VoipLoc: Compromising location-privacy via acoustic side-channel attacks

Anonymous

***Abstract*—We develop a novel location-privacy attack based on acoustic side-channels. The attack is based on acoustic information embedded within foreground-audio. We study how audio data, generated by secure messaging clients in voice-call mode, can be (ab)used to generate a location fingerprint. The attack leverages the pattern of acoustic reflections of human voice at the user's location. It does not depend on any characteristic background sounds. The attack can be used to compromise location privacy of participants of an anonymous VoIP session. Or, to carry out confirmation attacks that verify if a pair of audio recordings originated from same location regardless of the speakers. We show that location fingerprinting is exacerbated by codecs, and is robust to network jitter. Evaluation in real-world settings confirms that recording locations can be fingerprinted and detected remotely over VoIP channels with low false-positive rate.**

I. Introduction

Secure messaging clients are popular means for resisting widespread surveillance. While such tools have been examined for security properties such as end-to-end encryption and inspected for their integrity via source-code audits, accidental ways that expose user information are a constant worry. Secure messaging clients providing audio calling technologies are used everyday by people to communicate with friends, family members, colleagues, and strangers. In this work, we examine audio calling mechanisms from *first principles* and uncover a significant privacy weakness.

Messaging tools with audio calling technology inadvertently unicast unique location fingerprints in the acoustic channel, severely compromising the location-privacy of their users. During an audio call, the speaker's voice is modulated by the sound reflection characteristics of the location, and the modulated signal is captured as an electronic signal, compressed, and transmitted to the receiver. At the receiver, the signal is decoded and regenerated whereupon the receiver can examine the signal with respect to various attributes apart from the speakers's direct voice to ascertain location-induced uniqueness: the timing and shape of reflections, the energy contained in various bands, and their decay rate.

Our main finding is that, it is possible to recover this information in practical conditions. Endusers unknowingly transmit location information in addition to the intended (voice) content is a privacy compromise. Furthermore, obvious countermeasures such as filtering background noise or signal-dampening have little impact on attack efficiency.

While constructing active attacks for location fingerprinting is widely known to be possible – injecting ultrasound squeaks such as those used by bats – no passive attacks have been proposed to the best of our knowledge. In this context, we ask the following research question: can human speech recorded by off-the-shelf devices be used to acquire a *location fingerprint* in a practical setting? How unique is the fingerprint? In order to be practical: fingerprints should be *unique* – distinctiveness from fingerprints of other rooms or locations; fingerprint acquisition should be *stealthy* i.e unobservable to any entity within the room being fingerprinted; be *time invariant* – not rely on the identification of background sounds for fingerprinting, such as proximity to an airport; be robust to environmental noise – crowds or people, whirring computers, blowing fans, passing trains, or noisy coffee shops; and, not require specialised hardware.

The fingerprinting attacks proposed in this paper meets these requirements. The intuition behind our techniques is that room fingerprinting can be carried out passively by analysing recorded human speech, which encodes signal transformation due to the "acoustic behaviour" of a room i.e the transformation a room applies to any sound waves via a sequence of reflections, absorbtions, and scattering processes due to room geometry, construction material used, and the physical objects within the room [22]. An attacker who can access output audio but has no physical access to the room can analyse recordings for similarities induced by acoustic behaviour. Thus, given an audio recording what can possibly be understood regarding where the sound was recorded? Can multiple audio recordings of human conversations be linked to the same location of recording? Is it possible to develop a binary test which confirms or denies whether a conversation was recorded in a particular room in a building?

Linking audio streams based on the recording location can have significant privacy implications. An attacker can compare audio data to confirm or deny whether they were also recorded at the same location as a known sample. The attacker can compile a database of tuples each containing a sample recorded at the location, which can be used to to detect "matches". We report three types of privacy attacks: In a *location traceability* attack, such a matching technique can be used to link multiple sessions of a VoIP user, passively using the sound content transmitted. If a database of audio samples with labelled locations is known, then a *location confirmation* attack to reveal the geo-location of the victim can also be carried out. It is worth noting that the labelled audio samples don't have to be from the victim.

Previous approaches to acoustic location-fingerprinting consider background noise sources. A critical limitation is that background sources can usually be filtered out via high-pass filters, these are often used by compressive codecs to favour voice transmission. In contrast, VoipLoc depends only on the

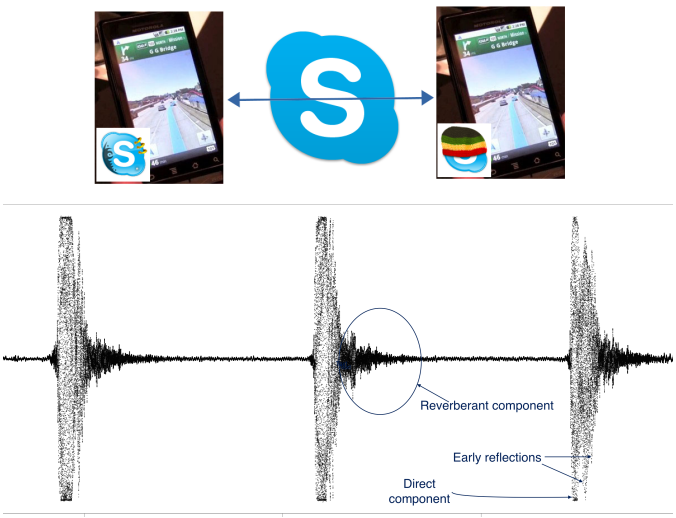


Fig. 1. Sound recording: x-axis is time and y-axis is pressure (or voltage)

speaker's voice, which cannot be filtered out during an audio call. This is our core contribution.

A detailed review of prior art can be found in Section VII. SurroundSense [1] – combines multiple sensors including measurement of acoustic background sounds (and combines it with visual background and cellular signals). In contrast, VoipLoc depends exclusively on the microphone and offers better localisation even in the presence of compressive codecs which filter out ambient sounds. Acoustic Background Spectrum (ABS) [31], is another technique that also focuses on background sounds in the low-frequency bands, particularly the buzz produced by electrical devices. The signal characteristics on which ABS depends are easily filtered out since its overlap with human speech is very little. In fact, most speech codecs operate above 300Hz which emulates the application of a high-pass filter rendering ABS ineffective. SoundSense [19] uses ambient sounds for activity detection such as steps on a staircase.

The layout of this paper is as follows. First, we will cover the necessary background on sound acoustics. Second, we discuss the privacy implications of room fingerprinting, describe the threat model, and justify its use. Third, we study passive and active attacks. Fourth, we study the impact of audio compression on attack efficiency via popular codecs whilst adjusting for varying network conditions that impact their functionality. Fifth, we control for different speaker frequency ranges and decibel level.

A. Threat model

The threat model is that the adversary has access to the audio stream such as recorded speech after it reaches a communication endpoint. For example, an intelligence agency analysing streaming audio from a dissident website or tracing the location of a dissident activist giving a phone interview. From a communications perspective, the adversary is an insider, who is engaged in a VoIP conversation with the victim and would like to determine the location of the latter

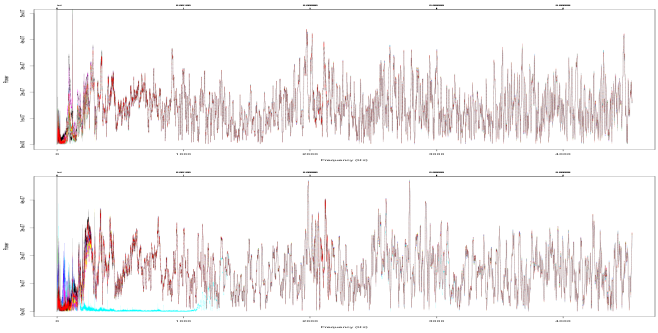


Fig. 2. Voice-energy distribution of two identically sized rooms with minor customisation differences. Note that the significant difference in the locations of peaks and troughs.

party. 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 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 salesman for a firm dealing in radioactive materials or surveillance equipment might be contacted by attackers posing as prospective clients, who can gain information about the salesman's clients after compromising his/her location privacy. Snowden's revelations famously revealed the tracking of sales personnel by tracking the victim's mobile phone. As opposed to the macro-level location information provided by cell-tower localisation, we report attacks that can be used to carry out fine-grained indoor-location identification, down to a specific room or corridor the victim used to make a voice call. As another example, an attacker attending an online meeting set up by an NGO activist could compromise operational security of the NGO's protest.

We have specifically chosen to study passive adversaries in order to demonstrate the power of the attacks. Even a passive attacker can confirm the source-location of an audio call. An active attack would likely be more successful by the injection of optimal sound waves (such as sharp impluse signals) these are considerable less stealthy. Optimal audio signal waveforms are short bursts of high-amplitude signals that are rarely found in human speech generation. On the other hand, the passive attacker model studies the lower bound of attacker success given the worst possible circumstances for location identification. It's noteworthy that while active methods have received some attention in the literature, there's little prior art on passive attack methods.

II. Background

A room's characteristic property is it's wave reflection behaviour. Sound is a pressure wave, and the room acts as an energy propagation system. Upon sound production, the pressure wave expands radially outward, reaching obstructions where the energy is absorbed or reflected. The pattern of reflections, each represented by the time of arrival and strength

in various frequency bands, can be used to fingerprint the room.

The sketch below depicts the sound received by a single listener as a function of time. The direct sound received is followed by distinct reflected sounds and then a collection of many reflected sounds which blend and overlap into what is called reverberation. The direct sound component is a baseline version of the original transmitted signal while the reflected waves carry information about the recording environment which aid in room fingerprinting.

Thus, recorded sound consists of three components. First, the *direct* sound as transmitted from the source to the recorder is recorded. Second, early reflections are recorded. Third, the reflected waves interfere with incident waves producing patterns of interference known as the *reverberant* component.

A. Model of room reflection behaviour

The recorded sound at any location within a room is composed of the three components described earlier: direct sound, early reflection, and the reverberation.

Direct sound is the speech signal on the direct path from the speakers mouth to the microphone. This corresponds to the signal from the sound source alone and no reflections. Early reflections (of the direct sound) are defined as those that arrive via a predictable directional path. They contain evidence of the geometrical surroundings of the receiver. We define the amount of time required for the early-reflections to diffuse into the room as the *mixing time* of a room.

After the early reflections, we have the reverberation signal. Reverberation is composed of higher order reflections. They can be modeled as an exponentially decaying, ergodic, stochastic process, with a Gaussian distribution and a mean of zero. The volume of a room dictates the frequencies (lower bound) above which a stochastic model of reverberation is valid. The Schroeder frequency, is approximated as

$$f_{Schroeder} \approxeq 2000\sqrt{\frac{RT_{60}}{V}}(Hz)$$

, where V is the room volume in cubic meters. The stochastic model for the reverberant component is therefore only valid after the mixing time and for frequencies above the Schroeder frequency [15]. RT_{60} is the length of the reverberant part of the signal – i.e the time taken for direct sound to diffuse into the room until signal power is reduced to -60dB (the threshold of silence). Larger rooms typically have longer reverberation times. This can lead to overlapping of reverberation signal from the previously spoken word with the direct sound of the consecutive word.

Above the Schroeder frequency the sound correlation between two locations in a room is dependant on the wavenumber and the distance between them . Therefore, the assumption that the reverberant-path components are uncorrelated, i.e., for a typical recording with a 20-cm spaced microphone pair is only valid for frequencies greater than approximately 1 kHz.

III. Primitives

As explained in the previous Section II-A, recorded sound is composed of direct, early reflections, and late reverberation components. These signal components are useful to analyse depending on the task at hand.

The first component, direct sound, is wholly a function of the speaker (or sound source). It's useful to isolate direct sound or to make the signal anechoic by removing room-dependent features, to anonymise the signal or increase acoustic analysis resistance. The third component, late reverberation, is useful to isolate when the analysis is based on a *summary* of room reflection behaviour. In contrast, early reflections, provide a fine grained access to reflection behaviour. An early reflection is a distorted version of the direct sound where the distortion is defined by the reflecting surface within a room. Therefore it's useful to isolate early reflections when the analysis is targeted at understanding reflection timing or frequency-amplitude distributions to characterise the precise location of the recording microphone within a given room.

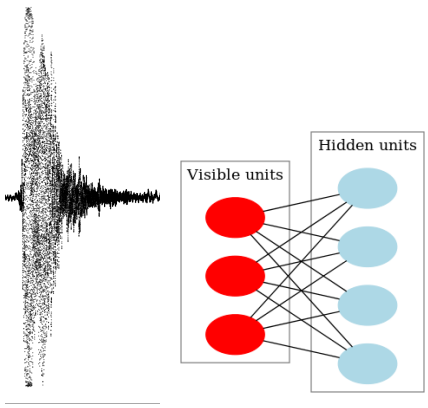
A. Audio segmentation

Speech segmentation, obtains discrete chunks of speech for processing. There are many voice activity detectors, silence detectors, and turn-taking options in the literature [13], [25]. We used a combination of volume, spectral energy, fundamental frequency (F_0), and spectral flatness for creating a predictor for speech segments. The spectral flatness can be used for characterizing an audio spectrum for how tone-like a sound is, and hence can eliminate signals with a large mixture of sources (such as multiple people talking at once, or music or TV in background). Speech segmentation preceeds the application of the rest of the primitives.

B. Isolating Wave Interference Pattern (reverberation component)

The key challenge in isolating the reverberant component is to isolate the complex overlapping sound patterns arising from multiple voice components from one or more speakers. In order to achieve speaker independent fingerprinting, it is critical that any direct sound components are removed from the reverberant component. For any speaker, a recorded speech segment contains several direct-sound components, each corresponding to a syllable uttered. As the speaker utters a syllable, the early and late reverberation components corresponding to previously uttered syllables, temporally overlap the direct sound of the current utterance. Thus the challenge of isolating direct sound component is to remove the temporally overlapping reverberation components from previous utterances. We will assume that the technique (Section III-B) for isolating late reverberation has already been applied on a given speech segment, so we now focus on removing the early reflection component.

We observe that the shape and form of early reflections are, by definition, very similar to direct sound. As such early reflections can be modelled as direct sound convoluted with the transformation applied by the reflecting surface. We leverge this observation to to reverse the wave interference process to arrive at a reverberant component that's free of direct sound and its reflections, hence achieving speaker independence. Recorded sound is represented as an algebriac combination



of repeating sound patterns where repetition is due one or more direct sounds reflecting off surfaces in the recording location. Thus direct sounds can be isolated via search for self-similiar statistical patterns, and removed from the reverberant component.

Isolating the primary signal so that room information doesn't get encoded into the recorded sound wave is a *deconvolution* problem. We applied non-negative matrix factorisation to deconvolve the sound waveform into the primary sound wave and its secondary and tertiary reflections. (The technique is weakly related to PCA [28] but doesn't enforce orthogonality between the basis vectors).

We split the input speech matrix O into a matrix of **hidden patterns** H (direct sound and transformation noise) and a **weight matrix** W. Recorded sound can be recomputed by linearly combining the pattern matrix H with weights W to obtain reflections. The given speech segment is expressed in terms of the hidden sound patterns and mixing weights as:

$$O = WH \tag{1}$$

Essentially, the observed data during each time interval is a function of the patterns (columns of H) and the extent of activation of each pattern given by rows of the weight matrix W: $O_{ij} = \sum_{k=1}^{K} W_{ik} H_{kj}$. The key to implementing a notion of compression is to represent as much of the speech segment using the sparsest and fewest number of sound patterns. This is achieved by standard application of the NMF algorithm. We start with randomly initialised positive-valued matrices W and H, and update them iteratively using a gradient descent algorithm until convergence is achieved. We chose to apply multiplicative update rules from optimisation literature [18]. Once convergence is achieved, the columns of H contain direct-sound signals.

$$\begin{aligned} W_{ik} &= \frac{W_{ik} \sum_k \frac{O_{ij}}{(WH)_{ij}} H_{kj}}{\sum_l W_{jk}} \\ H_{kj} &= H_{kj} \sum_i W_{ik} \frac{O_{ij}}{(WH)_{ij}} \end{aligned}$$

IV. Fingerprint computation

The fingerprint is computed over the reverberant component of the audio signal. We start with the following hypothesis. For two or more sound recordings: the reverberant components are at least partially correlated for recordings from the same room, but uncorrelated to reverberant component of a recording from a different room.

Formally, the reverberant components are uncorrelated $E\{r_i(t), r_j(t)\} = 0$ for rooms i and j. However, they are correlated for any locations within the same room i.e $E\{r_i^p(t), r_i^q(t)\} \neq 0$ for locations p and q within room i.

Room fingerprint involves several steps which we expand on in the following sections. In summary these are as follows: preprocessing to remove noise, detecting the onset of the reverberant component, aggregation of signal power in relevant bands to produce the fingerprint vector, and finally training and classification to examine detection accuracy.

A. Preprocessing

Noise plays a large part in the success of fingerprint detection and also for reverberation estimation. After capturing the audio, we perform normalisation to remove the DC offset and to keep the maximum amplitude capped at 1.0dB. All environments have some level of baseline noise typically from the HVAC. Since most signals will contain non-speech (over 95% in our experience), we build a noise model of the uniform noise in the room and use spectral subtraction on the speech segments. We used a Wiener noise suppressor [24]. Their approach uses harmonic regeneration noise reduction (HRNR) to refine the SNR a priori to compute a spectral gain to preserve speech harmonics. More sophisticated machine learning-based noise subtractors can extend this approach.

B. 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. In simple terms, we measure relative signal attenuation as a function of room geometry.

To design a fingerprint function, one design consideration is the choice of frequencies to consider. It is worth noting that the correlation of sound pressure for different locations within a given room is inversely related to the frequency. Specifically, the correlation in sound pressure is given by the following, for frequencies above the Schroeder frequency [25]. For two locations, i_l and i_m within the same room:

$E[i_l, i_m] = \frac{sin(2\pi f/c)x}{(2\pi f/c)x}$

As a result of the above relationship, high frequencies are only correlated at close proximity, while low frequencies (above Schroeder frequency) can be correlated at relatively longer distances. Thus, signal power at low frequencies is ideal for room fingerprinting while high frequencies within the reverberant component may be used to specify the microphone location within the room itself. We carry out frequency analysis to compute power in various bands and retain the spectral information related to low frequency bands (20 Hz to 500hz).

The second design aspect relates to the frequency analysis approach over the chosen frequency bands. 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 Contralto 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 a 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 freqency band, we use the Constant-Q transform [3]. 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 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 technique described in Section III-B. The output of the transform is a vector which contains the signal power in each frequency band. We call this the *fingerprint vector*.

C. Normalisation and aggregation

The role of the normalisation step is to enable comparison of reverberation signal for different excitation signals i.e voice from different speakers. Without this step, biases in the direct-sound induced by individual speakers would cause inaccuracies in comparing the spectrum of the reverberant signals. Normalisation emulates the application of a constant amplitude sweep across the signal frequency range.

The process of normalisation involves computing the signal strength in each frequency band of the direct-sound signal. Note that this is the input excitation signal and the later-reverberant signal is the corresponding *output* produced by the room. The fingerprint vector computed thus far is the output-signal strength for each frequency, this is normalised by the input signal strength; the output is undefined for a frequency where the input signal strength was zero. Normalisation is carried out by 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 fingerprint vectors from each speech segment. The necessity for this is that unlike a sine-sweep function [26], each speaker's voice only covers a subset of possible frequencies. Hence we need to merge multiple CQT vectors computed over the late-reverberation component, to maximise the range of frequencies for which a response is analysed. The design of the merge function is straightforward. Vector f contains the normalised aggregated fingerprint. For each segment i and frequency j,

$$D_{ij} > 0 : F_{ij} = R_{ij} n D_{ij} \tag{2}$$

$$D_i \leq 0 : F_{ij} = 0 \tag{3}$$

$$f_j = \sum_i^n F_{ij} \tag{4}$$

D. Fingerprint classification

We then input the fingerprint vector into an SVM classifier, which partitions the input into classes, where each class corresponds to a unique recording location.

We use a support vector machine (SVM) for classification because of the large feature size. The LIBSVM library is used for both training and testing. An SVM identifies the optimal separating hyperplane that maximises the *margin of separation* between linearly separable fingerprint vectors. The data points which lie closest to the separting hyperplane are called support vectors. For generalisation, we use a non-linear kernel function; linearly non-separable data points in input space can be mapped into a higher dimensional (possibly infinite dimensional) feature space through a nonlinear mapping function, so that the input fingerprints become almost linearly separable. The kernel function used in our work 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.

For location classification we used a standard methodology from ML literature called the one-versus-all technique. To link audio signals amongst N rooms, we use N SVMs. The ith SVM is trained using half the samples in the ith room with a positive room label and half of all the remaining samples with a negative room label. The other half of the samples are used for evaluation. This is nested within a k-fold over cross-validation process to avoid problems such as overfitting the data.

V. Evaluation – Coarse-grained fingerprinting (feasibility analysis)

To evaluate the effectiveness of our techniques we first experimented in a variety of real-world settings over a very simple range of parameters to establish the feasibility of voice based location-fingerprinting. Hence there are no human participants in this section of the evaluation. Human participants are involved in fine-grained analysis (next section). Further, we selected locations with markedly different acoustic characteristics owing to significant differences in physical size, shape, volume, and construction material. The rationale for these evaluation choices was simply to establish feasibility of the attack in the first instance. In the next section, we will present a fine-grained evalution of the technique by comparing attack efficiency across rooms with relatively similar shape,

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 centre of the university. The reverberation length is fairly long. Volume is $12000m^3$.
Typing Room	Partitioned typing room in the British Museum.
Office rooms	38 rooms from a university office building
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 panelled 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 I
INDOOR LOCATIONS CONSIDERED

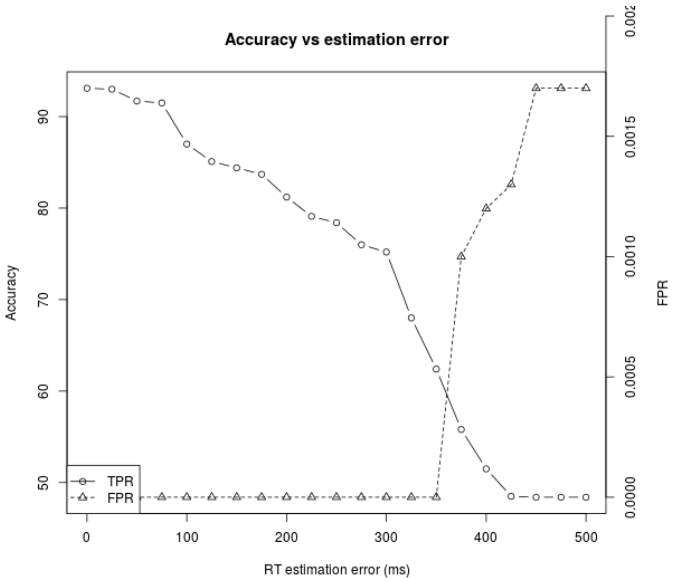


Fig. 3. Location accuracy vs RT estimation error

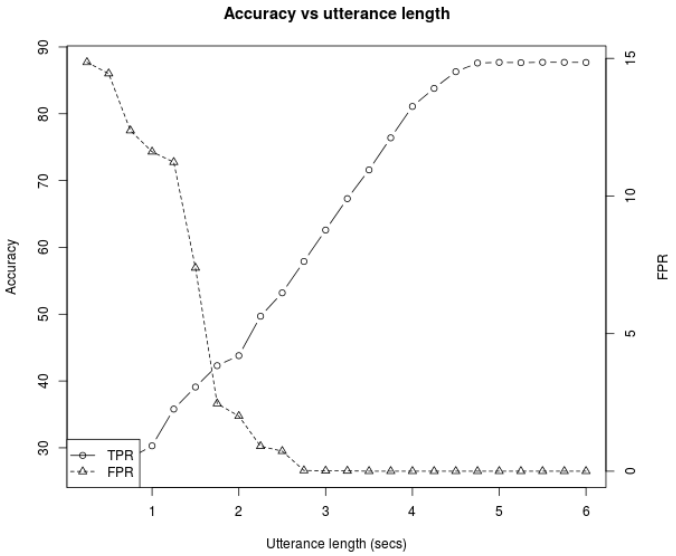


Fig. 4. Location accuracy vs utterance length

size, and volume, and a comparitive analysis of codecs via human participants engaged in VoIP conversations.

Speech production is carried out by playing anechoic pre-recorded audio clips of human speech using a bias-free (power output per input frequency) loudspeaker system. We use the EmoDb speech dataset. **Voice diversity:** The dataset contains spoken utterances by 10 different actors (5 male, 5 female) using a variety of emotions. The utterances are various short phrases recorded within a non-reverberant (non echoic) chamber. Each voice sample is played at 70dB (average sound power corresponding to normal (non-agitated) conversation **Intra-location diversity:** For each location, we divided the room into a 3x3 grid and the speaker was placed at a randomly chosen point on the grid. The idea is to rule out bias aring from placement of speaker within a fingerprinted location. The microphone was placed at a distance of 5cm in front of the speaker. **Audio sample generation:** A VOIP session is set up using Skype messenger between the sender at the given location and a recipient over mobile broadband link. Audio is transmitted in a single direction from the speaker location to the receiver. The raw audio signal is encoded by the speech codec on the sender side, using a cardoid microphone placed two inches in front of the speaker. Next, the encoded signal transmitted over the network via to the receiver is decoded and the resulting audio stream is recorded.

In this section, our goal is to understand whether sound samples from the same room can be linked to each other from a large population of recordings that come from various indoor venues. Our dataset consists of sound samples collected in locations given in Table I.

To resist over-training our classifiers, we carried out a ten-fold cross validation over the audio samples. The dataset is partitioned randomly into ten subsets with nine being used for training and the tenth being used for validation, repeated ten times with each possible sub-sample being used for validation. As we require multiple classes (each class corresponds to a location being fingerprinted), we follow a one-vs-all approach wherein a classifier is trained for each location using positive and negative labels from the nine subsamples per location. Note that all training and evaluation samples are post-codec audio signals, which contains noise due to codec operation as well as network jitter. The results are shown in Fig. 4 and Fig. 3, and we discuss the results.

Parameter estimation The classifier depends on accurate estimation of reverberation time to isolate the reverberant com-

ponent. 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 microseconds (ms) from the ground truth. The results are shown in Figure 3. 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 classifer, 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 300ms but suddenly increases after RT error reaches 350ms. 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 categorised together leading to a significant rise in the false-positive rate.

Utterance length . A second parameter is the minimum signal power conveyed by the utterance length. If the utterance length is very small, there simly isn't enough signal strength to carry out the necessary discriminative tests. 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 varing lengths and observed classifier accuracy. We used the EmoDb speech sample dataset again. In figure 4, 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%.

VI. Evaluation – Homogenous locations (Fine grained analysis)

The experiments from the previous section demonstrate the feasibility of room fingerprinting and identification in a real world environment consisting of a wide variety of locations. In this section, we demonstrate how well it performs using VoIP conversations in locations drawn from a relatively narrow distribution of acoustic characteristics. Simply stated, we consider geometrically identical rooms (with similar construction material) such as those found within a modern build. This scenario allows us to study the quality of location identification in a relatively controlled setting where we can carefully understand the parameter space outside of coarse-grained features such as room geometry and construction material and focus on fine-grained features such as network jitter, speaker orientation and position within the room, and the speaker's voice.

The 38 rooms in the study belong to the computer science department of a university. These rooms are customised by their occupants using room furnishings such as desks, bookshelves, monitors, and other objects that affect room acoustics. The rooms also have typical noise sources which could be continuous (such as air conditioning systems, heater fans, and fridges) or intermittent (road traffic). They also have open or closed doors. These customisations form the basis for location identification. Our goal is to understand the extent to which our techniques can exploit these differences in acoustic absorbtion characteristics and tolerate noise sources to aid location identification.

For this experiment, we requested sixteen volunteers to conduct VoIP sessions in each of the rooms and recorded them in two ways. First, the raw signal captured by the microphone. And second, the signal reproduced at the receiver after passing through the audio codec. For each room, we seated the volunteers at nine different positions located at the intersections of a 3x3 grid (rectangular). For each position, the volunteer was placed facing four different orientations (facing each of the room's walls in turn). Volunteers were instructed to remain in a neutral tone and hold a conversation by reading out from a script. Each volunteer sitting in one of the experiment rooms interacts with an experimenter via a LAN link to a room on the same campus. The scripts are from the NXT Switchboard Corpus [4] consisting of telephone conversations between speakers of American English, is one of the longest-standing corpora of fully spontaneous speech. We used to 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, while allows us to study classifier accuracy for this parameter.

A. Codecs

Up until now, we have analysed raw speech signals. We now consider the impact of speech codecs used in VoIP systems. 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 to remote fingerprinting due to the potential loss of relevant signal information that can be crucial to the effectiveness of the fingerprinting process.

Modern codecs involve a number of functions. *Voice activity detection* – generates a measure of speech activity by measuring the signal-to-noise ratio, often across multiple frequency bands (ranges). A variable cutoff *high-pass filter* removes ambient sounds – low frequency background sounds and breathing noise. The variable cutoff ensures that for high-pitched voices the cutoff is correspondingly higher, allowing all but the lowest harmonics to be filtered out. *Error coding* is

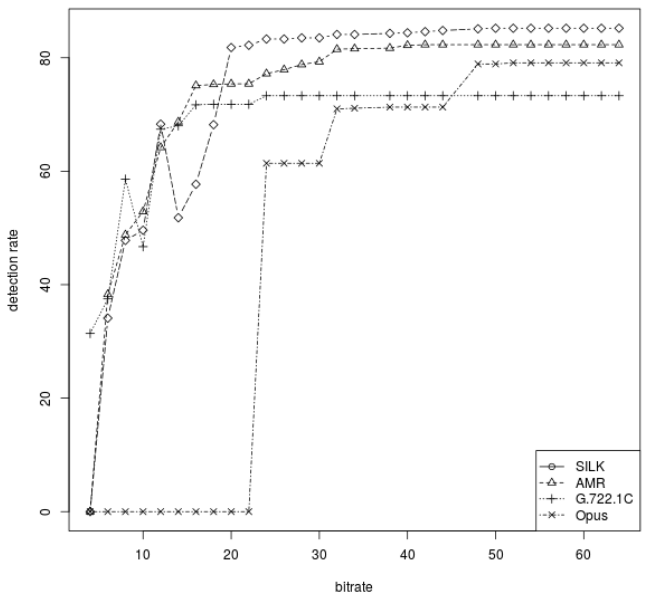


Fig. 5. Location accuracy vs bitrate

deployed by some codecs to absorb the random loss of packets. This is usually done by forward packet coding techniques, where in information about the current packet in the queue is probablistically injected into a fraction of subsequent packets. Finally, *Noise shaping* is used to exploit the properties of the human auditory system. In modern codecs, the encoder determines the speech signal by minimising the perceptual error rather than actual reconstruction error. The decoder then uses a postfilter on the reconstructed signal to supress spectral regions where the reconstruction error (quantisation noise) is expected to be higher relative to the speech signal. Most codecs will use some form of a perceptual model [16], [29] to make decisions on shaping the signal reconstruction error and appropriately boost or supress spectral regions within the signal. The use of specific models characterises a codec, since its operation reduces bandwidth consumption whilst decreasing noise and making the signal more predictable.

SILK codec(Skype) The widely used Skype VoIP services uses SILK codec [14] as well as other proprietary voice codecs for encoding high frequencies in the range of 16Khz. [this should come from Opus comparison paper from Nokia]. 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 8 kHz 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 5 shows the impact of bitrate on fingerprinting efficiency (detection rate). We observe a rather low efficiency 40%, at very low bitrates in the range of 6–10kbps. Around 14–16kbps we observed an increase in attack efficiency to 71%. 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, stabilising 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 to the receiver which significantly improves fingerprinting.

AMR-WB (Blackberry AIM) AMR is an audio data compression scheme optimised for speech coding in telecommunications systems such as GSM and UMTS. It employs Linear Predictive compressive coding, just like SILK does in Narrowband mode. However, unlike SILK AMR employs LP throughout, to support signal frequencies from 50Hz to 7000Hz. AMR supports bitrates from 4.75 kbps onwards upto 23.05 kbps. AMR is somewhat dated as a codec in terms of design principles and applications, we use it as a baseline to compare against relatively modern codecs.

722.1c 722.1c is a low-delay generic audio codec [35] standardised by the ITU. It is deployed widely in hardware-based VoIP devices particularly the Polycom series. It offers supports sampling rates above 32kHz offering a wider band than AMR, whilst supporting bitrates of 24, 32, and 48kbps. Results in figure 5 show that at least 32kbps is required to achieve reasonable fingerprinting success.

a) Opus (Facebook Messenger): The Opus codec is a framework for composing previously known high quality codecs, SILK [14] and CELT [21]. 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; it's best suited for audio signals. We observe a detection rate of less than 50% at less than 10kbps in LP mode. At 12kbps, we observe a significant improvement of 19% in fingerprinting efficiency to 70% (which is in the realm of usefulness). This is the threshold when Opus switches to Hybrid (wideband) mode, 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. This is the threshold at which Opus internally shifts from a lossy compression to lossless compression. As the bitrate increases the signal bandwidth increases leading to a greater fidelity on the decoder. At 14kbps, the Opus codec shifts from LP to hybrid mode, however at this bitrate, room fingerprinting in hybrid mode performs much worse than in 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.

All tests were performed with a frame length of 20ms at constant bit rate. Opus supports short (2.5ms) and long (60ms) frame lengths. Shorter the frame, higher the bitrate. Opus also 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 packetisation. 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. 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, framelength, jitter, look ahead, and training and testing on different conditions influenced by these parameters.

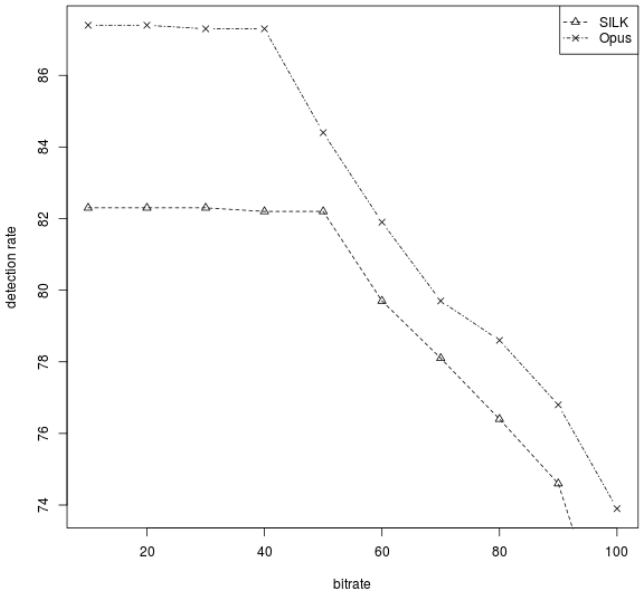


Fig. 6. Location accuracy vs room occupancy

B. Time invariance – Reflection characteristics

Any location fingerprinting method should be robust to random (non-adversarial) noise i.e temporary changes in the environment. We study the effects of two types of noise. The number of humans in a location can be time variant. For instance, the occupancy rates of meeting rooms, lecture rooms, and coffee shops can vary, changing both ambient noise as well as the physical characteristics of the room. We evaluate the impact of room occupancy on refl first.

First, we study the impact of *room occupancy* on reflection characteristics, i.e changes in the reflection/absorbtion characteristics of a room depending upon the number of people in it. Second, we will study the impact of *ambient noise* on attack efficiency, in the following subsection.

Conti et al. [7] showed that human bodies are good multi-directional reflectors assisting in the mixing of sound within a room. They demonstrated that the extend of reflection is soley described by the mass of the human which acts as a rigid water-filled ellipsoid. And, absorbtion by the human body is largly dominated by the amount and type of clothing — naked bodies will reflect entirely while think winter coats will increase signal absorbtion. Since the fingerprinting attack is primarily sensitive to the absorbtion 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 an absorbtive surface such as a couch is blocked by room occupants, it could change the room's fingerprint.

We examined the impact of room occupancy on reflection characteristics, we ran experiments in six different lecture rooms with various rates of occupancy. Training was carried out using audio traces recorded in a near-empty occupation state when only a few students were present. Testing was carried out at various rates of occupancy. For each room, we recorded four to eight different speakers lecturing or giving a talk over a period of 30 to 50 minutes. For each speaker, we collected sound traces specifically for training by having them read a standard script under the pretext of testing the AV system. Each session was recorded after taking consent of the speakers.

Figure 6 shows that the detection rate is decreases gradually after around 60% occupancy. We note a decrease of around 12% in the detection rate in case of SILK, from a high of 83% to 69%. In the case of Opus, a decrease of 14% from 87% to 73% is noted. Interestingly, the false-positive rate of detection is stable throughout for both SILK and Opus codecs. These results indicate that occupancy has some 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 allowing the sound waves ample access to most parts of the room.

However to confirm our hypothesis that attack accuracy is sensitive to absorbtion, we carried out another experiment. In this we examine the efficiency of attack when a speaker is surround by other people. For instance, when a hostage is making a statement whilst surrounded by kidnappers, or team participants are huddled around a microphone during a conference call. In this case the speaker is stitting down while being surrounded by people. All other test parameters remained the same. We observed that the detection rate incurs a steep fall to 50% after around 40% of the surrounding space is blocked. As the surrounding space is progressively blocked, the impact of the walls on the room behaviour is replaced by the sound scattering properties of people. In the worst case, only a minimal amounts of reverberation is created. The detection rate does not decrease to zero because sound being

Packet loss%	SILK	OPUS
0	82.30/1.63	87.40/0.71
5	80.31/1.89	85.60/0.94
10	70.18/4.22	80.45/0.97
20	51.91/14.23	74.71/3.66
50	34.00/18.90	46.95/8.07

TABLE II
IMPACT OF PACKET-LOSS ON ATTACK EFFICIENCY

a pressure wave can bend around obstacles albeit at a reduced amplitude.

C. 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. The impact of all these on the codec is that packets that arrive too late at the destination are not played out (discarded). This in turn impacts attack efficiency since the loss of audio information can negatively impact the generation of a reliable fingerprint — audio is damaged to the extent that the reverberant signal components are missing.

Since packet delays and losses damage audio quality, most codecs used by secure messaging systems implement a (packet) loss concealment strategy. This attempts to maintain a perceptual level of voice quality despite any residual packet loss. Often, this is implemented by some form of modification to the signal waveform involving the generation of missing audio segments that are used to replace the speech segments that are lost or delayed due to noisy channels. One class of techniques is the use of insertion schemes, that replace missing speech segments with silence or a copy of a recently delivered segment with minor modifications. An alternate approach used by both OPUS and SILK codecs, is the replacement signal is generated using the frequency spectrum of recent segments as this results in better perceptual quality. 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. A hybrid approach is to play out the segment that is still on time. 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. A Pica8 3920 SDN switch used for routing flows between source and destination pairs. The switch was configured to drop packets from the source-destination flows at different rates of packet loss.

For packet loss rates of around 5-10%, we find that the attack efficiency is fairly high with low FPR and reasonable detection rates of above 80% in the case of Opus. In the case of SILK codec, the detection rates are around 80% until 5% jitter and then reduce to 70% for 10% jitter. For higher rates of jitter, 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 jitter (10%), increasing only to around 8% for 50% jitter. The attack efficiency degrades faster when operating via the SILK codec for increasing jitter; beyond 10% jitter, FPR degrades to 14–18% which is very 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.

D. Discussion

The diffraction properties of sound waves can be a source of new attacks on communication privacy. Sound waves can record information about the location of the speaker as well as the recorder. We found that it is possible to distinguish between indoor recording locations even when the rooms were identical in geometry.

First, we experimented in locations with significant diversity from warehouses to car parks and a church. Our main finding was that fingerprinting and location detection via a trained classifier that has access to samples from all said locations was not just possible, but could be carried out with detection rates of between 85% and 90% and a false-positive rate that's close to zero in good conditions. Good conditions are not ideal conditions: the speaker must speak for a minimum of two to four seconds for the detection rate to be high and the false-positive rates to be close to zero. We think this is quite reasonable and most conversations often last longer than this period of time.

Second, we experimented with locations with relatively little diversity, i.e similar sized rooms in an office building, that are customised (room contents) by their occupants. Given the very low false-positive rates (0.0003%), 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 carry out background noise removal that negatively influences detection. In most cases, the steady state detection rates are between 60 and 88%, with a room occupancy of less than 50%, and a reliable network connection with 5–10% network jitter.

Our experiments confirm the hypothesis that reverberant sound component, if extracted completely and without being polluted by direct sounds and early reflections can be used to generate location fingerprints with high detection and low false-positive rates of detection. We have proposed a suitable technique for extracting and exploiting the reverberant component and successfully demonstrated how to build location fingerprints.

VII. RELATED WORK

Users who wish to hide their location whilst online, use anonymous communication techniques to hide their IP address.

However, user location information can still leak via side channels. Acoustic information in In the architectural acoustics community [17], it is well known that the broad physical characteristics – physical dimensions of the room and the furnishings used – can affect the quality of sound heard by a human. It is worth noting that we are our focus is on passive sensing of location information, i.e no explicit addition of sound is allowed by the attacker within our threat model.

A. Passive approaches

There is prior art in using acoustic information for indoor localisation. Azizyan proposed SurroundSense [1] which distinguishes neighbouring shops on the basis of sound amplitude. They combine this with camera and accelerometer inputs to detect overall shop ambience (sound, light, and decor). They report fairly low accuracy levels on the basis of amplitude distribution alone.

Tarzia et al. proposed ABS [31], which measures location ambience using low-frequency background sounds such as whirring of computers and fans, and the buzz of electrical equipment. This is the closest paper to our work since their technique is fully passive. ABS divides time into equal sized time windows and computes the power spectrum (amplitude in each frequency band) in each window, obtaining one result vector per window. It then chooses the most frequently occuring power spectrum vectors from the set of all result vectors, to filter out transient sounds such as human voices. ABS does not function well unless using raw sound traces. When subject to the compressive effects of modern speech codecs, most background sounds except those of a transient nature are removed. This is especially aggressive when using VoIP applications to transmit sounds, which apply significant amounts of filtering and compression to support low-latency communication. ABS is also unsuitable as a side-channel technique because of the increasing use of noise-cancelling microphones which filter out background sound. Noise cancellation is also applied by some VoIP applications such as Skype [14], as well as by sound recording applications, in order to increase the quality of recorded speech and music.

Another relevant work is SoundSense [19], proposed by Lu et al. SoundSense applies a classifier to partition a given sound sample into a sequence of events. It identifies transient background sounds such as passing trains and associates each location with a set of identifiable background sounds which forms the location fingerprint. To determine the location, it performs a reverse look up on this database given an audio stream. SoundSense is ill-suited for localisation, since it can only work in locations where the 'vocabulary' of external sounds uniquely defines the location. Thus it cannot distinguish between two quiet locations. Or, between adjacent rooms in the same area, unless they are characterised by one or more frequent, repeating, transient sounds. SoundSense and similar works from the signal processing field, can however distinguish between broad classes of locations such as distinguishing a street from an airport and perform comparably with human listeners [10]. These works propose feature sets relevant for partitioning sound [8] and advanced signal decomposition techniques [6] for acoustic context (hence location) identification.

In contrast, VoipLoc depends on unamplified human speech as opposed to background noise which ensures it does not get filtered out using noise-cancelling functions in microphones or signal processing software. Codecs are designed to be carry human voice and therefore is hard to filterout, indeed the whole purpose of audio codecs is to reject all but speech. It is highly robust to compressive codecs except at very low bitrates, where the transfer of the primary audio signal itself is poor. Location identification using human voice can be fairly challenging given the variable signal amplitudes and energy envelope across frequencies. Despite this, with appropriate normalisation, and by filtering out ambient sound we are able to achieve fairly reasonable rates of location identification.

Another set of works, from the digital forensics community leverage room models to determine the authenticity of a given audio recording by linking a recorded audio to a physical location. Kraetzer et al. [15] showed that it was possible to distinguish between recording locations using the recorded music. They developed a set of 63 statistical features similar to those used in SoundSense [19], consisting of 56 mel-cepstral domain based features and 7 time-domain based features. Their work showed that it was possible to distinguish between recording locations at a macro level, such as between: a large office, a small office, a bathroom, a laboratory, a lecture hall, a parking lot, and a corridor. Follow up work in audio forensics by Usher et al. [34] generalised 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 [20]. .

In the content-based audio processing community, audio fingerprinting (also called acoustic fingerprinting) is used to map an audio recording to a unique identifier. These techniques [5], [12], [20], [36] are suitable for searching for a noisy song snippet within a music database. These techniques have little to do with location fingerprinting, and instead focus on fingerprinting audio content (identifying the speaker or the music clip).

B. Active approaches

There's also prior art from the architectural forensics community. These works [23], [27], [32] apply forensic measurement techniques to develop room models to build auralisation systems — synthetic generation of sound in simulated locations. Or, to develop a high-fidelity location acoustic model in order to be able to replicate acoustic effects of a room over any anechoic signal [11]. The principle idea is to inject well designed impulse signals into a location and measure the impulse response to develop a room model. The most commonly used impulse pattern is a 'pop' impulse composed of a single maximum-amplitude signal at maximum frequency for a very short period of time, typically a few milliseconds. The impulse response appears as a series of reflections most of which are temporally non-overlapping with others. Extracting a room model from the impulse response is effectively a

fingerprinting process. These techniques depend on the use of powerful speaker equipment to actively inject a unit impulse signal and record the location response using an array of several carefully placed microphones. The state-of-the-art technique [9] requires four microphones spaced exactly 1 metre apart. In comparison, VoipLoc is fully passive technique and does not depend on the use of sophisticated equipment or specific configurations of microphone arrays at the attack location. It uses human voice as the impulse signal and a single microphone — which is significant challenging not just due to complex interference patterns caused by overlapping signals, but also due to compressive codecs and network jitter.

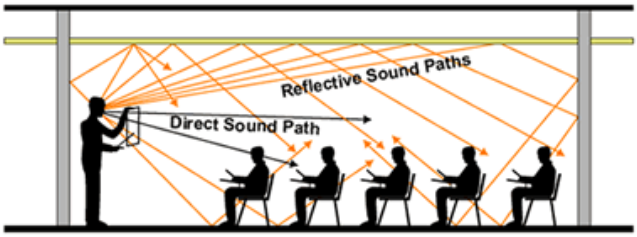
Sonar systems are capable of location fingerprinting using mapping techniques. Both passive [2], [33] and active techniques [30] have been proposed using ultrasonic sonars for above ground as well as underground location mapping. While the output of these systems can very well be used for location fingerprinting, the omnidirectional, unfocused, and insensitive microphones built into modern computing devices are not serviceable for building a rich sonar system. Thus traditional sonar techniques, while excellent at range mapping from a distance are of little use in our context where the victim is only equipped with a humble computer microphone.

C. Future work: privacy implications of location fingerprinting

Applications supporting and accepting voice based communications are very popular today. Humans record and exchange audio data on a planetary scale. For instance, 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. Voice also serves as an increasingly popular method of user-device interaction in smart devices, for instance to dictate emails to a smartphone. For these reasons, it's important to resist side-channel attacks instituted via audio-calling technologies.

'See behind the camera' attacks: The typical audio conference call involves a human speaking in front of a camera which records a video stream of the information in front of the camera. This is streamed electronically along with the speaker's voice to one or more destinations. The speakers voices are a series of sound signals at various frequencies. Both the primary wave and subsequent reflections are recorded and transmitted to other parties. Since the subject is facing the camera, most of the sound energy is directed in the opposite direction to which the camera is facing. The resulting sound reflections carry information about the areas behind the camera. This can have privacy implications since information about the scene behind the camera is being captured.

Location identification within a given room: VoipLocate is a macro-location fingerprinting and identification technique. However, users of audio-calling or audio-recording technologies may want to specifically hide their location within the room as well. For instance, consider a citizen journalist attending a secret meeting where illegal activities are being planned. The journalist wants to secretly record and publish the proceedings. In this case, would the audio-reflection characteristics give away his/her location within the room? Since both the primary wave from the speaker and the reflections are being recorded and made available over the Internet, the timing pattern between consequtive reflections of the primary wave might reveal information about the specific location of the recorder within the room. Specific features include time of arrival and sound pressure level. Once the location of the recorder is known, the adversary could consult CCTV recordings of the room to identify the set of people who were standing at the said location during the time of the meeting, substantially compromising the citizen journalist's identity.



VIII. Conclusion

Location information is embedded into human voice due to the laws of acoustic physics — the reflections of direct sound interfere with each other and with direct sound resulting in a rich interference pattern carried by encoded human voice, which forms the basis of location fingerprinting. Given the wide usage of smartphones and VoIP tools among the wider public to record and transmit audio, this work may have important implications for user privacy.

In contrast to past work that focused on extracting information from background sources, side-channels that exploit human voice as the carrier signal are hard to filter-out as transmission of human speech is the whole point of the communication channel. Modern codecs used in secure messaging clients detect and retain relevant signal components (such as speech or music) and subsequently apply a range of compressive filtering, and whitening techniques. These techniques are designed to remove the impact of steady background sounds and transient interference sources, on listening quality. However, voice itself must be transmitted with high fidelity since that is the primary goal of codec operation. Further, the complex wave interference patterns are a fundamental function of the location, and cannot be trivially reversed due to the time-overlapping nature of the direct sound with its higher order reflections (see section II-A). Finally, the attacks are completely passive, as there is no requirement for active agents to inject marker patterns.

References

[1] 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.

[2] Gaetano Borriello, Alan Liu, Tony Offer, Christopher Palistrant, and Richard Sharp. Walrus: Wireless acoustic location with room-level resolution using ultrasound. In *Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services*, MobiSys '05, pages 191–203, New York, NY, USA, 2005. ACM.

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

[4] 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.

[5] P. Cano, E. Batle, T. Kalker, and J. Haitsma. A review of algorithms for audio fingerprinting. In *Multimedia Signal Processing, 2002 IEEE Workshop on*, pages 169–173, Dec 2002.

[6] 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.

[7] 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.

[8] 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.

[9] Ivan Dokmani, 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.

[10] 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.

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

[12] Jaap Haitsma. A highly robust audio fingerprinting system. In *Proceedings of the ISMIR*, pages 107–115, 2002.

[13] 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.

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

[15] 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.

[16] E. Kruger and H.W. Strube. Linear prediction on a warped frequency scale [speech processing]. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(9):1529–1531, Sep 1988.

[17] H. Kuttruff. *Room Acoustics, Fifth Edition*. Taylor & Francis, 1973.

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

[19] 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.

[20] 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.

[21] 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.

[22] Erwin Meyer. Reverberation and absorption of sound. *The Journal of the Acoustical Society of America*, 8(3):155–161, 1937.

[23] G.M. Naylor. Odeonanother hybrid room acoustical model. *Applied Acoustics*, 38(24):131 – 143, 1993.

[24] 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.

[25] 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.

[26] Theodore Rappaport. *Wireless Communications: Principles and Practice*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 2001.

[27] Jens H. Rindel and Claus L. Christensen. Room acoustic simulation and auralization how close can we get to the real room ? *The Eighth Western Pacific Acoustics Conference*, (1025), 2003.

[28] Jonathon Shlens. A tutorial on principal component analysis, December 2005.

[29] Hans W. Strube. Linear prediction on a warped frequency scale. *Journal of the Acoustical Society of America*, 68(4):1071–1076, October 1980.

[30] 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.

[31] 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.

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

[33] S. Tucker and G. J. Brown. Classification of Transient Sonar Sounds Using Perceptually Motivated Features. *IEEE Journal of Oceanic Engineering*, 30(3):588–600, July 2005.

[34] 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.

[35] Minjie Xie, Dave Lindbergh, and Peter Chu. From itu-t g.722.1 to itu-t g.722.1 annex c: A new low-complexity 14khz bandwidth audio coding standard, 2009.

[36] Sascha Zmudzinski and Martin Steinebach. Psycho-acoustic model-based message authentication coding for audio data. In *Proceedings of the 10th ACM Workshop on Multimedia and Security*, MM&Sec '08, pages 75–84, New York, NY, USA, 2008. ACM.