VoipLoc: Establishing VoIP call provenance using acoustic side-channels

Shishir Nagaraja and Ryan Shah

{shishir.nagaraja,ryan.shah}@strath.ac.uk
University of Strathclyde

Abstract. We develop a novel technique to determine call provenance in anonymous VoIP communications using acoustic side-channels. The technique exploits location-attributable information embedded within audio speech data. The victim's speech is exploited as an excitation signal, which is modulated (acted upon) by the acoustic reflection characteristics of the victim's location. We show that leading VoIP communication channels faithfully transfer this information between sender-receiver pairs, enabling passive receivers to extract a location fingerprint, to establish call provenance. To establish provenance, a fingerprint is compared against a database of labelled fingerprints to identify a match. The technique is fully passive and does not depend on any characteristic background sounds, is speaker independent, and is robust to lossy network conditions. Evaluation using a corpus of recordings of VoIP conversations, over the Tor network, confirms that recording locations can be fingerprinted and detected remotely with low false-positive rate.

1 Introduction

Resisting widespread censorship has pushed users towards using VoIP over anonymous communication channels. In this paper, we develop a novel technique, based on acoustic side-channels, which exploits location-attributable information embedded within audio-speech data transmitted by most VoIP clients.

The VoIP-call provenance problem is depicted in Figure 1, involving two or more participants initiating a VoIP session. They are concerned about the risk of being deanonymised by a malicious intermediary or end-participant, who might geolocate them using their source IP address. Hence, they use a low-latency anonymous communication network such as Tor, to contact the VoIP backend service. Every participant uses off-the-shelf equipment, such as a smartphone, to participate in the call.

We now establish requirements for practical VoIP call provenance techniques:

- *Stealth* – Provenance should be established via passive methods, in order to prevent detection by call participants. Previous work includes active methods for establishing call provenance in anonymous VoIP channels, such as ultrasound squeaks injected by malware [insert citation]. However, no passive methods have been proposed to the best of our knowledge. With an obvious stealth advantage, passive attacks are more powerful compared to active attacks. They do not expose themselves, whilst active attacks can be detected by the enemy, or be removed by aggressive high/low-pass filters employed by audio codecs.

– *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. Existing approaches that leverage call-route [2], IP-geolocation, or cell-tower proximity, can establish provenance to a large geographical area, i.e. establish coarse-grained provenance.
– *Uniqueness* – A provenance fingerprint must be distinct from other sources.
– *Time invariant* – Provenance fingerprints should not rely on leveraging background sounds for fingerprinting, such as proximity to an external noise source that might be unavailable, resulting in a unreliable provenance technique. A number of existing approaches [1, 27, 18] consider background noise sources, but do not meet the above criteria.
– *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. Aside from unreliability, an additional disadvantage is that VoIP clients employ aggressive filtering of background noise sources, in order to reduce traffic loads and improve call quality. Thus, previous approaches [1, 27, 18] are unsuitable for VoIP call provenance.

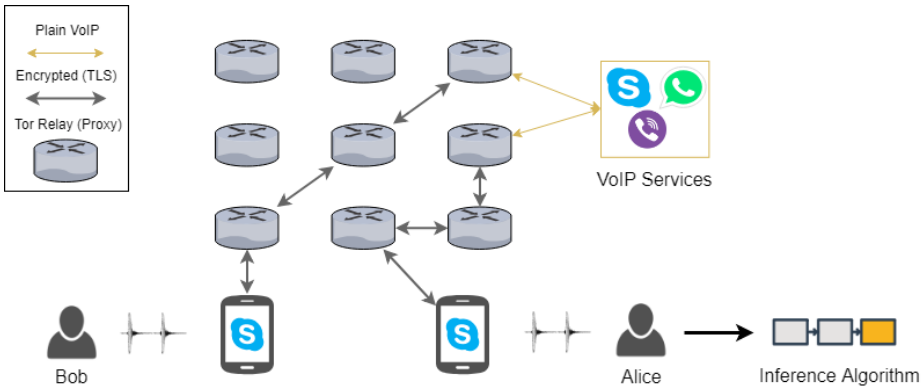

Fig. 1: The VoIP-call Provenance Problem

In this work, we develop a new audio-call provenance technique that works on anonymous VoIP channels. This technique is fully passive, time-invariant, robust to temporal conditions, and establishes fine-grained provenance down to a specific room, with a low false-positive rate, by leveraging the sound-reflection behaviour of the call location.

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 analyzing 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 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, 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 localization, we report attacks that can 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 secrecy of the NGO. We are specifically interested in passive adversaries as the undetectability of the attacks poses a greater threat then active adversaries.

3 Attack Technique

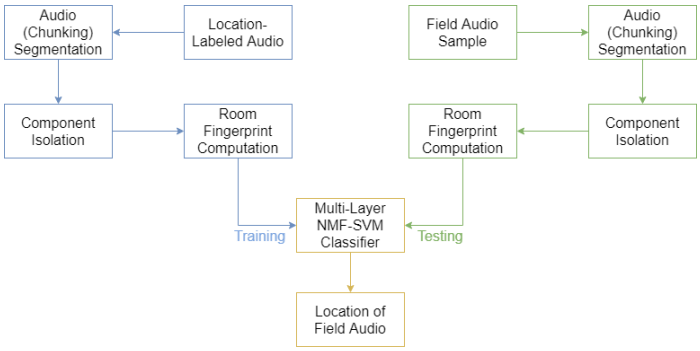

Fig. 2: Attack workflow

The high-level architecture of our attack technique is given in Figure 2. The reverberant component is extracted from the speech signal at the receiver using a multi-layer NMF-SVM classifier which maps the reverberant component to a physical location.

Sound components: Speech signal at the receiver consists of three components (see Figure 3). First, the *direct sound*, is the sound transmitted in a direct path from the speaker to the microphone with no reflections. Second, *early 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.

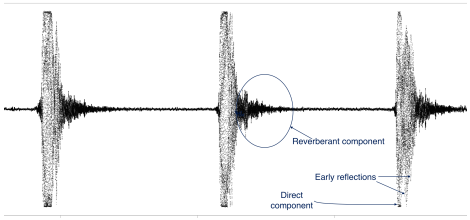

Fig. 3: Components of a sound sample

To fingerprint a location, the reverberant component is the most relevant, as it is a stable function of acoustic information diffused throughout the location. The reverberant component

3.1 Audio segmentation

Speech segmentation, breaks up the speech signal into chunks such that each chunk corresponds to a single *utterance* (the smallest unit of human speech) from one speaker. Speech segmentation involves a number of algorithms, namely a voice activity detector and a silence detector [24]. We used a combination of volume, spectral energy, and spectral flatness for creating a predictor for speech segments, following standard practice [12].

3.2 Isolating the reverberation component

After chunking, the next step is to isolate the reverberant component in each of the chunks. Each chunk contains several direct-sound components, emerging from the speaker's voice box (Larynx). The challenge here is to extract the reverberant component in its purest form, by removing early-reflections and direct-sound that are time-overlapped over the reverberant component. This is necessary, as the direct sound is speaker dependent (by definition), while the reverberant component is location dependent. Overlapping occurs in two scenarios. First, the isolation of the sound components within an utterance. And, second across utterances for eg. early reflections from one utterance may mix (and pollute) the reverberant component of a previous utterance that's still above the noise floor.

To reverse the effects of overlapping and remove any random background noise sources, VoipLoc uses a compressive deomposition technique (Fig. 4), which partitions recorded audio into direct sound, early reflections, and reverberant component. Partitioning leverages the observation that the shape and form of early 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 transformation noise. The input audio segment is therefore represented by an algebraic combination of direct-sound and the reverberant component. Partitioning is carried out using a deep NMF-SVM classifier that combines location fingerprinting with classification, it has three stages: (a)*K layers of nested NMF decomposition with pooling*; (b) *Fingerprint generation function*; and (c) the *SVM classifier*.

Recorded audio (time-domain signal) is loaded into an input speech matrix O. Each row corresponds to an audio trace for each utterance; the time-series of signal amplitude is in dB.

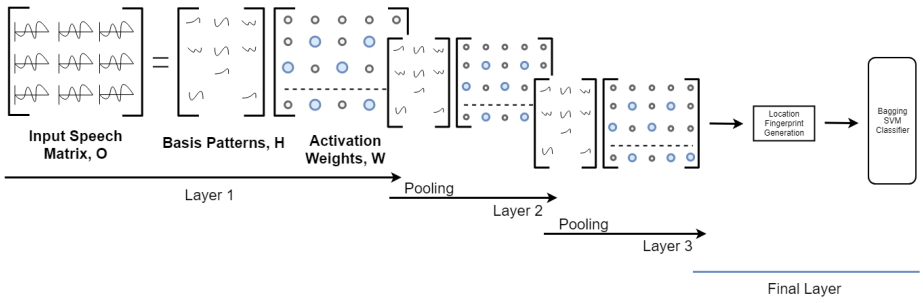

Fig. 4: Multi-layer NMF-SVM classifier

$$\begin{aligned}\text{Initial decomposition: } O^1 &= H^1 W^1 \\ \text{Second decomposition: } W^1 &= H^2 W^2 \\ &\vdots \\ \text{Final decomposition: } W^{(K-1)} &= H^K W^K\end{aligned}$$

In the initial decomposition, the audio trace is partitioned into *compressed*-patterns (columns of H) and their activation weights. Input audio is expressed as a linear combination of patterns (H) and mixing weights (W) as: $O_{ij} = \sum_{k=1}^{K} H_{ik} W_{kj}$. The key intuition here is to represent the speech segment using the sparsest and fewest number of sound patterns. 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 times/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 multiplicative update rules [17] (this is standard practice):

$$H_{ir} = \frac{H_{ir} \sum_r \frac{O_{ij}}{(HW)_{ij}} H_{rj}}{\sum_l H_{jr}}$$

$$W_{rj} = W_{rj} \sum_i H_{ir} \frac{O_{ij}}{(HW)_{ij}}$$

Multiple decomposition layers further promote sparsity and this is motivated by the power and accuracy of Deep Neural Networks (DNN) [15]. In subsequent decomposition steps (Eqn. 1), sound components are partitioned into sub-components, sub-sub-components, and so on. Thus the output of the final decomposition layer can be formally stated as $O^1 = H^1(H^2(H^3(....H^K W^K))))$, and the optimisation function as:

$$O^1 - H^1(H^2(H^3(...(H^K W^K)))).$$

To create robustness to small variations in the attack, such as the positionality of the speaker and their movement within the location, we include a pooling step. Pooling

is commonly found in DNNs [15]. In our technique, each decomposition step is followed by max-pooling, a moving-window function which takes as input the rows of the weight matrix W^k, and replaces a subset of the row by the maximum value of the subset. This approach was first suggested by Boureau et al. [3]. 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.

Input: $O \in \mathbb{R}^{m \times n}$
The number of layers K
The number of columns to pool across, poolsize
Output: Traffic decomposition at each layer k
for k *in* $1 : K-1$ **do**
  while $\epsilon > 0.05$ **do**
    $H^k = H^k \odot \frac{O^k (W^k)^T}{H^k W^k (W^k)^T}$
    $W^k = W^k \odot \frac{(H^k)^T O^k}{(H^k)^T H^k W^k}$
    $\epsilon = \sqrt{\sum_i^m \sum_j^n (O^k_{ij} - (H^k W^k)_{ij})^2}$
    $O^{k+1} \leftarrow F(W^k)$
  end
end

Algorithm 1: Multilayer decomposition

The nested signal-decomposition algorithm is given in Algorithm 1. 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 using Eqn. 1: $O' = H^1 W^X$. 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 reverant signal and noise), there is no upper limit. We set the number heuristically at 100 in all our experiments.

3.3 Fingerprint computation

To derive a fingerprint from the reverberant component, several steps are involved as shown in Figure 5: preprocessing to remove noise, aggregation of signal power in relevant bands to produce the fingerprint vector, before the classification layer.

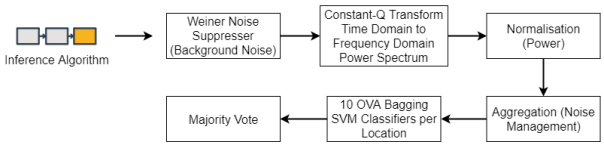

Fig. 5: Computing the fingerprint after multi-layer NMF decomposition

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 [23], 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.

Reflection measurement The key idea underlying location fingerprint computation, is to compute the signal power in each frequency band within the reverberant component normalized 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 signal power for different locations within a given room is inversely related to the frequency. High frequencies are only correlated at close proximity, while low frequencies (above Schroeder frequency) can be correlated anywhere within the given location [24]. Thus, signal power at lower frequencies (20Hz — 2Khz) is ideal for room fingerprinting. See Appendix A for theoretical background underlying the range selected.

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 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 [4]. 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 technique described in Section 3.2. The

output of the transform is a vector which contains the signal power in each frequency band. We call this the *fingerprint vector*.

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. The design of the merge function is straightforward. Vector p contains the normalized 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 Fingerprint classification

The final layer is a supervised classification layer. The fingerprint vector is input into an ensemble of weak classifiers called the *bagging* Support Vector Machine (SVM) classifier, which maps the input to a location. The classifier is trained positively with location traces and negatively against traces from other locations (and unlabelled traces).

Prefiltering for contamination-resistance: Since the classifier is trained negatively against unlabelled fingerprints, it is important to ensure that this is not contaminated with positive samples. Otherwise, the classifier would fundamentally fail having been trained both positively and negatively against fingerprints from the same location. To prevent this, we first cluster all the fingerprints using the (parameter free) Xmeans algorithm. All unlabelled fingerprints that are clustered with a positive fingerprint are assigned a positive label.

Classifier choice: SVM was chosen due to the high dimensionality of fingerprint vectors. An SVM identifies the optimal separating hyperplane that maximises the *margin of separation* between fingerprint vectors p. 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. A conventional approach to applying SVM would involve using N classifiers, one for each location.

However, 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. However, vanilla SVM classifiers are highly sensitive to the presence of unlabelled data. Therefore, we address the instability caused by unlabelled data 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 i^{th} room are trained using a tenth of the samples from the training set for the i^{th} 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 ten classifiers for i^{th} room are aggregated with a majority vote. Overall, for n rooms we train a total of $n \times 10$ SVM classifiers, which is still $O(n)$ classifiers as in conventional application of SVM. This method of combining classifier output is based on the technique first introduced by Mordelet and Vert [21].

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 from a number of datasets.

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 [5] 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. The variable cutoff ensures that for high-pitched voices the cutoff is correspondingly higher, allowing all but the lowest harmonics to be filtered out.

SILK codec (Skype): The widely used Skype VoIP services uses SILK codec [13] 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 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 %. 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 to the receiver which significantly improves fingerprinting.

AMR-WB (Blackberry AIM): AMR-WB (Adaptive Multi-Rate Wideband) is an audio data compression scheme optimized 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 up to 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: is a low-delay generic audio codec [30] standardized 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 6 show that a bitrate of at least 24–32Kbps is required to achieve reasonable fingerprinting success.

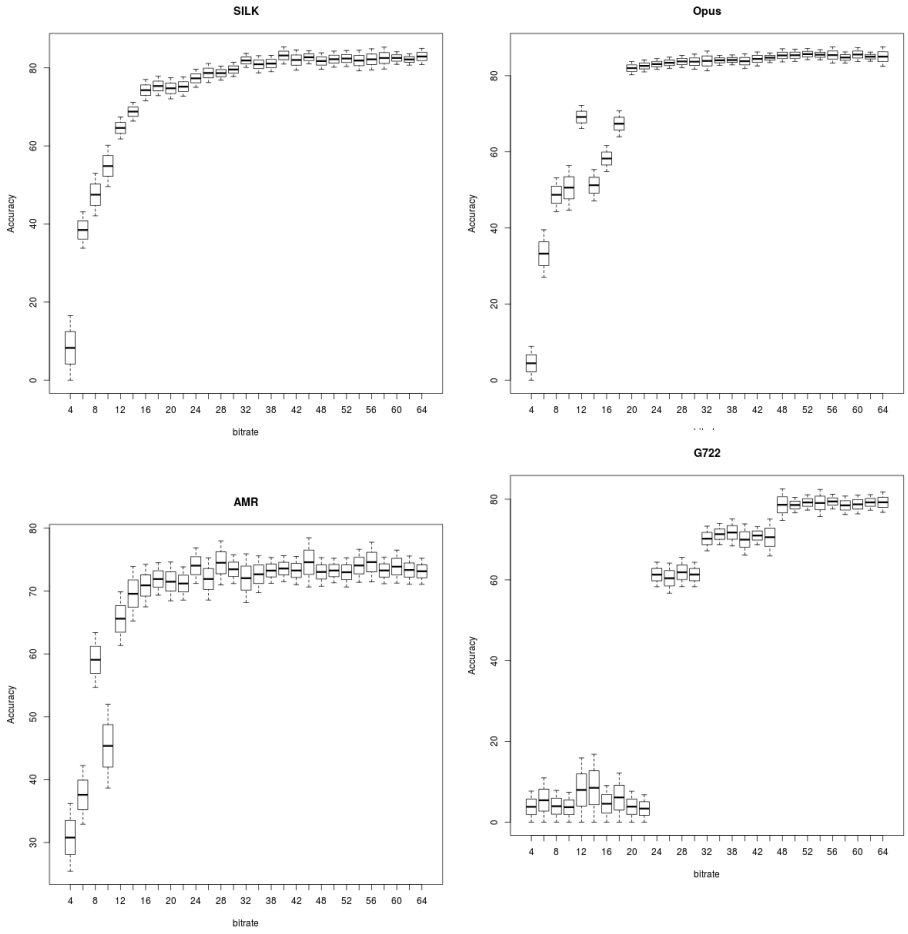

Fig. 6: Location accuracy vs bitrate

Opus (Facebook Messenger): The Opus codec is a framework for composing high quality codecs, namely SILK [13] and CELT [20]. 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 less than 50% at 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 (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. Shorter the frame, 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 4.3.

4.3 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). This in turn impacts attack efficiency since the loss of reverberant 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 reduce 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 modifying the signal waveform by synthetically generating missing audio segments. 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.

Packet loss%	SILK (TPR%/FPR%)	OPUS (TPR%/FPR%)
10	83.32/0.55	91.47/0.50
15	82.15/1.89	85.60/0.86
20	70.10/3.75	80.45/0.93
25	54.47/14.86	74.71/2.59
30	32.09/17.00	46.95/8.01

Table 1: Impact of packet-loss on attack efficiency

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.

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.

5 Discussion

During a VoIP call, human speech recorded at the speaker's location contains captures the reflection characteristics of the room used by the speaker. These are preserved by audio codecs and transmitted to the recipient. Our attack technique allows participants of a VoIP call to extract the reflection charateristics from recorded audio and fingerprint the location of call participants down to the specific room used to make a call. We found that routine customisations by occupants to make the room 'their own' is enough to be able to distinguish between indoor recording locations even when the rooms were identical in geometry.

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.

Location confirmation: VoipLoc may be used in conjunction with macro-geolocation techniques such as PinDr0p [2] and cell phone tracking techniques. While these tech-

niques point out the rough geolocation coordinates, VoipLoc can be used to provide fine-grained location 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.

Indoor vs Outdoor application: Outdoor locations typically demonstrate poor reverberance characteristics and require excitation signals of higher amplitude than typical of human voices whilst in a VoIP conversation. For this reason, the applicability of this work is primarily of significance in fingerprinting indoor locations. However, where voice signals are amplified such as in outdoor gatherings, VoipLoc may be usable without requiring changes to the technique if any reflections are recorded.

6 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. It is well known that the echo-location characteristics of a location affects the quality of sound recorded [16]. Despite this, no study of passive location-privacy attacks has been carried out thus far.

6.1 Passive approaches

Azizyan proposed SurroundSense [1] which distinguishes neighboring shops on the basis of sound amplitude. They combine this with camera and accelerometer inputs to detect overall shop ambiance (sound, light, and decor). They report fairly low accuracy levels on the basis of amplitude distribution alone.

Tarzia et al. proposed ABS [27], a fully passive technique, to identify a location using low-frequency background sounds such as whirring of computers and fans, and the buzz of electrical equipment. ABS computes the power spectrum vector (amplitude in each frequency band). It filters out transient sounds such as human voices. ABS does not function well unless using raw sound traces. When subject to the compressive effects of aggressive VoIP speech codecs, most background sounds except those of a transient nature are removed. ABS is also unsuitable as a side-channel technique because of the

increasing use of noise-canceling microphones that filter out background sound. Noise cancellation is also applied by VoIP applications such as Skype [13] and professional audio recorders to increase the quality of recorded speech and music.

Kraetzer et al. [14] propose a location identification technique that relies on repetitive patterns of music played in a location. Lu et al. (SoundSense) [18] generalises this to use background sounds such as passing trains and associates each location with a set of identifiable background sounds. To determine the location, they perform a reverse look up on this database given an audio stream. SoundSense and other similar works [9, 7, 6] are ill-suited for localization, they depend on a 'vocabulary' of external sounds that uniquely defines the location. While this is useful in distinguishing between a street from an airport it isn't serviceable for confirming (or detecting) the location in which a VoIP conversation took place. Usher et al. [29] 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 [19].

In contrast, VoipLoc depends on unamplified human speech as opposed to background sounds. Hence, acoustic data critical to fingerprint generation is not impacted by noise-canceling or compressive codec functions unlike past works. VoipLoc is 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 is fairly challenging given variability in signal amplitudes and power spectrum. Despite these challenges it is possible to achieve reliable location identification at scale with a well designed source separation method combined with appropriate normalization and noise filtering techniques.

In the context of multimedia copyright techniques, audio fingerprinting (also called acoustic fingerprinting) is used to map an audio recording to a unique identifier. These techniques [31, 19] 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).

There's also related work in the telephony-spam and mobile-fraud detection literature. PinDrop [2] 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. Since telephony spam assume static devices and location, VoipLoc can be used to label an entire call-centre room as associated with spamming activity instead of waiting for each device within the room to be used for spamming at least once. Another difference is that PinDrop examines the influence of device and network characteristics to develop "source biometrics", while VoipLoc focuses on the impact of location characteristics on speaker voice. Thus in the context of VoIP over anonymous communication networks, PinDrop's fingerprint could be damaged by the use of path-altering proxies such as Tor routers. For these reasons, it is fair to state that VoipLoc is resilient to the changes arising from network congestion. Another interesting difference is that PinDrop generates a device specific location fingerprint, whereas VoipLoc is device independent.

6.2 Active approaches

Recent works on privacy and acoustic channels include sensory malware, notably Soundcomber [25] 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 [11].

Another approach is to inject well designed impulse signals into a location and measure the impulse response [22, 28, 26] 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 [10]. The state-of-the-art technique in this space [8] requires four microphones spaced exactly one meter apart. In comparison, VoipLoc is a fully passive technique and does not depend on the use of sophisticated equipment. 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.

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 eachother 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.

References

1. Azizyan, M., Constandache, I., Roy Choudhury, R.: Surroundsense: Mobile phone localization via ambience fingerprinting. In: Proceedings of the 15th Annual International Conference on Mobile Computing and Networking. pp. 261–272. MobiCom '09, ACM, New York, NY, USA (2009). https://doi.org/10.1145/1614320.1614350, http://doi.acm.org/10.1145/1614320.1614350
2. Balasubramaniyan, V.A., Poonawalla, A., Ahamad, M., Hunter, M.T., Traynor, P.: Pindr0p: Using single-ended audio features to determine call provenance. In: Proceedings of the 17th ACM Conference on Computer and Communications Security. pp. 109–120. CCS '10, ACM, New York, NY, USA (2010). https://doi.org/10.1145/1866307.1866320, http://doi.acm.org/10.1145/1866307.1866320
3. Boureau, Y.L., Ponce, J., Lecun, Y.: A theoretical analysis of feature pooling in visual recognition. In: 27TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, HAIFA, ISRAEL (2010)
4. Brown, J.C.: Calculation of a constant Q spectral transform. Journal of the Acoustical Society of America **89**(1), 425–434 (1991)

5. Calhoun, S., Carletta, J., Brenier, J.M., Mayo, N., Jurafsky, D., Steedman, M., Beaver, D.: 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 (Dec 2010). https://doi.org/10.1007/s10579-010-9120-1, http://dx.doi.org/10.1007/s10579-010-9120-1
6. Chu, S., Narayanan, S., Kuo, C.C.: Environmental sound recognition with time;frequency audio features. Audio, Speech, and Language Processing, IEEE Transactions on **17**(6), 1142–1158 (Aug 2009). https://doi.org/10.1109/TASL.2009.2017438
7. Davis, S., Mermelstein, P.: 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). https://doi.org/10.1109/TASSP.1980.1163420
8. Dokmani, I., Parhizkar, R., Walther, A., Lu, Y.M., Vetterli, M.: Acoustic echoes reveal room shape. Proceedings of the National Academy of Sciences **110**(30), 12186–12191 (2013). https://doi.org/10.1073/pnas.1221464110, http://www.pnas.org/content/110/30/12186.abstract
9. Eronen, A.J., Peltonen, V.T., Tuomi, J.T., Klapuri, A.P., Fagerlund, S., Sorsa, T., Lorho, G., Huopaniemi, J.: Audio-based context recognition. IEEE Transactions on Audio, Speech and Language Processing **14**(1), 321–329 (2006). https://doi.org/10.1109/tsa.2005.854103, http://dx.doi.org/10.1109/tsa.2005.854103
10. Farina, A., Ayalon, R.: Recording concert hall acoustics for posterity. In: 24th AES Conference on Multichannel Audio, Banff, Canada. pp. 26–28 (2003)
11. Hanspach, M., Goetz, M.: On covert acoustical mesh networks in air. JCM **8**(11), 758–767 (2013). https://doi.org/10.12720/jcm.8.11.758-767, http://dx.doi.org/10.12720/jcm.8.11.758-767
12. Ishizuka, K., Nakatani, T., Fujimoto, M., Miyazaki, N.: Noise robust voice activity detection based on periodic to aperiodic component ratio. Speech Communication **52**(1), 41 – 60 (2010). https://doi.org/http://dx.doi.org/10.1016/j.specom.2009.08.003, http://www.sciencedirect.com/science/article/pii/S0167639309001277
13. Jensen, S., Vos, K., Soerensen, K.: SILK Speech Codec. Tech. Rep. draft-vos-silk-02.txt, IETF Secretariat, Fremont, CA, USA (Sep 2010), http://www.rfc-editor.org/internet-drafts/draft-vos-silk-02.txt
14. Kraetzer, C., Oermann, A., Dittmann, J., Lang, A.: Digital audio forensics: A first practical evaluation on microphone and environment classification. In: Proceedings of the 9th Workshop on Multimedia & Security. pp. 63–74. MM&Sec '07, ACM, New York, NY, USA (2007). https://doi.org/10.1145/1288869.1288879, http://doi.acm.org/10.1145/1288869.1288879
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (May 2017). https://doi.org/10.1145/3065386, http://doi.acm.org/10.1145/3065386
16. Kuttruff, H.: Room Acoustics, Fifth Edition. Taylor & Francis (1973), http://books.google.co.uk/books?id=X4BJ9ImKYOsC
17. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: In NIPS. pp. 556–562. MIT Press (2000)
18. Lu, H., Pan, W., Lane, N.D., Choudhury, T., Campbell, A.T.: Soundsense: Scalable sound sensing for people-centric applications on mobile phones. In: Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services. pp. 165–178. MobiSys '09, ACM, New York, NY, USA (2009). https://doi.org/10.1145/1555816.1555834, http://doi.acm.org/10.1145/1555816.1555834
19. Malik, H., Farid, H.: Audio forensics from acoustic reverberation. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. pp. 1710–1713 (March 2010). https://doi.org/10.1109/ICASSP.2010.5495479
20. Maxwell, G., Terriberry, T., Valin, J.M., Montgomery, C.: Constrained-Energy Lapped Transform (CELT) Codec. Tech. Rep. draft-valin-celt-codec-02.txt, IETF Secretariat,

Fremont, CA, USA (Jul 2010), http://www.rfc-editor.org/internet-drafts/draft-valin-celt-codec-02.txt
21. Mordelet, F., Vert, J.P.: A bagging svm to learn from positive and unlabeled examples. Pattern Recogn. Lett. **37**, 201–209 (Feb 2014). https://doi.org/10.1016/j.patrec.2013.06.010, http://dx.doi.org/10.1016/j.patrec.2013.06.010
22. Naylor, G.: Odeonanother hybrid room acoustical model. Applied Acoustics **38**(24), 131 – 143 (1993). https://doi.org/http://dx.doi.org/10.1016/0003-682X(93)90047-A, http://www.sciencedirect.com/science/article/pii/0003682X9390047A
23. Plapous, C., Marro, C., Scalart, P.: Improved signal-to-noise ratio estimation for speech enhancement. Audio, Speech, and Language Processing, IEEE Transactions on **14**(6), 2098–2108 (Nov 2006). https://doi.org/10.1109/TASL.2006.872621
24. Ramirez, J., Segura, J.C., Gorriz, J.M., Garcia, L.: Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition. Trans. Audio, Speech and Lang. Proc. **15**(8), 2177–2189 (Nov 2007). https://doi.org/10.1109/TASL.2007.903937, http://dx.doi.org/10.1109/TASL.2007.903937
25. Schlegel, R., Zhang, K., Yong Zhou, X., Intwala, M., Kapadia, A., Wang, X.: Soundcomber: A stealthy and context-aware sound trojan for smartphones. In: NDSS. The Internet Society (2011), http://www.isoc.org/isoc/conferences/ndss/11/
26. Tarzia, S.P., Dick, R.P., Dinda, P.A., Memik, G.: Sonar-based measurement of user presence and attention. In: Proceedings of the 11th International Conference on Ubiquitous Computing. pp. 89–92. UbiComp '09, ACM, New York, NY, USA (2009). https://doi.org/10.1145/1620545.1620559, http://doi.acm.org/10.1145/1620545.1620559
27. Tarzia, S.P., Dinda, P.A., Dick, R.P., Memik, G.: Indoor localization without infrastructure using the acoustic background spectrum. In: Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services. pp. 155–168. MobiSys '11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/1999995.2000011, http://doi.acm.org/10.1145/1999995.2000011
28. Tsingos, N.: Pre-computing geometry-based reverberation effects for games (2004)
29. Usher, J., Benesty, J.: 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). https://doi.org/10.1109/TASL.2007.901832
30. Xie, M., Lindbergh, D., Chu, P.: From itu-t g.722.1 to itu-t g.722.1 annex c: A new low-complexity 14khz bandwidth audio coding standard (2009)
31. Zmudzinski, S., Steinebach, M.: Psycho-acoustic model-based message authentication coding for audio data. In: Proceedings of the 10th ACM Workshop on Multimedia and Security. pp. 75–84. MMSec '08, ACM, New York, NY, USA (2008). https://doi.org/10.1145/1411328.1411343, http://doi.acm.org/10.1145/1411328.1411343

A Background: the stochastic model for acoustic wave propagation

The characteristic property of a room is its 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 out, where obstructions absorb energy from a subset of the frequency spectrum of the wave and reflect the rest. 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 stochastic model for the reverberant component is only valid after the mixing time and for frequencies above the Schroeder frequency defined as follows:

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

where V is the room volume in cubic meters. RT_{60} is the length (duration) of the reverberant part of the signal – i.e the time taken for direct sound to diffuse into the room until signal power reduces 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, creating a primary challenge for any acoustic fingerprinting process based on the reverberant component.

Therefore it's useful to isolate early reflections when the analysis is targeted at understanding reflection timing or frequency-amplitude distributions to characterize the precise location of the recording microphone within a given room.

Above the Schroeder frequency the sound correlation between two locations in a room is dependent on the wavenumber and the distance between them. Therefore, the assumption that the reverberant-path components are uncorrelated, is only valid for frequencies greater than approximately 1 kHz.

Consider a sound wave from a human source. Let $m(t)$ be the recorded signal. $m(t)$ can be described by the acoustic convolution between the sound source signal $s(t)$ and the L_r-length direct-path coefficients ($d_{*,*}$) and summed with the convolution of $s(t)$ with the $(L-L_r)$-length reverberant-path coefficients ($r_{*,*}$), as shown below:

$$m(t) = \sum_{k=0}^{L_r-1} s(t-k) d_{i,k} + \sum_{l=L_r}^{L} s(t-l) r_{i,l}$$

The direct-path coefficients are the first L_r samples of the L-length energy wave recorded at the receiver. These correspond to the primary wave and the early reflection waves. These reflections arrive via a predictable non-stochastic directional path giving evidence of the immediate geometrical surroundings of the receiver. The reverberant-path coefficients are the remaining $(L-L_r)$ samples of the energy wave. The reverberations are sound reflections which can be modeled as an exponentially decaying, ergodic, stochastic process, with a Gaussian distribution and a mean of zero. Both the early reflections and the reverberant reflections carry different types of information about spatial geometry.

The amount of time required for the early-reflections to diffuse into the room is referred to as the *mixing time* of a room. The stochastic model of sound field is only valid after the mixing time.