for a sampling rate of 44.1 KHz. However, this alone is inadequate since the signal at higher frequencies will have better resolution than those at lower frequencies.

To provide constant frequency-to-resolution ratio for each frequency band, we use the Constant-Q transform [7]. This is similar to the Discrete Fourier Transform but with a crucial difference – it permits the use of a variable window width to achieve constant resolution, enabling effective coverage across the spectrum. Constant resolution is achieved via a logarithmic frequency scale. The CQT transform of the reverberant component is computed, after due isolation using the technique described in Section 3.2. The output of the transform is the *fingerprint vector* which contains the signal power in each frequency band.

3.3.3 Normalisation. The fingerprint vector is normalised to remove biases arising from variability in the input-signal amplitude; some speakers speak louder while some speak softly, amplitude variance can also arise from speaker movement. The fingerprint vector computed thus far is normalised by the signal amplitude of the direct sound component in the corresponding band via element-wise division of the CQT transform of the reverberant signal (R) by the CQT transform of the direct-sound signal (D). For each speech segment i, matrix R_{i*} is the CQT transform vector of the reverberant component, and matrix D_{i*} stores the CQT transform vector of the direct-sound component. We then aggregate the normalised vectors from each speech segment, to maximise the range of frequencies that can be used in the fingerprint. Thus we merge multiple CQT vectors computed over respective reverberation components, as follows. Vector p contains the normalised aggregated fingerprint. For each segment i and frequency j, we compute $P_{ij} = \frac{R_{ij}}{D_{ij}} | D_{ij} > 0$ and $P_{ij} = 0 \forall D_i \leq 0$. The signal power is then added up $p_j = \sum_i^n P_{ij}$.

3.4 Final Layer – Non-linear separation

The final layer is a supervised classification layer that maps the outputs of the fingperprint generator onto a non-linear multi-dimensional space. The fingerprint vector is input into an ensemble of weak *bagging* classifiers. Their role is to map the input to a location label. The classifier is trained positively with location traces and negatively against traces from other locations.

Pre-filtering for contamination-resistance: It is important to ensure that negative training sets are minimally contaminated with positive samples to avoid overlapping label definitions. To achieve this, we applied the (parameter free) Xmeans algorithm to cluster vectors in the negative training set on a per-set basis. Any cluster with labelled positive fingerprints is discarded from the negative training set.

Classifier choice: We chose a non-linear kernel function within an SVM classifier. The kernel function uses exponentiation of the Euclidean distance to ensure linearity in the locality of a fingerprint vector: $K(x_i, x_j) = e^{(-\gamma ||x_i - x_j||^2)}$, where γ is the width of the Gaussian function. Our classifier choice is driven by the following reasons: first, the dataset is *small* as only a few samples per room-location are likely to be available for training. Second, the dataset is *sparse* due to the high dimensionality of fingerprint vectors. Thus an classifier like SVM will have no problem identifiing separating hyperplanes that maximises the *margin of separation* between data vectors p with a low overlap across class boundaries. Third, the classifier must be robust to the presence of unlabelled data during testing. For instance, in the context of deanonymising a VoIP caller using an anonymous communication channel, a call could be made from a number of locations that are not in the fingerprint database. This requirement can be addressed within an SVM framework by using a bagging approach – instead of using just one classifier, we train ten classifiers per location, in One-vs-All (OVA) mode where each of the ten classifiers for the ith room are trained using a tenth of the samples from the training set for the ith room with a positive label, and a randomly chosen tenth of all the remaining samples (of the training set including unlabelled samples) with a negative room label. In other words, random resamples (of a 10th) of the unlabelled data are drawn and the classifiers are trained to discriminate the positive room sample from each resample. Resampling unlabelled locations induces variability in classifier performance which the aggregation procedure, used to combine the outputs of individual classifiers, can then exploit. **Aggregation:** The results of the classifiers for i^{th} room are aggregated with a majority vote. Overall, for n rooms we train a total of $n \times 10$ classifiers, which is still $O(n)$ classifiers. This method of combining classifier output is based on the technique first introduced by Mordelet and Vert [38].

4 EVALUATION

In this section, we evaluate the attack technique using VoIP conversations in a diverse set of locations, codecs, network jitter, and speech characteristics using a corpus of recordings.

4.1 Real-world dataset

Our first dataset consists of audio recordings of VoIP sessions conducted over the Tor network, from 79 rooms of identical geometry of a university computer science department. Occupants customise these rooms using furnishings such as desks, bookshelves, monitors, and other objects that affect room acoustics but are otherwise identical. The impact of these customizations on the reverberant component of recorded audio forms the basis for location identification. Our goal is to understand the extent to which our VoIPLoc exploits differences in acoustic absorption/reflection characteristics whilst tolerating acoustic and network jitter. The rooms have typical (acoustic) noise sources which could be continuous such as air conditioning systems, heater fans, and fridges, or intermittent noise from road traffic or human subjects in the vicinity. The dataset was generated in 2016 and used the public Tor network for experiments.

A VoIP session (see Figure 1) is set up over the Tor network between the sender (UK) at the given location and a recipient on the other end (Davis, CA, USA). At the receiver the resulting audio stream is recorded. We recruited sixteen volunteers to conduct VoIP sessions in each of the 79 rooms and recorded the audio at the recipient end. The (sender) volunteers were selected for diversity in voice pitch (*8 male and 8 female*). For each room, we seated each of the volunteers at nine different positions located at the intersections of a 3x3 grid (rectangular) control for position-specific bias.

Volunteers were instructed to remain in a neutral tone and hold a conversation, moving naturally, whilst reading out from a script from the NXT Switchboard Corpus [8] consisting of telephone conversations between speakers of American English. It is one of the longest-standing corpora of fully spontaneous speech. We used the MS-State transcript of the corpus, and all volunteers read the same transcript for consistency. The corpus has transcripts that support conversations of different lengths for diversity.

Training and testing sets: Our dataset contains a significant number of audio traces (288) per location, most of which are utilised for testing and a small fraction are used for training (a credible attack cannot depend on more than a couple of audio samples). We partitioned the dataset into k different non-intersecting sets (by random allocation with uniform probability) on a per-location basis. For each location, one set is used for training and $k-1$ are used for testing, this is repeated $k=50$ times so that every subset is used for testing (standard k foldover cross-validation).

4.2 Codecs

A wide variety of VoIP clients are in popular use thus we are interested in the impact of speech codecs on fingerprinting. Codecs apply a range of techniques such as compression and variable sampling rates to efficiently encode as much of the speech information as possible under assumed steady state network conditions. The resulting compression presents a significant challenge for remote fingerprinting due to the potential loss of relevant signal information. Indeed many codecs apply a variable cutoff *high-pass filter* to remove ambient sounds – low frequency background sounds and breathing noise.

SILK codec (Skype): The widely used Skype VoIP services uses SILK codec [23] as well as other proprietary voice codecs for encoding high frequencies in the range of 16KHz. The key parameter for our purposes is the target bitrate. SILK's signal bandwidth (frequency range) varies in time depending on network conditions. When throughput is low, lower bitrates are used, and the codec enters a Narrowband mode wherein the sampling rate is set to 8KHz and the signal bandwidth covers 300-3400Hz. In this mode, higher frequencies are not transmitted, potentially affecting the performance of our fingerprinting techniques. The range of frequencies skipped in this manner depends on the network throughput. Internally, SILK supports 8, 12, 16, and 25KHz resulting in bitrates from 6 to 40 Kbps.

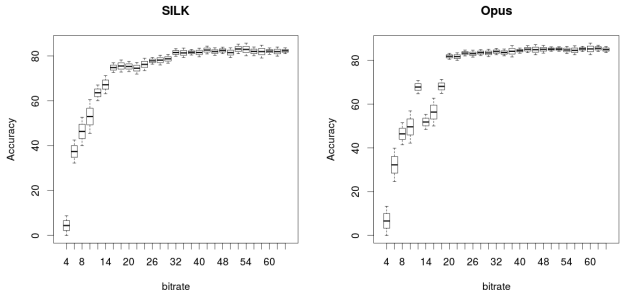


Figure 6: Location accuracy vs bitrate

Figure 6 shows the impact of bitrate on fingerprinting efficiency (detection rate) for various codecs. We observe a rather low efficiency 38%, at very low bitrates in the range of 6–10Kbps. At around 14–16Kbps, we observed an increase in attack efficiency to 73%. This is interesting and linked to the increase in sampling rate to 12KHz around 10–12Kbps when SILK assumes a wider signal bandwidth of 6KHz. A steady improvement in attack efficiency is noted, as the sampling rate improves transmitting a greater part of the signal bandwidth to the receiver. Another increase (to 77%) is noted at 24Kbps, when the codec switches to 24KHz sampling rate internally, also known as the superwideband mode in SILK parlance, stabilizing to 82% at 40Kbps. At higher bitrates, SILK is able to support a wider band of frequencies, allowing a larger fraction of the signal features to be transmitted which improves fingerprinting.

Opus (Facebook Messenger/Zoom): The Opus codec is a framework for composing high quality codecs, namely SILK [23] and CELT [34]. It operates in three modes: SILK mode, a new hybrid mode, and CELT mode. In the SILK mode, it supports narrow to wide frequency bandwidths, with relatively low-bit rates. The CELT mode is a high-bitrate consuming codec offering a greater bandwidth than the SILK mode. We observe a detection rate of 50% at 10Kbps in LP mode. At 12Kbps, we observe a significant improvement of 20% in fingerprinting efficiency to 70% (which is in the realm of usefulness). This is the threshold when Opus switches to Hybrid (wide band) mode i.e from lossy to lossless compression, once again confirming the importance of mid-range frequencies in the accuracy of room fingerprinting. This is of interest, since it's meant to fill the gap between LP mode and the MDCT mode. As the bitrate increases, the signal bandwidth increases, leading to greater fidelity at the receiver. At 14Kbps, the Opus codec shifts from LP to hybrid mode, entering a lossy compression stage onece again, resulting in reduced attack effectiveness compared to the LP mode. At around 18Kbps, the codec recovers to the same level as LP mode at 12Kbps. A second threshold increase is noted at 20Kbps as the hybrid mode starts to support the super-wideband frequency range. Gradual further improvement is noted to 85% which is fairly close to the baseline (no compression) figure of 87% accuracy. This is achieved when the bitrate is high enough (> 48Kbps) to allow lossless compression in CELT mode super-wideband.

A frame length of 20ms, at constant bit rate, was used in all experiments. Opus supports short (2.5ms) and long (60ms) frame lengths. The shorter the frame, the higher the bitrate. Further, Opus supports redundant information, which improves quality at a cost of higher bitrate allowing the decoder to recover against frame losses due to random faults. In addition to frame length adjustment and redundant information, Opus also supports multiple frame packetization. This improves coding efficiency by reducing the number of packet headers needed per second at the cost of additional delay. Overall, we have focused our analysis on the impact of bitrate and assumed the network path is free of significant variations in jitter and other error conditions. We relax this assumption in Section 4.4.

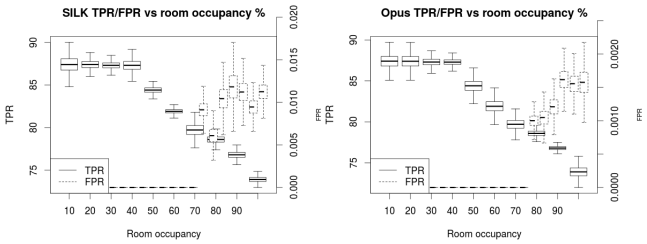
This is sustainable under the assumption that the significant parameter is variable network bandwidth available to the VoIP application resulting in variable bitrate. As part of future work in the area, we plan comprehensive analysis involving other parameters, namely redundancy, frame length, jitter, look ahead, and training and testing on different conditions influenced by these parameters.

4.3 Impact of room occupancy

Next we study the impact of time-variant properties of the location such as the number of people present and the *speaker movement*.

Conti et al. [11] showed that human bodies are good multi-directional reflectors assisting in the mixing of sound within a room. They demonstrated that the extent of reflection is solely described by the mass of the human which acts as a rigid water-filled ellipsoid. Absorption by the human body is largely dominated by the amount and type of clothing — naked bodies will reflect entirely while thick winter coats will increase signal absorption. Since the fingerprinting attack is primarily sensitive to the absorption characteristics of the room, the expected impact is minimal in most circumstances. However, as reflectors, the scattering of sound from the human body can cause new influences. For instance, if occupants block a reflective surface such as a whiteboard, it could change the room's fingerprint.

To study the impact of room occupancy on location fingerprinting, we ran experiments on a set of 18 meeting rooms with various rates of occupancy – this is the number of people in the room measured as the proportion of the maximum seating capacity of the room. We then tested VoIPLoc's efficiency at various rates of occupancy. For each room, we set up a VoIP call over Tor over a period of 8 to 12 minutes. We collected sound traces specifically for training by having each speaker read a standard script (with consent).



(a) SILK **(b) Opus**

Figure 7: Location accuracy vs room occupancy

Figure 7 shows that the detection rate gradually decreases at around 50% occupancy. We note a detection rate of over 85% for all occupancies below 50%, decreasing to above 70% for 100% occupancy. Our observations indicate that occupancy has impact but the attack is still a credible threat even at high occupancy rates. In the case of lecture and meeting rooms, all occupants were seated and the speaker was standing.

To confirm our hypothesis that attack accuracy is sensitive to absorption, we carried out another experiment. We examined the efficiency of attack when the VoIP sender (speaker) is surrounded by other people. For instance, when a hostage speaking whilst surrounded by kidnappers, or team participants huddled around a microphone during a conference call. We observed that the detection rate incurs a threshold decrease to 50% after 40% of the surrounding space is blocked. Further, as the surrounding space is progressively blocked, the impact of the walls on the room behaviour is *fully* replaced by the sound scattering properties of the occupants. In the worst case, only a minimal amount of reverberation is created. The detection rate does not decrease to zero because sound being a pressure wave can bend around obstacles although this attenuates the signal.

4.4 Robustness to network jitter

VoIP traffic flows are routed over the Internet as a sequence of packets. In the process, flows can experience variability in the inter-arrival times of packets (jitter), experience loss of packets, and variation in throughput due to dynamic router work-loads. Packets that arrive too late at the destination are not played out (discarded).

Packet loss %	SILK (TPR%/FPR%)	OPUS (TPR%/FPR%)
10	83.64/0.60	92.10/0.42
15	82.14/1.94	86.97/0.74
20	70.40/3.32	83.32/0.84
25	55.46/16.34	75.19/1.67
30	33.11/17.46	45.24/5.88

Table 1: Impact of packet-loss on attack efficiency

Since packet delays and losses reduce audio quality, most codecs used by secure messaging systems implement a (packet) loss concealment strategy to maintain a perceptual level of voice quality despite any residual packet loss. Both OPUS and SILK codecs generate a replacement signal using the frequency spectrum of recent segments. For instance, substituting the missing signal with another signal with identical frequency spectrum, whilst replicating the pitch waveform from a recently received speech-segment signal. We note that jitter by itself does not affect the attack efficiency, since the packets arriving late can still be leveraged for fingerprint construction, although they are not played out. Hence, the focus of our analysis is on missing packets rather than delayed ones.

We introduced packet losses at various rates and observed changes in attack efficiency using a configurable router. A Pica8 3920 SDN switch was used to routing flows between source and destination pairs. The switch was programmed to drop packets from the source-destination flows at a selected rate of packet loss according to a poisson distribution, which has shown to be a realistic assumption for standard network traffic in applications such as arrival of HTTP/VoIP sessions [35, 39]. For packet loss rates of 10%, we find that the attack efficiency is fairly high with low enough FPR and reasonable detection rates of above 90% in the case of Opus. In the case of SILK codec, the detection rates are around 80% for 10% loss, which then reduce to 70% for a loss rate of 20%. For higher rates of packet loss, attack efficiency is severely degraded in both cases to less than 50%. More, importantly we note that the FPR in Opus is relatively stable, being less than 1% until medium levels of loss (10%), increasing only to 8% for 30% loss. The attack efficiency degrades faster when operating via the SILK codec for increasing losses; beyond 10% loss-rates, FPR degrades to 14–17% which is high. The reason for the higher attacker efficiency via Opus is because of a dynamic jitter buffer. When frames arrive after the length of the jitter buffer they are discarded. In the case of Opus, the codec adapts to lossy network conditions by embedding packet information into subsequent packets allowing significantly better reconstruction rates and hence enhanced attack efficiency in comparison with SILK.

To understand packet loss rates, Guéguin et al. [19] studied the mean opinion score, a well respected metric to measure speech quality. Packet losses of less the 1% are required to ensure no audible losses. A loss rate of 10% is considered poor with users being considerably annoyed at and beyond this point. We note therefore that if the VoIP connection delivers good to excellent voice quality then it is a fit candidate for fingerprinting purposes with low FPR.

4.5 Robustness to location diversity

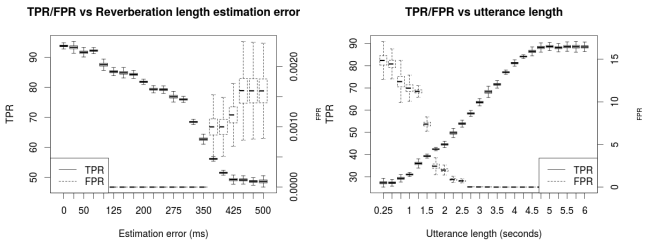
In previous sections, we have discussed the effectiveness of fingerprinting very similar locations which is the harder case for fingerprinting. However, the real world is much more diverse. Therefore we now add *diverse* locations with very different characteristics

Room	Description
Building Atrium (enclosed courtyard)	Semi-enclosed Atrium of the CS department. The courtyard is enclosed on three sides by partition board walls and on the third side by a glass wall. Volume is roughly $24000m^3$.
University Sports Centre	A large sports hall in the sports center of the university. The reverberation length is fairly long. Volume is $12000m^3$.
Typing Room	Partitioned typing room in the British Museum.
Office rooms	79 rooms from a university CS department
Stairway	Stairway within an office building.
Concrete studio	Bare room with plastered walls, concrete floor, and concrete ceiling.
Underground car park	Cemented, with pillars, $8500m^3$
Domestic living room	Wood and lathe, $1100m^3$.
Wood paneled room	Wood paneled studio room, $1400m^3$
Stanbrook Abbey Malvern	Large hall within a former monastery, $61000m^3$.
Factory warehouse	Large empty warehouse inside the Cadbury factory, $4800m^3$.

Table 2: Indoor locations considered

to the fingerprint database and evaluate the effectiveness of the technique with this change. We selected locations with markedly different acoustic characteristics owing to significant differences in physical size, shape, volume, and construction material. Our goal is to understand whether sound samples from the same room can be linked given a population of VoIP traces from a diverse range of indoor venues. Our dataset consists of sound samples collected in locations given in Table 2.

Location diversity also gives us the opportunity to study additional parameters. Reverberation length (the length of the reverberant component) is shorter in meeting rooms as opposed to warehouses or lecture rooms. Diverse room sizes also induce variability in reverberation length so it's important to accurately estimate this. Additionally, we also consider the impact of utterance length i.e the amount of time for which signal power is above the threshold of silence. An utterance corresponds to one or more words spoken together such that the signal amplitude does not fall to the noise floor (0 dB). Long utterances have higher aggregate signal power compared to shorter utterances which makes fingerprinting easier.



(a) Impact of RT estimation error **(b) Impact of utterance length**

Figure 8: Impact of Parameters on Accuracy (Opus)

Parameter estimation The classifier depends on accurate estimation of reverberation time to isolate the reverberant component. In practice, this parameter must be estimated without having any information other than the signal itself. Our method estimates this parameter, with some errors. The amount of error depends on the noise level within the signal. Hence, we evaluate how poor reverberation-time estimation impacts overall accuracy. We compare the error in room identification as a function of estimation error. We vary the error in milliseconds (ms) from the ground truth. The results are shown in Figure 8a. The results show that if there is a very small error ($100ms$) in estimation, then location identification accuracy is above 87%. If the estimation error is larger ($200ms$—$300ms$), then the accuracy drops to around 75%. In indoor offices and residence halls, we found the estimation error to be within $60ms$. In the stairway, the error was $50ms$, and around $400ms$ in the warehouse. The key insight here, is that the quality of separation between datapoints corresponding to different classes (separating hyperplane) was of sufficiently high quality as to enable a high detection rate that decayed linearly with quality of capture of the reverberant component (higher the RT estimation error, lower the quality of input datapoint into the classifier, hence lower the classification output). A second insight is that we observe a threshold effect in the false-positive rate; the false-positive rate is zero until RT error is $350ms$ but experiences a threshold increase at $375ms$. Until the threshold value, quality separation between the various data-point categories (location classes) cushions the impact of the estimation error on the false-positive rate. Beyond the threshold, the input fingerprint candidate vectors are simply noise and get categorized together leading to a significant rise in the false-positive rate.

Utterance length . A second parameter is the minimum signal power conveyed by the speaker in a single word or sentence (utterance). If the utterance length is very small, the signal strength is too low for fingerprinting. A human utterance is a consecutive set of speech segments. During data collection, we observed that the length of an utterance varies in duration (a well established fact in the literature). Our classification technique works by extracting the statistics of the reverberant component across the length of an utterance. Conversations with long utterances increase accuracy of the classification. To understand the impact of utterance length, we used the conversations with varying lengths and observed classifier accuracy. In figure 8b, we observe that if the length of utterances is greater than 3.8 seconds, we obtain over 78% accuracy for the classifier, while for utterances lasting less than 1 second (monosyllable words) the accuracy is less than 28–30%.

4.6 Scalability

The ability to fingerprint a location is only as useful as the number of locations that can be uniquely fingerprinted. One concern is that VoIPLoc may perform less accurately with a larger number of rooms as it may become easier for fingerprints to "collide".

To get a sense of performance over a larger number of locations we used a large-scale synthetic location dataset comprising rooms with realistic room and furniture layouts. Each room is manually filtered for its realism by a set of three volunteers recruited via Mechanical Turk. If at least two volunteers vote in favour of a synthetic room layout, it is considered realistic. After filtering, the number of realistic furnished rooms in the dataset are 404,058. We generated the synthetic traces as follows: for each room, we generated the room's characteristic function using the Lehmann-Johansson method [30]. We convolved the function with an anechoic audio recording to generate the audio traces containing direct sound, overlapped with reflections, further overlapped with a reverberant component. Each audio trace is therefore a realistic emulation of the room constructed with real building materials with the exclusion of background noise. The Lehmann-Johansson method uses the geometric dimensions (length, breadth, and height) of a room, along with speaker position and orientation, and the reflection coefficients of the room surfaces. We set the reflection coefficients using industry standard values for various labelled materials [57] as per the dataset. As with the real-world dataset in Section 4.1, we vary orientation and position.

We used the voices of 16 speakers from the NXT corpus [8] (also used in section 4.1, to obtain 1.4 billion. Each recording is transmitted via a secure messaging client (Skype) communicating to a recipient over a Tor circuit. The location fingerprint is computed over the audio data at the recipient. The traces are then input to the classifier. As before (4.1), we partition the dataset into k non-overlapping subsets, with one subset being used for training and $k-1$ for testing. Each subset contains 5 traces per location. The results are averaged over k iterations as in a k-foldover cross-validation.

Codec	Room count	% FP	% Detected
SILK	100	0.000	87.57
	1000	0.047	87.65
	10000	0.042	88.76
	100000	0.008	88.93
	404058	0.007	89.98
Opus	100	0.001	90.76
	1000	0.037	91.28
	10000	0.024	91.41
	100000	0.003	91.52
	404058	0.003	91.77

Table 3: Scalability results on 404,000 rooms

Codec	Room count	% FP	% Detected
SILK	1000	0.053	83.24
	10000	0.014	84.38
	100000	0.009	84.59
Opus	1000	0.026	88.27
	10000	0.018	89.36
	100000	0.007	89.39

Table 4: Scalability — results with partial traces

Overall, we found that VoIPLoc scales well with the size of the location-database, with performance remaining stable as the number of locations increases. For example (in Opus), the false-positive rate (FPR) for 10000 rooms is 0.02% (Table 3), while for 100000 rooms this rate is 0.003%. The FPR decreases roughly by a factor of 10, which is equal to the scale-up factor between the two experiments, indicating that the actual number of false-positives remains the same.

Training only with partial traces: In the experiments we have performed so far, the attacker has access to audio traces generated from various victim positions and orientations. However, in practice such a diverse coverage of a room's acoustic characteristics may not be available for sampling. It is therefore useful to evaluate how well VoIPLoc would work with datasets where only a fraction of the possible audio traces are known for training. The training set is as follows: conservatively, we assumed the victim is at one randomly chosen position of nine and just three randomly chosen orientations out of twenty. This gives reasonable room for head movment in a VoIP session. We therefore removed 97% of the audio traces from our dataset. Table 4 documents the effects of partial data. While the detection rate falls, VoIPLoc still compromises the location-privacy of over 83–88% of users with high reliability. In practice, the number of audio traces available per location will likely be higher.

Effects of unlabelled data: Up until now, we have trained VoIPLoc with a label for each location it can expect to be tested with. However, in the real-world this won't be the case since all possible locations won't have been fingerprinted. It is therefore important to consider how missing locations will impact VoIPLoc's performance. We simulate testing in the open-world setting by removing a fraction of locations from the training phase whilst retaining them during the testing phase. Table 5 documents the impact on VoIPLoc's performance. We removed an increasing number of location labels from fingerprints in the training set but retained them as unlabelled fingerprints in the testing set. Broadly, we find that algorithmic efficiency is not impacted by low to medium levels of noise from unlabelled locations. Notably the FPR is 26–50 per 100000 locations. Efficiency reduces by around 10% compared to the fully labelled set. We posit the unlabelled set is contaminated with fingerprints that are actually positive, causing a reduction in the efficiency of the classifier. To reduce the impact of this issue, the pre-filtering step in Section 3.4 removed positive fingerprints from the unlabelled set which prevents large-scale contamination.

Codec	Missing location count	% FP	% Detected
SILK	101000	0.049	78.46
	202000	0.040	79.53
	303000	0.040	79.29
	383800	0.047	79.80
Opus	101000	0.019	83.76
	202000	0.024	83.03
	303000	0.026	83.95
	383000	0.026	83.81

Table 5: Open-world – results with 404k locations with both unlabelled and partial traces

5 DISCUSSION

Our experiments confirm the hypothesis that the reverberant sound component can be used to generate location fingerprints, with high reliability and low false-positive rates of detection. We evaluated with rooms of identical geometry which are differentiated only by the customisation introduced by their occupants, such as the placement of monitors and the number of books on their shelves. The attack technique uses a deep NMF-SVM classifier, which was trained on a few samples per room, and then tested extensively against samples recorded in different parts of the location. This indicates that a fingerprinting technique can be used to reliably link an audio traces recorded at different parts of the same location. Given the low false-positive rates (0.003%), we documented the location accuracy in terms of the detection rate alone. While we expected aggressive audio compression employed by the codecs to significantly damage the detection rates, we found that low-bitrate codecs such as SILK and Opus carry out an important function that improves detection rate: they remove background noise that negatively influences detection. In most cases, the steady state detection rates are between 60% and 88%, with a room occupancy (% of maximum seating capacity) of less than 50%, and a reliable network connection with 5–10% network jitter.

VoIPLoc does not depend on background sounds within a location or the voice of a specific speaker. Thus passive countermeasures such as filtering techniques will have little impact on attack efficiency since the fingerprint is computed over a basic VoIP-channel property — delivering the speaker's voice to the receiver with integrity.

Degrees of freedom: Scalability of the fingerprinting process is an important aspect to study. We evaluated VoIPLoc against a synthetic dataset comprising 404,058 rooms. Our experiments show that VoIPLoc can scale to a large number of locations with a couple of audio samples per room. Thus we can say that VoIPLoc presents at least five degrees of freedom to the attacker. collect at scale.

Tradeoffs between compression and privacy: The room structure and its interaction with compressive techniques employed by modern codecs plays an important role in the communication of characteristic information that can be leveraged for fingerprinting. At the same time, the benefits of compression such as reduced delay penalties are an important incentive for their use in VoIP design. As part of future work, we plan to study the tradeoff between compression and the ability to avoid location-detection, and whether there exist fundamentally stealthy codecs that can mask fingerprintable information but that are also serviceably compressive.

Location confirmation: VoIPLoc may be used in conjunction with macro-geolocation techniques such as PinDr0p [5], as a way to combine coarse-grained and fine-grained tracking.

Countermeasures: Defenders may consider a number of approaches. First, defenders may use acoustic jitter to damage fingerprint information. For instance, a constant amplitude signal a the room's characteristic frequencies between 50Hz and 2KHz (the discriminating subset of the location fingerprint) can cause a significant decrease in VoIPLoc's performance. This is essentially an acoustic jamming strategy which will deny access to the reverberant component of the channel to the attacker (receiver). On the other hand, any acoustic interference strategy will need to avoid jamming the communication channel itself or causing substantial disruption. However this is hard to achieve as even small amounts of audible noise will negatively impact voice quality (hence unlikely to be deployed). Alternately, network jitter can be used to induce packet latencies encouraging standard codec implementations to drop packets containing reverberant components. If accurately executed, this countermeasure could be fairly effective in preventing the sender from extracting a credible room fingerprint. As much as it would be effective against a standard implementation, the attacker could retain late packets (instead of dropping them) and access the reverberant component. Further, the reverberant component may be encoded into other packets as standard codec implementations often encode previous audio data into transmitted packets to mitigate packet losses.

Classifier design: VoIPLoc uses a deep NMF-SVM classifier. It combines techniques from DNNs (Deep Neural Networks) with a robust classical approach for finding decision boundaries. From DNN literature, we used multiple-layers, pooling, and normalisation, which are among the promising components of deep neural networks. Multiple-layers enable fine-grained partitioning between the first, second, and multi-order reflections via hierarchical decomposition that involves no training of weights (hence no backpropagation); Pooling reduces the impact of noisy audio traces; and, normalisation enables comparison across audio traces by normalising out the effect of varying amplitudes of direct sound. The VoIPLoc classifer architecture has multiple decomposition layers with the final layer composed of support vectors. The reason for using SVM as the final layer is that VoIPLoc requires a robust classifier that works with small datasets. Unlike SVMs, DNNs require large training datasets [14] which are not available in our problem setting —- one may need to distinguish just ten locations from a large number of unlabeled location fingerprints. To increase the quantity of training data, data augmentation techniques are commonly used for DNNs. However, each network of neurons will still require a significant amount (eg. 1000 samples per class for Imagenet classification [49]) of labeled data for it to train before data augmentation can assist. Applying DNNs is therefore challenging in the call provenance problem, where only a few tens of samples per location are available for training in the best case and a handful in the typical case. Aside from poor performance over small datasets, a second reason for choosing SVM is the appropriateness of the tool. DNNs are appropriate for high-dimensionality problems such as image classification [49] where the feature set for a 120×120 pixel RGB image is 43200, and feature selection is left to the classifier. This isn't the case with VoIPLoc, where the attack is based on a specific feature of the audio traces, namely the reverberant component for location inference. Thus acoustic location fingerprinting based on the reverberant component would not draw on one of the main strengths of DNNs – the ability to perform better in a high-dimensionality setting as compared with SVMs [37].

Bagging vs Boosting: An alternate approach to combine the output of different classifiers is *Boosting* (classifiers with votes weighted by accuracy) as opposed to the *Bagging* approach used by VoIPLoc (weak classifiers trained on a subset of the training data). Boosting requires a large training set which is typically unavailable in room fingerprinting whereas Bagging can be successfully trained with a small sample set. Also, Bagging performs relatively better with noisy data than Boosting approaches [25]. At the same time, Bagging approaches increase complexity. So are they worth the additional complexity? We tested VoIPLoc with and without Bagging. We observed that across the experiments, there was a 5% improvement in TP detection rate and a small decrease in the FPR. This indicates that while the noise tolerance of Bagging contributes in a positive manner, the improvement achieved is not very significant ($< 10\%$), and therefore we thing it is not worth the additional complexity.

Open-world vs closed-world: We evaluated VoIPLoc in both a closed-world environment as well as an open-world scenario (unlabeled fingerprints). As long as the proportion of unknown locations is less than 85% of the fingerprint database, the FPR is serviceable. Beyond that, unlabelled data penetrates through the contamination-resistance filters, decreasing detection efficiency.

Indoor vs. Outdoor application: Outdoor locations typically demonstrate poor reverberance characteristics, and require excitation signals of higher amplitude than human voices whilst in a VoIP conversation. For this reason, the applicability of this work is primarily of significance in fingerprinting indoor locations.

6 RELATED WORK

A number of works have attempted to derive location information via side-channels. The techniques can be broadly classified into passive approaches and active approaches. VoIPLoc is the first passive technique to achieve fingerprinting via echo-location characteristics of call origin.

6.1 Passive approaches

A number of passive approaches use static features of background noise. SurroundSense [4] combine sound amplitude with camera and accelerometer inputs to distinguish between indoor locations via overall ambiance (sound, light, and decor). ABS [53], identifies a location using low-frequency background sounds (computers, fans, buzz of electrical equipment). Next, we look at passive approaches that use dynamic features of background noise. Kraetzer et al. [26] propose a location identification technique that relies on repetitive patterns of music played in a location. Lu et al. (SoundSense) [31] generalises this to use background sounds such as passing trains and associates each location with a set of identifiable background sounds. Usher et al. [55] generalized this a bit further by replacing music with the voice of a single human speaker. Malik et al. carried out a small study over four very differently sized rooms and showed that differently sized rooms had different length and decay rate of reverberation [32]. Parhizkar and others extend this to echo-based approaches that leverage background sound as impulse signals [10, 12, 15, 24, 28, 43].

Vaidya et al. [56] describe re-identification attacks, wherein an attacker can infer location data from underlying audio, by analysing packet size distribution. While these approaches can distinguish a street from an airport it isn't serviceable for confirming the location of a VoIP user. The main challenge, as before, is that aggressive compressive encoding filters out all background signals.

Finally, within passive approaches, we review a number of complimentary techniques that leverage network characteristics instead of acoustic side-channels for localisation. Wang et al. [58] fingerprint locations based on channel-state information (CSI) from WiFi network interface cards. The use of CSI for localisation shows to be more successful than previous machine learning approaches which typically use stored received signal strengths (RSS) that have been described to have high variability for fixed locations and are very coarse [46, 59]. There are also works in the telephony-spam and mobile-fraud detection literature on mobile device provenance [24, 36] which is complementary to VoIPLoc's location provenance. PinDrop [5] leverages distinguishing acoustic characteristics arising from the device used, such as peak activity, choice of codecs, and double talk. It also uses characteristics induced by the network path on the codecs involved such as packet-loss rates in different networks. Similar ideas of leveraging packet-delay metadata has been proposed by Abdou et al. [3]. Unfortunately, Pindrop's fingerprint (and Abdou's work) being based on path characteristics, is vulnerable to the effects of network components such as Tor routers. VoIPLoc is less impacted since it doesn't fingerprint using network qcharacteristics as it directly focuses on location characteristics. While empirical testing with Tor has been carried out in this work, VoIPLoc's ability to remain undeterred where other solutions such as OnionPhone [17] (*successor* of TORFone [18]) and Phonion [21] may be employed has to be experimentally verified. Further, PinDrop's fingerprint is coarse grained by its very nature of being path and equipment centric characteristics. VoIPLoc's fingerprint is fine-grained being a function of the room's physical characteristics. This suggests that a unified fingerprinting approach that combines network-path, equipment, and room echolocation characteristics may be the best way forward. Malik et al. [33] was the first to apply machine learning to the problem (using SVM) but their fingerprinting technique suffered fundamentally. It measured the power spectrum (MFCC coefficients) without separating out reflections leading to two problems. First, contamination by background noise. Second, the fingerprint becomes a function of a specific spot within a room owing to contamination by direct-sound and early-reflections.

6.2 Active approaches

Recent works on privacy and acoustic channels include sensory malware, notably Soundcomber [50] and its variants which extract sensitive information such as credit card numbers from acoustic channels on a smartphone. There have also been reports of malware that communicate across air gaps via ultrasound squeaks [20]. A research paper based on this idea is SonarSnoop [9] which claims undetectability as no noise or vibrations are induced. However, health and well-being apps [1, 2] that record ambient 'energy' levels on a per-source basis will register these ultrasound beeps. A number of other works leverage the reverberation component for a function similar to fingerprinting. The state-of-the-art technique in this space for signal separation — isolating a high-resolution version of the reverberant component [13] – requires four microphones spaced exactly one meter apart. A first-order approximation using two microphones can be obtained using the technique of Pradhan et al. [47]. VoipLoc is the first to achieve credible separation of signal components to isolate the reverberant component using a single microphone without the need for a synthetic impluse signal. In the audible range, Shumailov et al. [51] propose an active acoustic side-channel attack using the device microphone to record sound waves which present over a touch-screen when it has been tapped by a user. A more sophisticated approach is to inject an optimal (short-high amplitude) impulse signal into the target location and measure the impulse response [42, 52, 54] apply forensic measurement techniques to develop high-fidelity acoustic model of a location to simulate the effects of a room over a given anechoic signal [16]. Murgai et al. [40] estimate the decay time and amplitude of the reverberant component to estimate room volume but not necessarily a specific room. In comparison, VoIPloc can differentiate between rooms of identical geometry by isolating the frequency components within a reverberant component of audio recorded with a single microphone.

7 CONCLUSION

Applications supporting and accepting voice based communications are very popular. Humans record and exchange audio data on a planetary scale. VoIP is a popular and important application used by dissidents, police, journalists, government, industry, academics, and members of the public. Thus privacy for VoIP applications is an important requirement. Beyond VoIP, voice-control is an increasingly popular method of user-device interaction in smart devices, which might reveal the fine-grained information about a user's location down to which part of the building they occupy.

Location information is embedded into human voice due to acoustic wave propagation behaviour, which forms the basis of location fingerprinting — the reflections of direct-sound interfere with each other and with direct-sound resulting in a rich interference pattern carried by encoded human voice. Given the wide usage of smartphones and VoIP tools, among the wider public to record and transmit audio, this work has important implications for anonymous VoIP communication, and more generally on user expectations of privacy within their homes as fine-grained user-location information can be derived from audio-interfaces to IoT devices. In terms of machine learning, while DNNs are popular, we learned that they are not the best option universally. In the call provenance challenge, given the sparse availability of traces to compute a fingerprint, less is more. Consequently, a hybid SVM-centric approach does better than a neural network approach.

ACKNOWLEDGMENTS

The authors are indebted to the numerous volunteers who helped with the field experiments from the University of Birmingham CS department and School of Informatics, University of Edinburgh. A particular note of thanks to Jon Weekes. The authors are thankful to the reviewers for their insightful comments. The first author is supported by a UKIERI grant (UKIERI2018-19-005). The second author is funded by EPSRC (EP/11288S170484-102) and NPL's Data Science program.

REFERENCES

[1] https://venturebeat.com/2011/06/19/narattes-zoosh-enables-nfc-with-just-a-speaker-and-microphone/. https://venturebeat.com/2011/06/19/narattes-zoosh-enables-nfc-with-just-a-speaker-and-microphone/. Accessed: 2020-03-13.

[2] Ultrasound detector android app. https://play.google.com/store/apps/details?id=com.microcadsystems.serge.ultrasounddetector&hl=en. Accessed: 2020-03-13.

[3] AbdelRahman Abdou. Internet location verification: Challenges and solutions, 2018.

[4] Martin Azizyan, Ionut Constandache, and Romit Roy Choudhury. Surroundsense Mobile phone localization via ambience fingerprinting. In *Proceedings of the 15th Annual International Conference on Mobile Computing and Networking*, MobiCom '09, pages 261–272, New York, NY, USA, 2009. ACM.

[5] Vijay A. Balasubramaniyan, Aamir Poonawalla, Mustaque Ahamad, Michael T. Hunter, and Patrick Traynor. Pindr0p: Using single-ended audio features to determine call provenance. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, CCS '10, pages 109–120, New York, NY, USA, 2010. ACM.

[6] Y-Lan Boureau, Jean Ponce, and Yann Lecun. A theoretical analysis of feature pooling in visual recognition. In *27TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, HAIFA, ISRAEL*, 2010.

[7] J. C. Brown. Calculation of a constant Q spectral transform. *Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

[8] Sasha Calhoun, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. The nxt-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Lang. Resour. Eval.*, 44(4):387–419, December 2010.

[9] Peng Cheng, Ibrahim Ethem Bagci, Utz Roedig, and Jeff Yan. Sonarsnoop: Active acoustic side-channel attacks. *International Journal of Information Security*, pages 1–16, 2018.

[10] S. Chu, S. Narayanan, and C.-C.J. Kuo. Environmental sound recognition with time;frequency audio features. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(6):1142–1158, Aug 2009.

[11] Stephane G. Conti, Philippe Roux, David Demer, and Julien de Rosny. Measurement of the scattering and absorption cross sections of the human body. *Applied Physics Letters*, 84(5):819, 2004.

[12] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, Aug 1980.

[13] Ivan Dokmanić, Reza Parhizkar, Andreas Walther, Yue M. Lu, and Martin Vetterli. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences*, 110(30):12186–12191, 2013.

[14] Simon S. Du, Yining Wang, Xiyu Zhai, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Aarti Singh. How many samples are needed to estimate a convolutional or recurrent neural network?, 2018.

[15] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi. Audio-based context recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1):321–329, 2006.

[16] A. Farina and R. Ayalon. Recording concert hall acoustics for posterity. In *24th AES Conference on Multichannel Audio, Banff, Canada*, pages 26–28, 2003.

[17] Van Gegel. Onionphone v0. 2a: Voip add-on for torchat.

[18] Van Gegel. Torfone: P2p secure and anonymous voip tool. http://torfone.org, 2012.

[19] Marie Guéguin, Régine Le Bouquin-Jeannès, Valérie Gautier-Turbin, Gérard Faucon, and Vincent Barriac. On the evaluation of the conversational speech quality in telecommunications. *EURASIP J. Adv. Signal Process*, 2008, January 2008.

[20] Michael Hanspach and Michael Goetz. On covert acoustical mesh networks in air. *JCM*, 8(11):758–767, 2013.

[21] Stephan Heuser, Bradley Reaves, Praveen Kumar Pendyala, Henry Carter, Alexandra Dmitrienko, William Enck, Negar Kiyavash, Ahmad-Reza Sadeghi, and Patrick Traynor. Phonion: Practical protection of metadata in telephony networks. *PoPETs*, 2017(1):170–187, 2017.

[22] Kentaro Ishizuka, Tomohiro Nakatani, Masakiyo Fujimoto, and Noboru Miyazaki. Noise robust voice activity detection based on periodic to aperiodic component ratio. *Speech Communication*, 52(1):41 – 60, 2010.

[23] Soeren Jensen, Koen Vos, and Karsten Soerensen. SILK Speech Codec. Technical Report draft-vos-silk-02.txt, IETF Secretariat, Fremont, CA, USA, September 2010.

[24] Constantine Kotropoulos and Stamatios Samaras. Mobile phone identification using recorded speech signals. In *2014 19th International Conference on Digital Signal Processing*, pages 586–591. IEEE, 2014.

[25] Sotiris Kotsiantis. Combining bagging, boosting, rotation forest and random subspace methods. *Artificial Intelligence Review*, 35(3):223–240, 2011.

[26] Christian Kraetzer, Andrea Oermann, Jana Dittmann, and Andreas Lang. Digital audio forensics: A first practical evaluation on microphone and environment classification. In *Proceedings of the 9th Workshop on Multimedia & Security*, MM&Sec '07, pages 63–74, New York, NY, USA, 2007. ACM.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.

[28] Nicholas D Lane, Petko Georgiev, and Lorena Qendro. Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 283–294, 2015.

[29] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562. MIT Press, 2000.

[30] E. Lehmann and A. Johansson. Prediction of energy decay in room impulse responses simulated with an image-source model. *Journal of Acoustical Soceity of America*, 124(1):269–277, June 2008.

[31] Hong Lu, Wei Pan, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell. Soundsense: Scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services*, MobiSys '09, pages 165–178, New York, NY, USA, 2009. ACM.

[32] H. Malik and H. Farid. Audio forensics from acoustic reverberation. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 1710–1713, March 2010.

[33] Hafiz Malik and Hong Zhao. Recording environment identification using acoustic reverberation. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1833–1836. IEEE, 2012.

[34] Gregory Maxwell, Timothy Terriberry, Jean-Marc Valin, and Christopher Montgomery. Constrained-Energy Lapped Transform (CELT) Codec. Technical Report draft-valin-celt-codec-02.txt, IETF Secretariat, Fremont, CA, USA, July 2010.

[35] Deep Medhi and Karthik Ramasamy. *Network routing: algorithms, protocols, and architectures.* Morgan Kaufmann, 2017.

[36] Jahanirad Mehdi. *Optimising acoustic features for source mobile device identification using spectral analysis techniques/Mehdi Jahanirad.* PhD thesis, University of Malaya, 2016.

[37] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *CVPR*, 518(7540):529–533, February 2015.

[38] F. Mordelet and J. P. Vert. A bagging svm to learn from positive and unlabeled examples. *Pattern Recogn. Lett.*, 37:201–209, February 2014.

[39] Samuel Muhizi, Gregory Shamshin, Ammar Muthanna, Ruslan Kirichek, Andrei Vladyko, and Andrey Koucheryavy. Analysis and performance evaluation of sdn queue model. In *International Conference on Wired/Wireless Internet Communication*, pages 26–37. Springer, 2017.

[40] Prateek Murgai, Mark Rau, and Jean-Marc Jot. Blind estimation of the reverberation fingerprint of unknown acoustic environments. In *Audio Engineering Society Convention 143*. Audio Engineering Society, 2017.

[41] Shishir Nagaraja and Ryan Shah. Clicktok: Click fraud detection using traffic analysis. In *Proceedings of the 12th Conference on Security and Privacy in Wireless and Mobile Networks*, WiSec '19, page 105–116, New York, NY, USA, 2019. Association for Computing Machinery.

[42] G.M. Naylor. Odeon—another hybrid room acoustical model. *Applied Acoustics*, 38(2–4):131 – 143, 1993.

[43] Reza Parhizkar, Ivan Dokmanić, and Martin Vetterli. Single-channel indoor microphone localization. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1434–1438. IEEE, 2014.

[44] The Guardian Paul Lewis. Snowden documents show nsa gathering 5bn cell phone records daily. https://www.theguardian.com/world/2013/dec/04/nsa-storing-cell-phone-records-daily-snowden, 2013.

[45] C. Plapous, C. Marro, and P. Scalart. Improved signal-to-noise ratio estimation for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(6):2098–2108, Nov 2006.

[46] Nico Podevijn, David Plets, Jens Trogh, Abdulkadir Karaagac, Jetmir Haxhibcqiri, Jeroen Hoebeke, Luc Martens, Pieter Suanet, and Wout Joseph. Performance comparison of rss algorithms for indoor localization in large open environments. In *2018 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–6. IEEE, 2018.

[47] Swadhin Pradhan, Ghufran Baig, Wenguang Mao, Lili Qiu, Guohai Chen, and Bo Yang. Smartphone-based acoustic indoor space mapping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–26, 2018.

[48] J. Ramirez, J. C. Segura, J. M. Gorriz, and L. Garcia. Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition. *Trans. Audio, Speech and Lang. Proc.*, 15(8):2177–2189, November 2007.

[49] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 525–542, Cham, 2016. Springer International Publishing.

[50] Roman Schlegel, Kehuan Zhang, Xiao Yong Zhou, Mehool Intwala, Apu Kapadia, and XiaoFeng Wang. Soundcomber: A stealthy and context-aware sound trojan for smartphones. In *NDSS*. The Internet Society, 2011.

[51] Ilia Shumailov, Laurent Simon, Jeff Yan, and Ross Anderson. Hearing your touch: A new acoustic side channel on smartphones. *arXiv preprint arXiv:1903.11137*, 2019.

[52] Stephen P. Tarzia, Robert P. Dick, Peter A. Dinda, and Gokhan Memik. Sonar-based measurement of user presence and attention. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, UbiComp '09, pages 89–92,

New York, NY, USA, 2009. ACM.

[53] Stephen P. Tarzia, Peter A. Dinda, Robert P. Dick, and Gokhan Memik. Indoor localization without infrastructure using the acoustic background spectrum. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, MobiSys '11, pages 155–168, New York, NY, USA, 2011. ACM.

[54] Nicolas Tsingos. Pre-computing geometry-based reverberation effects for games, 2004.

[55] J. Usher and J. Benesty. Enhancement of spatial sound quality: A new reverberation-extraction audio upmixer. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(7):2141–2150, Sept 2007.

[56] Tavish Vaidya, Tim Walsh, and Micah Sherr. Whisper: a unilateral defense against voip traffic re-identification attacks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 286–296, 2019.

[57] Michael Vorla:ander. *Auralisation: Fundamentals of Acoustics, Modelling, Simulation, and Acoustic Virtual Reality.* Springer, 2007.

[58] Xuyu Wang, Lingjun Gao, Shiwen Mao, and Santosh Pandey. Csi-based fingerprinting for indoor localization: A deep learning approach. *IEEE Transactions on Vehicular Technology*, 66(1):763–776, 2016.

[59] Moustafa Youssef and Ashok Agrawala. The horus wlan location determination system. In *Proceedings of the 3rd international conference on Mobile systems, applications, and services*, pages 205–218, 2005.