Acoustic Eavesdropping through Wireless Vibrometry

Teng Wei[†], Shu Wang[†], Anfu Zhou[*†] and Xinyu Zhang[†]
[†]University of Wisconsin - Madison, [*]Institute of Computing Technology, Chinese Academy of Sciences
{twei7, swang367, azhou9}@wisc.edu, xyzhang@ece.wisc.edu

ABSTRACT

Loudspeakers are widely used in conferencing and infotainment systems. Private information leakage from loudspeaker sound is often assumed to be preventable using sound-proof isolators like walls. In this paper, we explore a new acoustic eavesdropping attack that can subvert such protectors using radio devices. Our basic idea lies in an acoustic-radio transformation (ART) algorithm, which recovers loudspeaker sound by inspecting the subtle disturbance it causes to the radio signals generated by an adversary or by its co-located WiFi transmitter. ART builds on a modeling framework that distills key factors to determine the recovered audio quality. It incorporates diversity mechanisms and noise suppression algorithms that can boost the eavesdropping quality. We implement the ART eavesdropper on a software-radio platform and conduct experiments to verify its feasibility and threat level. When targeted at vanilla PC or smartphone loudspeakers, the attacker can successfully recover high-quality audio even when blocked by sound-proof walls. On the other hand, we propose several pragmatic countermeasures that can effectively reduce the attacker's audio recovery quality by orders of magnitude.

Categories and Subject Descriptors

C.3 [**Special-Purpose and Application-Based Systems**]: [Signal processing systems]

Keywords

Acoustic Eavesdropping; Acoustic-radio Transformation; WiFi Sensing; Vibration Sensing

1. INTRODUCTION

Despite powerful security-proof measures in this digital communication age, sound – the primitive medium of unencrypted human communication and side-product of many private activities – remains a vulnerable source of information leakage. The pursuit for high-fidelity sound capturing has motivated more powerful acoustic hardware on consumer devices. Meanwhile, it creates a looming threat of acoustic eavesdropping that impinges on people's everyday privacy [1]. Even subtle acoustic emanation from keystrokes [2], printers [3], and PC electronic components [4], can be exploited to decode sensitive information. However, such acoustic attacks all require a tampered microphone in the vicinity of the victim sound source which, to some extent, thwarts eavesdroppers who cannot approach the target's physical space.

MobiCom'15, September 7–11, 2015, Paris, France.

ACM 978-1-4503-3619-2/15/09 ...$15.00.
http://dx.doi.org/10.1145/2789168.2790119.

We explore a new eavesdropping mechanism that uses wireless transceivers to decode loudspeaker sounds from afar. Our basic idea lies in a translation between acoustic vibration and radio signal fluctuation. Audio emission causes small vibration of the loudspeaker body. Such minute motion pattern is invisible to human eyes. But it can resonance with radio waves reflected by the loudspeaker, or originating from a wireless transmitter co-located with the loudspeaker. The contaminated radio waves can be captured by a tampered receiver and processed to recover the original audio played by the loudspeaker. We refer to such a remote sound acquisition/recovery system as *wireless vibrometry*.

Built on commonly available radio devices, and harnessing the better penetration of radio signals, such an acoustic eavesdropping system raises alarming issues in security and privacy. Today, loudspeakers are widely used in communication and infotainment systems, *e.g.*, tele-conference call, VoIP hangout and home theater. In such usage scenarios, we envision two categories of threats, categorized according to how the target loudspeaker modulates the radio waves: *(i) Reflective vibrometry:* The adversary is a pair of radio transmitter and receiver. The transmitter continuously sends radio signals, while the receiver decodes sound vibration from the signals reflected and disturbed by the loudspeaker vibration. *(ii) Emissive vibrometry:* The adversary is a radio receiver. The target loudspeaker is co-located with a WiFi radio on the same platform, *e.g.*, a smartphone, smart TV, or speaker dock. The loudspeaker's minute motion causes tiny variation in the WiFi radio's outgoing signals, which can be overheard and leveraged by the adversary to recover the sound.

The concept underlying wireless vibrometry resembles laser radar (LADAR) [5], which is commonly used to test the stability of building structures. A LADAR projects laser beams towards vibrating objects and discerns the vibration patterns based on the shaking reflective beams. In contrast, the unique advantage of wireless vibrometry lies in its ability to penetrate opaque obstacles. This advantage comes with more challenges. The narrow field-of-view of LADAR enables them to measure even micron movements very accurately, because tiny motion can alter the reflecting angle. In contrast, radio devices are far less directive and suffer from severe reflection/diffusive loss. More importantly, under the aforementioned threat scenarios, adversary's radios may not even fall in the line-of-sight (LOS) of the victim loudspeaker and cannot "point" towards the vibrating loudspeaker.

In lieu of such challenges, we explore a new set of mechanisms to extract and boost the acoustic signals from radio channel dynamics.

(i) Basic acoustic-radio transformation (ART) algorithm, which harnesses the received signal strength (RSS) and phase information, readily available on typical radio transceivers [6], to "demodulate" acoustic signals from the target loudspeaker. Our basic idea is to model the procedure of audio vibration disturbing radio waves as a procedure of low-rate amplitude/phase modulation. With this model, we design a frequency-domain demodulator to isolate irrele-

vant radio signal components, extrapolate the audio signals, project them to the time-domain, which eventually become perceptible by human.

(ii) Diversity mechanisms to enhance reflective vibrometry. Whereas LADAR essentially discriminates the displacement of vibrating surface, we found that wireless vibrometry is mostly sensitive to the multipath overlapping patterns, with both analytical and experimental evidence. We design two diversity mechanisms to take advantage of this effect, so as to amplify the loudspeaker vibration.

First, we design a simple frequency selection mechanism to foster the case where multiple signal paths involving loudspeaker reflection can strengthen each other. *Second*, we adapt multi-antenna beamforming to amplify the signal paths reflected by the loudspeaker. Since loudspeaker cannot provide any location or anchoring signals to facilitate the MIMO beamforming, we adopt the strategy of *blind beamforming* that coherently combines the signals on multiple antennas, without prior knowledge of target. In addition to receiving beamforming similar to [7,8], we take advantage of transmit beamforming gain through a role-switching mechanism.

(iii) RSS sampling and amplification mechanisms for emissive vibrometry. We design packet processing algorithms to extrapolate RSS values from legacy WiFi packets emitted by uncontrolled wireless devices. Consequently, we can reduce the RSS noise by orders of magnitude, thereby substantially amplifying the tiny radio RSS disturbance caused by the target's audio vibration.

The feasibility and effectiveness of the above approaches is verified through a software-radio based eavesdropper implementation, for both the reflective and emissive vibrometry. Through comprehensive experiments, we identify various practical factors, such as spatial separation, obstacle blockage, loudspeaker model, and environment dynamics, that determine the threat level. The results demonstrate that a basic wireless vibrometry setup, with eavesdropper placed close to the loudspeaker, can easily transform acoustic vibration into quality sounds. Using a combination of the diversity mechanisms, the eavesdropper can decode the loudspeaker's sound from 5 *m* apart, even with a sound-proof wall in between. In emissive vibrometry, the eavesdropper can decode the audio from unmodified WiFi devices with loudspeakers, at a much higher level of fidelity owing to high radio signal power emitted (instead of reflected) by the target.

These results pose alarming challenges to securing acoustic information. Using sound-proof walls is no longer effective, and even a radio shield may not work because the eavesdropper can attack from a wide range of radio frequencies. However, we show that several pragmatic counter-measures may make the eavesdropping significantly hard and thwart adversaries. In particular, we analytically provide a guideline of safety distance for a reflective victim and propose PHY-layer power randomization mechanisms for an emissive victim, which can reduce the eavesdropper's recovered audio quality by orders of magnitude. The impact of nearby human movement is also quatitatively evaluated through experiments. Slow and minor movement, *e.g.*, breathing, has negligible influence on the ART attack; Rapid larger-scale body movement, *e.g.*, walking, will considerably undermine the eavesdropping quality.

Contributions. In contrast to prior laser or vision based approaches [5, 9] to remote acoustic sensing, this work is the first to thoroughly investigate vibrometry on wireless devices, from both theoretical and experimental perspectives. We design a simple ART algorithm to decode remote acoustic vibration on loudspeakers by monitoring the reflective/emissive radio signals disturbed by the loudspeakers. Through an analytical framework and extensive experiments, we distill the key factors that enable highly sensitive WiFi vibrometry, and empower the vibrometry with a blind beamforming and frequency selection framework. The enhancement techniques share similar spirit with diversity mechanisms in wireless communications, but a wireless link can use known preambles to synchronize/train the transmitter/receiver channel. From a security/privacy perspective, we are the first to examine the factors that determine how threatening wireless vibrometry can be when applied in acoustic eavesdropping in practice, and to propose countermeasures.

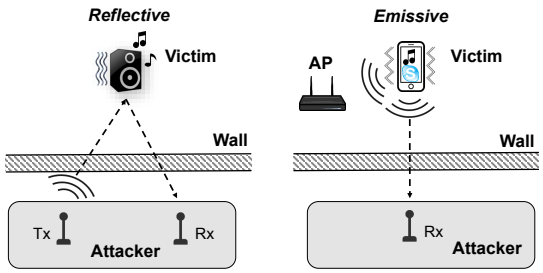

Figure 1: Two threat models based on wireless vibrometry: *Reflective and Emissive*.

2. WIRELESS VIBROMETRY: AN OVERVIEW

In this section, we brief the basic assumptions behind wireless vibrometry, given its two targeted scenarios (Figure 1). In general, both scenarios involve an eavesdropper (also referred to as attacker or adversary) and a target (or victim) loudspeaker. The target is assumed to be static. The eavesdropper controls radio transceivers that can run at a specific frequency band, with a practical transmit power level. The eavesdropping radios can be compromised devices within LOS of the target, or unfettered devices that are separated from the target by *sound-proof* obstacles (walls, windows, *etc.*). In any case, the eavesdropper can transmit/capture radio signals, but has no direct control over the target. However, as long as the eavesdropper's radio signals can reach the target, wireless vibrometry is feasible. Specifically, we explore a set of solutions, colloquially referred to as *audio-radio transformation* (ART), to enable wireless vibrometry.

(i) Basic ART (Sec. 3) directly decodes audio by processing the RSS/phase of radio signals disturbed by the loudspeaker. It is the basic decoding mechanism for both reflective and emissive vibrometry.

(ii) Enhanced reflective ART (Sec. 4) works in the reflective vibrometry scenario, where the adversary sends single-tone radio signals, and captures the signals reflected by the loudspeaker body. This enhanced mechanism harnesses diversity mechanisms, including multi-antenna blind beamforming and frequency selection to maximize the threat.

(iii) Enhanced emissive ART (Sec. 5) works in the emissive vibrometry scenario, where the target loudspeaker and a WiFi radio are co-located in the same device. The adversary captures the target's outgoing WiFi packets, extrapolates the RSS, and executes a set of amplification algorithms to enhance the decoded audio quality. Here the adversary has no knowledge about the target's WiFi radio identity and does not need to decode the actual content of the packets.

We will further propose countermeasures against ART in Sec. 6.

3. BASIC ART

In this section, we model wireless vibrometry as a process where audio vibrations modulate the radio signal magnitude/phase. Then we describe the basic acoustic-radio transformation (ART) algo-

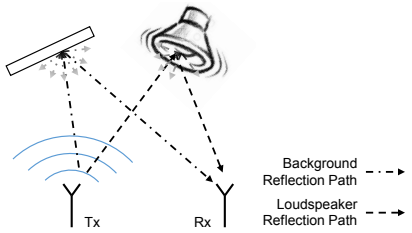

Figure 2: Illustration of loudspeaker modulation and multipath effects of reflective radio vibrometry.

rithm as the reverse demodulation process. We focus on reflective vibrometry, since the emissive vibrometry is a simplified special case. We further conduct experiments to evaluate the feasibility of ART, and identify the potential design knobs that warrant an enhanced ART.

3.1 Modeling Audio-Radio Frequency Translation

Let's consider a simple Line-of-Sight (LOS) scenario illustrated in Figure 2, where an adversary's radio Tx broadcasts wireless signals. Some of the signals undergo multipath reflections on the loudspeaker's surface. Part of the reflected signals will eventually be captured by the Rx. When the loudspeaker surface is vibrating, it will modulate the RSS and phase of reflected signal accordingly.

We first assume the loudspeaker is playing a mono sound with single frequency ω. The loudspeaker membrane and surface resonates and vibrates at the same frequency ω. Suppose the vibration displaces its surface by $d(t)$ at time t:

$$d(t) = k\sin(\omega t + \theta), \tag{1}$$

where k is the vibration magnitude, determined by the sound volume, and θ is the initial phase of sound. For simplicity, we omit the time index t of functions in later notations. Hence, $d \equiv d(t)$.

(a) Impact of vibration on radio RSS: We model the RSS of signals reflected from loudspeaker as:

$$RSS_L = \sigma A^2(d_0 + \hat{d}), \tag{2}$$

where $\hat{d} = d\cos\beta$, β being the angle between vibration direction and reflection direction. d_0 denotes the distance between antenna and loudspeaker. σ is the reflectivity (reflection gain) of loudspeaker surface ($0 < \sigma < 1$). $A(\cdot)$ is the channel gain function (equivalent to $1/$*pathloss*). The square operation over $A(\cdot)$ models the attenuation of signals due to round-trip propagation: one for path from Tx to loudspeaker and the other for path from loudspeaker to Rx. The loudspeaker surface is considered as a virtual transmitter that emits reflected signals.

For clarity, here we only consider the direct reflection path from loudspeaker. The analysis can easily incorporate the secondary-order reflection signals. The function $A(\cdot)$ can be assumed to follow any classical pathloss models in communications theory, *e.g.* free-space or log-normal [10]. Owing to continuity of the pathloss, we can expand Eq. (2) following Taylor's theorem:

$$RSS_L = \sigma[A^2(d_0) + 2A(d_0)A^{'}(d_0)\hat{d} + \cdots + o(d_0)\hat{d}^k], \tag{3}$$

in which $A^{'}(\cdot)$ is the first order derivative, and $o(d_0)$ is the Peano form of remainder.

The first term $A^2(d_0)$ is the DC component (carrier signals sent by the radio) and can be filtered out by a DC filter. The second term $2A(d_0)A^{'}(d_0)\hat{d}$ carries the sound frequency components, and $2A(d_0)A^{'}(d_0)$ can be considered as the *sensitivity*, or *gain* of audio-to-radio transformation. All remaining terms are the harmonic components of audio frequency. Eq. (3) implies that *the magnitude of harmonic frequencies is determined by the linearity of path loss function $A(\cdot)$ w.r.t. to distance*. The larger the path loss exponent, the more non-linear $A(\cdot)$ will be, and thus the stronger the harmonic becomes.

(b) Impact of vibration on radio signal phase: Assuming the adversary's Tx transmits a mono tone signal with frequency f_r, and up-modulated to carrier frequency f_0 (with wavelength λ_0). Then, the phase can be expressed as signal path length divided by wavelength:

$$Phase_L = \frac{2\pi(d_0 + 2\hat{d})}{\lambda_0} + \gamma \tag{4}$$

where γ is the initial phase of reflection path. The $\frac{4\pi\hat{d}}{\lambda_0}$ term contains the audio frequency from loudspeaker.

The above two sets of analysis imply that the *sensitivity of vibrometry is determined by the following factors*: *(i)* The vibration direction β relative to the reflection signal direction. The closer β is to 0, the stronger the vibrometry effect will be. *(ii)* The volume of the sound that determines the magnitude of displacement d, and hence the strength of vibrometry. *(iii)* The radio's carrier frequency f_0 or wavelength λ_0. In addition, for RSS based vibrometry, the strength is proportional to the reflectivity of the loudspeaker surface σ, and inversely proportional to the distance d_0 between radio adversary and the loudspeaker victim.

The emissive vibrometry model is similar, except that the signal source comes directly from the radio antenna vibrating together with the loudspeaker. We thus omit the details.

3.2 Demodulating the Transformed Audio

We proceed to design a basic ART that can demodulate the loudspeaker's audio out of the reflected radio waves. We focus on demodulation using the radio RSS (Eq. (2)), but the technique can be easily extended to phase based recovery.

Several challenges emerge here. First, the reflected signals become extremely weak as attacker-victim distance d_0 increases beyond a few meters. Second, the vibration magnitude d of the loudspeaker surface falls within sub-millimeter scale [11]. The variation it causes to the radio signal is orders of magnitude lower than the signal power itself. Third, the leakage signal directly coming from the adversary's Tx to Rx, and reflections from irrelevant background objects, can overwhelm the signals modulated by the loudspeaker alone.

Algorithm 1 Decoding the audio that is modulated by ART

1: **INPUT:** received radio samples $y[t]$
2: **OUTPUT:** recovered audio samples $d^*[s]$
3: /*Get one audio sample from every m radio samples*/
4: **foreach** segment s in set [0:S)
5: $\quad y_s \leftarrow y[sm+1, (s+1)m]$ /*Segment radio signals*/
6: $\quad z(v) \leftarrow \sum_{u=1}^{m} y_s(u) e^{\frac{-i2\pi vu}{m}}$ /*FFT analysis*/
7: $\quad g(s) \leftarrow |\frac{z(\lfloor \frac{f_r}{f_b} \times m \rfloor)}{m}|^2$ /*Pick RSS of CW's freq.*/
8: **endforeach**
9: $d^* \leftarrow filter_{bandpass}(g)$ /*Filter out the DC component*/

To address the challenges, we exploit a unique feature inside the audio-modulated radio signals, *i.e.*, the radio sampling rate is far beyond the audio frequency. Thus, we can enforce an averaging on the oversampled signals so as to amplify and promote the audio out of noise. In addition, we observe that the leakage and background signals are typically quasi-stationary compared to the audio vibration, and thus can be isolated from the ART signals through proper filtering. Algorithm 1 formalizes our basic decoder that incorporates these solutions. Below we detail the key steps.

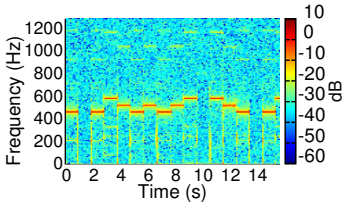

Figure 3: Decoding a sequence of piano sounds.

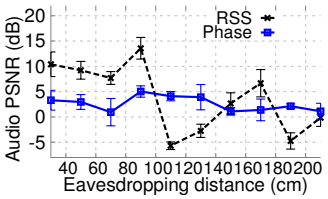

Figure 4: Audio PSNR over various distances to loudspeaker.

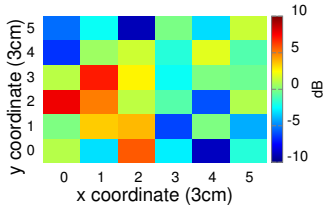

Figure 5: Audio PSNR over various locations of antenna.

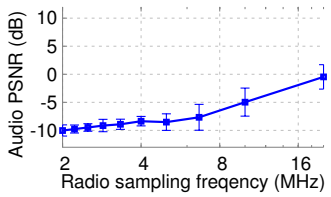

Figure 6: PSNR over different radio sampling frequencies.

Capturing audio-modulated radio samples: The eavesdropper transmits a baseband continuous wave (CW) with frequency f_r, comprised of T samples:

$$x(t) = \sin(2\pi f_r t), \quad t \in [0, T). \tag{5}$$

Meanwhile, the loudspeaker vibrates following Eq. (1). Based on the RSS modulation model in Eq. (3), we approximate the received radio samples as follows,

$$y(t) = \sqrt{RSS_L(t) + RSS_e}\, x(t) + n(t), t \in [0, T) \tag{6}$$

where $RSS_L(t)$ and RSS_e are signal strengths reflected from loudspeaker and relatively stationary objects in the environment. $n(t)$ is the noise. For clarity, we have omitted the phase term of signal components from multipaths.

Segmenting the samples: We divide all T radio samples into S segments, each containing m samples. Since audio vibrates at a much lower rate, audio signals within any small segment of radio samples can be considered stationary. Thus, we have $RSS_L(sm) \approx RSS_L(sm+u), \forall u \in [0, m)$ and $s \in [0, S)$. For the s^{th} segment y_s, we can approximate the *u-th* sample inside as:

$$y_s(u) \approx \sqrt{RSS_L(s) + RSS_e}\, x(sm+u) + n_s(u), \tag{7}$$

where $n_s(u) \triangleq n(sm+u)$. Samples in each segment are regarded as single-tone radio signals modulated by an audio of constant amplitude, plus a stationary component and noise.

Signal processing – time-frequency domain translation: To extract the amplitude modulated on top of radio signal, we apply the discrete Fourier transform (DFT) on each segment y_s. The resulting sequence of DFT coefficients is given by:

$$z(v) = \sum_{u=1}^{m} y_s(u) e^{-i2\pi vu/m}, \qquad v \in [0, m). \tag{8}$$

Since the transmitted signals is a single-tone of frequency f_r, we only need to inspect the coefficient of that frequency bin, and extrapolate the audio signal strength within this radio segment:

$$g(s) = |\frac{z(v^*)}{m}|^2 = RSS_L(s) + RSS_e \tag{9}$$

where $v^* = \left\lfloor \frac{f_r}{f_b} \times m \right\rfloor$ denotes the frequency-bin index of f_r, and f_b is the radio receiver's sampling rate.

Repeating the procedure above on all radio samples produces audio signals sampled at rate $\frac{f_b}{m}$. Clearly, a higher radio sampling frequency means a lager audio sampling frequency, which will lead to higher recovered audio quality.

Bandpass filter: The series of $g(s), s \in [0, S)$ still contains the DC component, which originates from background reflections, *e.g.*, the term RSS_e (Eq. (9)), and the term $\sigma A^2(d_0)$ in RSS_L (Eq. (3)). DC component can be eliminated via a bandpass filter. We empirically choose the lower and upper stopping frequencies as 20 Hz and 1500 Hz, which correspond to the range of human voice.

3.3 Feasibility of Basic ART

We have implemented the basic ART demodulator on a software-radio testbed with a custom-built FPGA core (see Sec. 7 for details). In this section, we conduct testbed experiments to verify the key factors that govern the performance of ART, and explore opportunities that warrant improvement.

Our basic experimental setup complies with the foregoing model assumptions. The adversary radio emits a single tone with 5 MHz baseband frequency, carrier-modulated to 2.485 GHz (WiFi channel 14). The target is a PC loudspeaker 2 *m* away. Figure 3 showcases a time-frequency plot of a piano sound decoded using the basic ART. The piano sound has three piano notes of frequencies 440Hz, 493.88Hz and 554.365Hz. We observe a close-to-perfect recovery and the decoded audio is clearly audible. In addition, the result also shows weaker harmonics of audio frequencies (*e.g.*, 880Hz), as predicted by our model.

Audio quality vs. distance: We now vary the distance between the eavesdropper and loudspeaker at steps of 20 *cm*. Peak-Signal-to-Noise-Ratio (PSNR) is a commonly used metric to quantify the decoded audio quality [9]. Note that for single-frequency audio, PSNR above -13dB is typically audible to human, while for multi-frequency audio like human speech, PSNR above 0dB is audible [12]. Here the loudspeaker is forced to play a 400 Hz tone, without loss of generality. For each location, we repeat the experiment 10 times, and plot the mean PSNR in Figure 4.

We make the following observations: *(i)* Either phase or RSS can be used to decode the audio. Moreover, performance of the RSS and phase based decoders does not show clear correlation. However, the highest audio PSNR from RSS is higher than that from phase. *(ii)* There does not exist a monotonic trend between PSNR and range. We suspect this is due to multi-path effects, which deserves a more in-depth investigation.

Audio quality variation over space: We highlight the multipath effect by varying the adversary radio location within a small 15*cm* $\times$ 15*cm* region, sampled at 3 cm granularity. Figure 5 shows that *the decoded audio quality is highly sensitive to adversary location*: even with a minor location change of 3 *cm*, the PSNR can differ by up to 10 dB. We will model the root cause and propose an enhanced ART algorithm that exploit this effect in Sec. 4.

Audio quality vs. radio sampling rate: Figure 6 shows that the audio PSNR improves proportionally with radio sampling frequency f_b. Hence, *a wideband eavesdropper presents a higher level of threat*. The frequency of human speech sounds typically falls below 500Hz [13]. To recover human speech, the audio sampling frequency should be at least twofold of the audio's frequency , according to Nyquist sampling theorem. A common WiFi device with 20Mbps bandwidth is good enough for eavesdropping human speech, since $f_a = \frac{20e6}{1024} \approx 19.5\text{KHz} \gg 1\text{KHz}$, when we use a typical segment length $m = 1024$.

4. ENHANCING REFLECTIVE AUDIO EAVESDROPPING

The above analysis/experiments hint that the performance of ART is closely related with radio signal/channel diversity, which we now explore to enable a more powerful ART for reflective vibrometry.

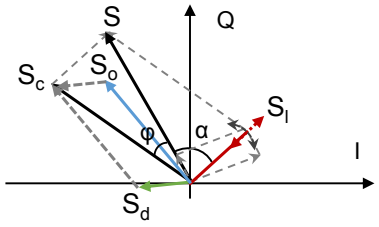

Figure 7: Model of multipath effects in baseband received signals. S_l denotes reflection path from loudspeaker. S_o and S_d represents signal from other objects and LOS path.

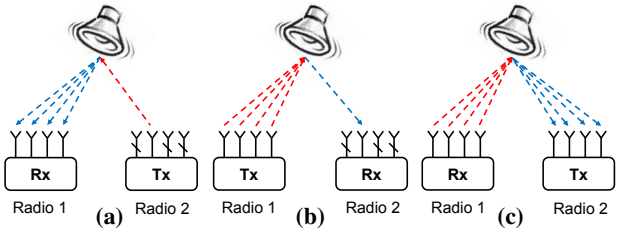

Figure 8: Three stages of role-switching beamforming algorithm. (a) Rx beam-searching (b) Rx-guided Tx beamforming (c) Tx-Rx beamforming.

4.1 Modeling Multipath Effects in ART

We first extend the analysis of basic ART to incorporate multipath reflection, which will become the foundation for the enhanced ART. As illustrated in Figure 2, the received signals S contain LOS component S_d between adversary Tx and Rx, reflection components S_l from loudspeaker and irrelevant component S_o from other objects. Since the transmitted signal is a single-frequency tone, these components can be analyzed on an I-Q plane (Figure 7):

$$\overrightarrow{S} = \overrightarrow{S_c} + \overrightarrow{S_l}, \text{ and } \overrightarrow{S_c} = \overrightarrow{S_o} + \overrightarrow{S_d}. \tag{10}$$

Note that reflection component S_l is a time-varying vector due to the modulation effect from audio vibration, thus the magnitude of S_l vibrates. We represent this strength vibration using the arrow along S_l in the figure. Similarly, the tangent arrow represents the phase vibration of S_l caused by audio vibration.

We again focus on signal magnitude without loss of generality. From wireless channel prospective, reflected signal from loudspeaker and other n paths can be formalized as:

$$S_l = h_0 x + n_1, \text{ and } S_c = \sum_{i=1}^{n} h_i x + n_2 = \boldsymbol{h} x + n_2, \tag{11}$$

where $|h_0|^2 = \sigma A^2$, σ and A are the material reflectivity and path loss model discussed in Section 3.1. x is the transmitted symbol. h_i is the channel coefficient factor, and n_i denotes the noise in channel. In classical channel models [10], $\boldsymbol{h}$ and n_i can be modeled as complex Gaussian, *i.e.*, $\boldsymbol{h} \sim CN(0, N_h)$ and $n_i \sim CN(0, N_0)$.

From Figure 7, we can derive the radio RSS by simple geometry:

$$|S| = |S_l| \cos\alpha + |S_c| \cos\varphi, \tag{12}$$

where α (φ) is the angle between S and S_l (S_c). In our ART decoder (Section 3.2), m radio samples are used to extrapolate one audio sample, and audio samples pass through a DC filter. Hence, the recovered audio PSNR is determined by the ratio of non-DC part of S_l over S_c, *i.e.*:

$$\begin{aligned} PSNR &= \frac{mpE[|S_l|^2]\cos^2\alpha}{E[|S_c|^2]\cos^2\varphi + N_0} \\ &\approx \frac{2mp\sigma A(d_0)A^{'}(d_0)k\cos\beta|x|^2\cos^2\alpha}{N_h|x|^2\cos^2\varphi + N_0}, \end{aligned} \tag{13}$$

where the approximation omits high-order harmonic terms. p denotes factor of energy loss due to NLOS eavesdropping, *e.g.* penetrating the wall.

The analysis unveils the following insights:

(i) The PSNR is determined by α and φ, which depend on the relative arriving angle between signals reflected by loudspeaker and other objects, which in turn depend on the adversary's location and frequency (wavelength). Thus, the PSNR can be improved by judiciously selecting position or radio frequency. This insight also corroborates the foregoing feasibility study: *the PSNR fluctuation is mainly attributed to the multipath coherence combining effect, rather than the minor distance changes between adversary and loudspeaker*.

(ii) The transmit symbol term x appears in both nominator and denominator, implying that *the adversary cannot achieve an arbitrarily high PSNR by increasing its radio signal power*. However, it can improve by reducing the term $N_h|x|^2\cos^2\varphi$. More specifically, this can be achieved by adopting wireless beamforming technology, which can focus the energy on a certain spot, *e.g.*, the loudspeaker surface.

4.2 Harnessing Diversity

We now exploit diversity techniques to enhance the reflective ART. Diversity has been well studied in communications theory [10]. However, we will show that it takes new facets in ART, particularly because the adversary has no control over the target loudspeaker.

4.2.1 Blind Beamforming to Amplify Reflective ART

Based on the prior analytical insights, we explore multi-antenna beamforming to enhance the reflective ART. Unlike traditional beamforming in wireless communications [10], which enforces channel training between active Tx/Rx radios using known preambles, here we need to *amplify signals from a passive loudspeaker with unknown channel signature*. One may consider existing blind beamforming algorithms [7, 10] that search and assign different weights to the Tx/Rx antennas to maximize SNR. However, due to slow time scale of audio playback, the accumulated time of such trials is formidable and thwarts any in situ eavesdropping attack. Moreover, previous algorithms [7] can have a cooperative target (*i.e.*, a target human performs special gestures) to estimate channel preamble, but in eavesdropping case the eavesdropper doesn't have any control on the loudspeaker.

We propose a *role-switching beamforming* (RSBF) algorithm to address these challenges. From a high level, RSBF comprises three steps, as illustrated in Figure 8: *(i)* One-time test transmission (lasting for a few seconds, just enough to obtain one audio segment). Any one antenna, *e.g.*, 2^{nd}, of radio 2 acts as transmitter. All antennas in radio 1 receive the data and run an *Rx beamforming* (RXBF) algorithm to compute their beamforming weights. *(ii)* Rx-guided Tx beamforming. Antennas of radio 1 act as transmitter and beamform to one antenna of radio 2. The Tx beamforming weights are directly derived from their Rx weights, with no extra runtime overhead. *(iii)* Tx-Rx beamforming, where antennas of radio 2 compute new Rx beamforming weights under Tx beamforming from radio 1. Algorithm 2 summarizes the RSBF procedure. Below we elaborate on the design.

Calibrating RF Heterogeneity. Due to hardware imperfection, antennas and RF chains may have different initial phase offsets and gains. Heterogeneity calibration first measures the normalized distortion of different RF ports *w.r.t.* the first one. All calibrating an-

Algorithm 2 Role-switching Beamforming

/*RF Heterogeneity Calibration*/
$\Gamma_r[n] = y_r[n]/y_r[1]$
$\Gamma_t[n] = y_t[n]/y_t[1]$
$\mathbf{W} = Rx\_BF(\mathbf{D}[1:N])$ /*Rx beamforming*/
$\mathbf{W_t} = \frac{\mathbf{W}\Gamma_\mathbf{r}}{\Gamma_\mathbf{t}}$ /*Rx-guided Tx weights*/
/*New Rx weights under Tx beamforming*/
$\mathbf{W_r} = Rx\_BF(\mathbf{D_T}[1:N])$

tennas can be placed on a circle around one transmit antenna. Since they have the same distance to transmit antenna, their received signals $\mathbf{Y}$ will experience similar distortion H. The difference of received signals are mainly caused by antenna heterogeneity $\Gamma_\mathbf{r}$:

$$\mathbf{Y} = H\Gamma_\mathbf{r}X,$$

where $\mathbf{Y} = [y_1, y_2, \ldots, y_N]$ and $\Gamma_\mathbf{r} = [f_{r1}, f_{r2}, \ldots, f_{rN}]$ are 1×N matrixs, and X is the transmit symbol. The normalized distortion $\overline{\Gamma_\mathbf{r}}$ is then:

$$\overline{\Gamma_\mathbf{r}} = [1, \frac{f_{r2}}{f_{r1}}, \ldots, \frac{f_{rn}}{f_{r1}}] = [1, \frac{y_2}{y_1}, \ldots, \frac{y_n}{y_1}] = \overline{\mathbf{Y}},$$

where $\overline{\mathbf{Y}}$ is the normalized received signals *w.r.t.* to the first antenna. In a similar way, we can estimate the transmit RF heterogeneity. $\overline{\Gamma_\mathbf{t}}$

Rx Beamforming without Channel Training. The objective of the Rx beamforming algorithm is to find the weights $\mathbf{W} = \{w[i] | 1 \leq i \leq N\}$ for radio signals arriving at different antennas, in order to maximize the audio PSNR:

$$\arg\max_{\mathbf{W}} \ \mathit{PSNR}(\sum_{i=1}^{N} w[i]\mathbf{D}[\mathbf{i}]), \tag{14}$$

where $\mathbf{D}[\mathbf{i}]$ is the signals at i^{th} antenna. Directly searching for the optimal weights is computationally inefficient as the search space lies in the continuous complex domain. We thus propose the following approximation. First, the amplitude and phase are discretized into range $[A_l : A_h]$ and $[P_l : P_h]$ with step A_s and P_s respectively. We choose $A_l = 0.5$ and $A_h = 2$, which cover $\pm 3dB$ magnitude range, and $P_l = 0$ $P_h = 2\pi$ to cover the whole phase range. Empirically, $A_s = 0.05$ and $P_s = 0.1$ is sufficient in terms of granularity.

Our Rx beamforming algorithm recursively computes the optimal complex weight for the i^{th} antenna *w.r.t.* to previous $i-1$ antennas. It ensures the weight for a new antenna will not degrade the audio PSNR when co-working with existing antennas. To avoid trapping into local maxima, we randomize the antenna ordering. Our algorithm differs from existing beamforming weight searching [7] that computes the weights for $N-1$ antennas *independently w.r.t.* to the first one, which cannot prevent audio PSNR from degrading when antennas are co-working.

Note that computation of the optimal complex weights is an offline procedure. In other words, we assess the beamforming weights by collecting and then processing a single short time series of radio signals, rather than real-time over-the-air transmission and processing. To compare the effects of different weight selections, we first use existing voice detection methods to identify if there is sound of interest in the decoded audio. Then we recurse on the weight selection with higher PSNR.

Rx-guided Tx Beamforming. The basic idea behind Rx-guided Tx beamforming is to leverage the channel reciprocity, *i.e.*, wireless link will distort signals in a similar way by reversing the role of Tx and Rx. Hence, Rx beamforming weights which focus on signals coming from certain position can also concentrate the energy onto the same spot when used as Tx weights. Mathematically, the rationale can be expressed by: $\mathbf{W_r}H = (H^T\mathbf{W_r}^T)^T$, where $(.)^T$ denotes the transpose of matrix. Left (right) part of the equation show the case when weights are used at receiver (transmitter) side. However, RF heterogeneity must be compensated before applying the weights:

$$\mathbf{W_t} = \mathbf{W_r}^T\Gamma_\mathbf{r}/\Gamma_\mathbf{t}, \tag{15}$$

where $\mathbf{W_r}$ represents Rx weights. The guided Tx (radio 1) beamforms to one of antennas at Rx (radio 2). To fully leverage the beamforming power, Rx runs another round of RXBF algorithm to maximize the audio PSNR.

Algorithm 3 Rx_BF – "Offline"Rx Beamforming

INPUT: received data stream $\mathbf{D}[i]$
OUTPUT: weights $\mathbf{W}$
$\mathbf{cur}_D \leftarrow \mathbf{D}[1]$, $m\_PSNR \leftarrow 0$ /*Initialize*/
foreach antenna i in set [2:N]
  $m_\psi \leftarrow 0$, $m\_\rho \leftarrow 0$
  foreach ψ in range [P_l:P_s:P_h] /*Search phase*/
    $\mathbf{T}_D \leftarrow \mathbf{cur}_D + \exp(1i * \psi)\mathbf{D}[i]$
    if $PSNR(\mathbf{T}_D) > m\_PSNR$
      $m\_PSNR \leftarrow PSNR(\mathbf{T}_D)$, $m\_\psi \leftarrow \psi$
  foreach ρ in range [A_l:A_s:A_h] /*Search amplitude*/
    $\mathbf{T}_D \leftarrow \mathbf{cur}_D + \rho * \exp(1i * m\_\psi)\mathbf{D}[i]$
    if $PSNR(\mathbf{T}_D) > m\_PSNR$
      $m\_PSNR \leftarrow PSNR(\mathbf{T}_D)$, $m\_\rho \leftarrow \rho$
  $W[i] \leftarrow m\_\rho * \exp(1i * m\_\psi)$
  $\mathbf{cur}_D \leftarrow \mathbf{cur}_D + W[i]\mathbf{D}[i]$
return $\mathbf{W}$

4.2.2 *Amplifying Multipath Effect via Frequency Selection*

Besides leveraging spatial diversity of beamforming, we can also harness the *frequency diversity* — a radio signal and its reflected version may either strengthen or weaken each other, depending on its frequency (or wavelength). This in turn affects the audio quality of ART decoding.

Typical wireless systems (*e.g.*, WiFi) adopt a two-level channelization: a transceiver first selects a residential band at a certain carrier frequency, and then splits the band into multiple bins called *subcarriers*. We examine the channelization impact on ART by fixing the adversary location. Figure 9 plots the PSNR of recovered audio when adversary sends an 1 MHz baseband sine-tone through WiFi channel 1 to 14 (corresponding to carrier frequency from 2.412 GHz to 2.485 GHz). Figure 10 further plots the PSNR when the sine-tone is sent through different subcarriers within one WiFi channel.

The figures show that, owing to wider frequency separation, *channel-level frequency diversity has much more significant impact than subcarrier-level diversity*. The former typically creates 20 dB of PSNR discrepancy among channels, versus 10 dB in the latter. Another observation is that higher radio channel gain does not always mean higher audio quality, which is consistent with our theoretical prediction.

Based on these findings, we design a two-level channel selection mechanism to foster the multipath strengthening effect. Specifically, when initiating eavesdropping, the adversary spies on each channel for a few seconds, and fixes on the one with highest PSNR. Then, within that channel it transmit the single-tone through the subcarrier with highest channel gain. Note that the adversary can estimate the channel gain of all subcarriers simultaneously by sending a single OFDM packet with a training preamble.

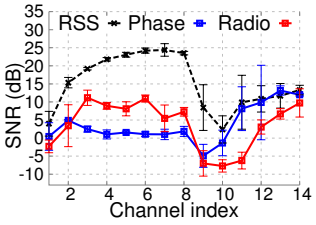

Figure 9: Audio and radio SNR over channel.

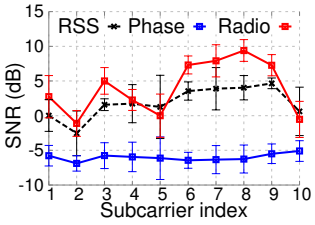

Figure 10: Audio and radio SNR over subcarrier.

5. EMISSIVE AUDIO EAVESDROPPING

Unlike the reflective case, an adversary in emissive vibrometry needs to decode audio by monitoring the RSS vibration of WiFi packets from the target. These packets are modulated across a wide spectrum, arrive randomly, and are beyond the control of the adversary. Our basic solution follows the same principle as the basic RSS-based ART (Sec. 3), but incorporates mechanisms to extract and amplify RSS values from the legacy WiFi packets.

Packet Detection: Following the 802.11 standard, each WiFi packet starts with a known preamble called Short Training Field (STF), which has a repetitive pattern in time domain. Hence, we follow the typical approach in [14], which can detect such periodic preambles using a self-correlation algorithm. We further use an 802.11 OFDM demodulator [15] to decode the unencrypted MAC header which conveys the data symbol modulation scheme (*e.g.*, BPSK).

RSS Estimation and Amplification: Suppose the target's MAC address is known to the adversary (we will discuss about how to relax this assumption). The simplest way to extract RSS is to single out target's WiFi packets, and extrapolate the time-domain signal power level across each packet to obtain one RSS value. However, we found this approach only leads to around -8 dB of PSNR even if the eavesdropper is 5 *cm* away from the target. The underlying reason is the high peak-to-average power ratio (PAPR) in 802.11's OFDM modulated packets, which renders a simple RSS averaging highly inaccurate.

To solve this problem, we first demodulate the OFDM-modulated packet into low-level data symbols. Let's first consider a simple case with BPSK or QPSK modulated data symbols, which all have the same magnitude in the constellation diagram. We can estimate the RSS by simply computing the average power over all N data symbols inside the packet:

$$RSS = \frac{\sum_{i=1}^{N} |s[i]|^2}{N}, \tag{16}$$

which circumvents the high PAPR. In general, data symbols may not have the same magnitude when using higher modulation order, *i.e.* 16QAM and 64QAM. We need to normalized the received symbols according to their constellation magnitude before computing the average power.

A typical WiFi packet has an N-value of thousand scale (*e.g.*, $N = 2048$ for a 512-byte QPSK modulated packet). Our RSS averaging method essentially reduces the noise level by N, thus potentially boosting the PSNR of ART by orders of magnitude. We thus refer to this approach as *RSS amplification*.

Reinterpolation of Non-uniform Packet Arrival: Since 802.11 MAC adopts CSMA, the timing gap between two consecutive packets is highly random. Thus, the RSS directly obtained through the above approach has non-uniform intervals. However, basic ART only works under uniform RSS samples. Our solution is to reinterpolate the audio signals based on non-uniform sampling theory [16], which essentially remaps the RSS samples to a grid of uniform intervals.

Params	Value	Params	Value	Params	Value
p	-8dB	N_0	-90dBm	β	0
m	4096	d	1mm	α	0
σ	1	$\|x\|^2$	100dBm	φ	90

Table 1: Parameters of counter measures for reflective eavesdropping.

Under non-uniform sampling, the effective sampling rate is determined by the maximum gap between samples. To sample the audio from the target with frequency of 500Hz (for human speech), we thus need a packet arrival rate of 1000 pkt/s. This rate can be easily achieved in modern WiFi protocols and applications.

Three points are worth further discussion regarding emissive ART.

(i) We have assumed the adversary knows the target's MAC address. In practice, the adversary can simply attempt to overhear packets from each MAC address nearby and run ART over it. The one resulting in a high PSNR and meaningful speech can be identified as the target.

(ii) Beamforming is still applicable as an amplification mechanism. However, in contrary to the reflective case, where beamforming aims to amplify one of the multipaths, in the emissive case, the beamforming objective is to amplify all signals sent from the target device. Hence, this problem can be reduced to traditional receiver beamforming [10]. We omit more details for the sake of space.

(iii) The emissive eavesdropping quality depends on the packet rate of the victim's smartphone. A very slow packet rate *e.g.*, 200 pkt/s as evaluated in Section 8, cannot provide sufficient sampling rate to recover the audio. However, this problem can be alleviated by applying techniques such as ARP attack [17] and RTS injection [18] that force "silent" idling WiFi device to "talk". We leave more exploration of such approaches as future work

6. COUNTER MEASURES

In this section, we propose counter measures that can thwart reflective/emissive eavesdropping in practice.

6.1 Combating Reflective ART

Distance between adversary and victim is a critical factor that determines the capability of reflective ART. Following an empirical analysis, we can derive the *safety distance* d_{safe}, *i.e.*, the upper-bound distance that an adversary can sustain to launch an attack.

Our analysis extends the model in Sec. 3.1. We use free-space propagation [10] as our pathloss model: $A(d_0)^2 = \frac{\lambda^2}{(4\pi)^2 d_0^2 L}$. Real world wireless signals typically undergo stronger attenuation than free-space model, and thus have a shorter d_{safe}. λ is the wavelength, *i.e.* 12.5cm for 2.4 GHz. $L(L \geq 1)$ is the system loss, which is set as 1, *i.e.* no loss, in our analysis. Transmit/receive antenna gains G_t and G_r are set to typical values for high-gain WiFi antennas [19]: $G_t = G_r = 12dB$. We assume the adversary needs to penetrate through drywall, with 8 dB of round-trip loss: p = -8dB [20]. Other parameters are optimistic values for the adversary and summarized in Table 1.

Based on above settings and our model in Eq. (13), we can derive the relation between d_{safe} and audio PSNR, which is plotted in Figure 11. Free-space model (pathloss exponent = 2) requires d_{safe} to be larger than 26.5m so that audio PSNR drops below -13 dB, *i.e.* the PSNR level below which even a single frequency becomes inaudible (page 5 of [12]). For practical multipath channels with higher pathloss (*e.g.* office environment with path loss exponent

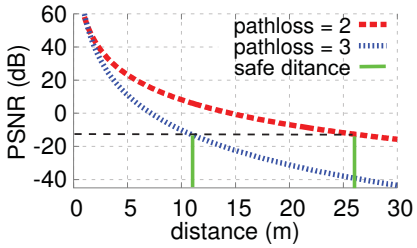

Figure 11: Safety distance vs. Audio PSNR. (Audios with PSNR lower than -13 dB cannot be detected).

3 [10]), safety distance can be reduced to 11m. It is worthy noting that the derived d_{safe} is conservative, which assumes optimistic parameters for adversary. In practice, the distance an adversary needs to recover recognizable audio has to be much smaller than d_{safe}.

Besides the safety distance, many other mechanical vibrations, *e.g.*, human activity, that have overlapping vibration frequencies with human sound can act as *interfering signals* to reduce the eavesdropping PSNR. However, it is worth noting that human movement does not *guarantee* keeping the victim from being eavesdropped. Part of the audio may still be recovered if victim temporarily slows down the movement. For minor human movements, *e.g.*, breathing, audio can still be recovered because they mainly cause low-frequency vibrations, and thus can be filtered by a bandpass filter. On the contrary, it is hard to recover audio signals when rapid movements, such as walking, is taking place nearby. We quantitatively evaluate the impact of human movement in Section 8. Overall, although human activity cannot completely thwart the ART attack, it is still a good approach to undermine the eavesdropping quality.

6.2 Combating Emissive ART

For the emissive case, the victim has a better control of the source of information leakage, and they can combat ART using more active mechanisms. Given this advantage, we propose a PHY layer countermeasure mechanism called *Transmission Power Randomization (TPR)*, which randomizes the transmission power across different radio packets transmitted by itself, so as to disturbs the RSS variation from acoustic vibration. Specifically, suppose the t-th packet has a legitimate transmit power of $P(t)$ (typically remaining as a constant in WiFi [21]). We randomize its power as $f(t) \cdot P(t)$, where $f(t)$ follows a normal distribution $N(1, \delta^2)$. The randomization intensity depends on the standard deviation δ, and determines the effectiveness of this countermeasure. δ should be kept small, so as not to affect the normal MAC/PHY protocol. However, we find that even a small value of $\delta = 0.05$ can already reduce the eavesdropper's PSNR by several orders of magnitude, as will be shown in our experiments (Sec. 8.3),

Similar rationale may be applied to thwart phase-based emissive ART. However, practical radio front-end already has a random initial phase offset for each packet, which naturally thwarts phased-based emissive ART.

7. IMPLEMENTATION

We prototype the eavesdropper based on the WARP software defined radio (SDR) [22]. The WARP radios are controlled by a laptop PC which implements the signal processing mechanisms that constitute basic/reflective/emissive ART (Sec. 3, 4 and 5). The radio transmitter directly takes baseband signals from the PC in the format of discrete I/Q samples, and then sends them through any WiFi channel. The receiver captures signals, converts them to baseband, and passes to the PC for processing. In addition, each WARP radio uses up to 4 antennas to run the RSBF algorithm in Sec. 4.2.

Our implementation of reflective ART sends/receives customized single-tone signals at 2.485 GHz, following the algorithms in Sec. 3 and Sec. 4. To realize the emissive ART, we have implemented an 802.11g/n-compliant OFDM communication library. Its receiver module consists of packet detection, synchronization, frequency offset compensation, and OFDM/symbol demodulation functions. It can directly decode OFDM symbols from legacy 802.11 packets, and then process the symbols using the RSS extraction and amplification module (Section 5).

Besides the 2.4 GHz WiFi channels, we also pair each WARP radio with WURC [23], a third-party RF front-end that enables transmitting/receiving signals over the UHF band. The UHF band experiments are conducted under an FCC experimental license, which allows us to use the vacant TV channels 40 and 41 (626 – 638 MHz) in our area.

Due to the interface latency between the PC and WARP [24], current WARP driver only allows transmitting/receiving bursts of I/Q samples at 3000 bursts/second, each burst containing 2^{10} samples. This translates to only up to 3 KHz of audio sampling rate (Sec. 5), and further reduces by 4 folds when using 4 beamforming antennas. We overcome the limitation by retrofitting the WARP FPGA kernel, and implementing a continuous streaming mode, which leverages WARP's internal memory to store extra data samples. This enables continuously writing/reading of up to 2^{28} I/Q samples, and the audio sampling rate can be as high as $\frac{20e6}{2^{10}} \approx 19.5$ KHz. The maximum time duration of a single recording can be $\frac{2^{28}}{20e6} \approx 13.4s$.

8. EXPERIMENTAL VALIDATION

Based on the prototype implementation, we evaluate reflective and emissive vibrometry, each following two steps. First, we conduct micro-benchmark validation of different enhancement mechanisms, separately. Second, we combine these mechanisms and perform blind-tests to demonstrate the overall eavesdropping threats under various practical scenarios. Finally, we validate the proposed countermeasures. Unless noted otherwise, there are working people inside the room, sitting roughly 2m away from the antennas and loudspeaker.

8.1 Validating Reflective Eavesdropping

8.1.1 Micro-benchmark of Diversity Mechanisms

We first verify the effectiveness of the diversity mechanisms in Sec. 4.2. By default, the loudspeaker plays a single-tone 400 Hz audio. Besides LOS, we create NLOS test case by separating the adversary and loudspeaker using an office partition. For each test case, we try 10 different eavesdropper locations 1 *m* away from the loudspeaker, and then measure the decoded PSNR.

Impact of blind beamforming: Recall the blind beamforming comprises 3 steps (Sec. 4.2.1): *(i)* Rx beamforming; *(ii)* Tx beamforming guided by Rx weights; *(iii)* TxRx beamforming. We first investigate the Rx beamforming performance as the number of antennas increases. From the experiments (Figure 12), we observe obvious improvement when the number of antennas increases from 1 to 2, but the gain saturates quickly. This is consistent with the log scale beamforming gain in traditional communications [10]. It also implies that the adversary can harness the majority of beamforming gain even with a small number of antennas.

Figure 13 depicts the eventual PSNR after the 3-step beamforming described in Sec. 4.2.1. we have two observations. First, the effect of beamforming is location-dependent: the beamforming gain ranges from 3 to 13.5 dB across the 10 locations. This is a direct

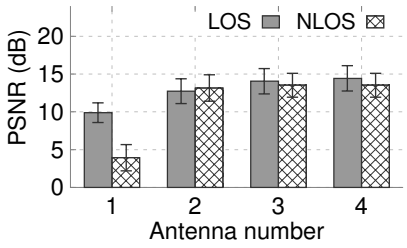

Figure 12: Audio quality gain vs. number of Rx beamforming antennas.

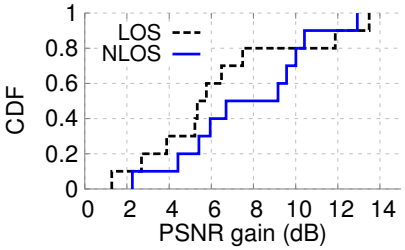

Figure 13: CDF of PSNR improvement of 4Tx-to-4Rx beamforming.

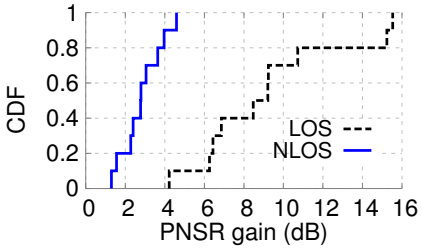

Figure 14: CDF of improvement by channel selection.

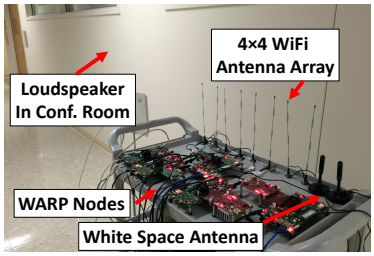

Figure 15: ART hardware platform. Testing ART outside a conference room.

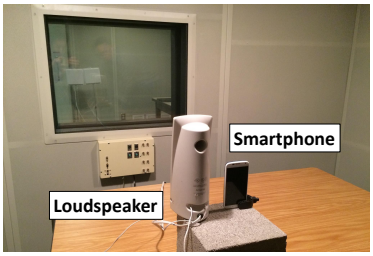

Figure 16: Testing ART performance. Loudspeaker is inside a soundproof room.

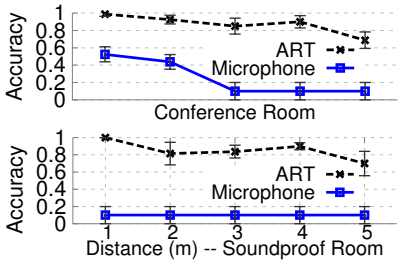

Figure 17: Through-wall recognition accuracy of ART compared with a microphone.

consequence of the location-dependent property in single-antenna ART (Sec. 3.3): when the multiple antennas all have strong PSNR and are uncorrelated, beamforming leads to highest amplification.

Secondly, there are more substantial benefits of beamforming at NLOS scenarios. On average, the NLOS beamforming gain is 8 dB across 10 locations, in contrast to 5.6 dB in LOS. This is because in NLOS, multipath diversity becomes dominant, leading to highly uncorrelated channel across antennas, which boosts beamforming gain.

Impact of channel selection: We run the two-level channel selection mechanism (Sec. 4.2) for each of the adversary location and plot the CDF of PSNR gain in Figure 14. As expected, channel selection plays a significant role in improving eavesdropping performance. We also observe that the average improvement ratio under LOS (*i.e.*, 60% and 9.5dB) is larger than that under NLOS (*i.e.*, 40% and 2.9dB), which is different from the beamforming case. Still, the reason is that when multipath diversity becomes dominant in NLOS, all channels degrade significantly. Moreover, the better channels under LOS have a larger deterioration since their dominating LOS paths with stronger signal are blocked. In consequence, channel selection yields a lower gain in NLOS.

8.1.2 Overall Eavesdropping Threat Analysis.

We now combine all the diversity mechanisms, and evaluate the overall eavesdropping threat under realistic settings.

Through-wall eavesdropping threats: We perform experiments in two closed rooms. The first is a typical conference room with drywalls all around, located in our office building (Figure 15). The second is a specialized soundproof room for conducting children behavior research (Figure 16). In both scenarios, we place a loudspeaker (for the reflective cases) and a smartphone (for the emissive cases) inside the room. The eavesdropper Tx/Rx are deployed outside.

To quantify the eavesdropping threat, we set up a blind test to evaluate the quality of recovered audio in comparison to a iPhone microphone. Specifically, the loudspeaker plays human speech sounds that pronounce random numbers between "zero" to "nine". Meanwhile, the microphone is placed at the same location as the eavesdropper and records the sound. We permute the numbers in each experiment, and invite 6 users to listen and transcribe the numbers recorded by the microphone and diversity-enhanced ART, respectively. We use the percentage of correctly-transcribed numbers as evaluation metric, referred to as *recognition accuracy*.

Figure 17 plots the recognition accuracy as we increase the distance between eavesdropper radios and the loudspeaker (with wall in between). Human ear can recognize the sound recovered by ART with almost 100% accuracy when the distance is less than 1 *m* outside conference room. Moreover, ART can keep a high recognition accuracy of more than 80% for up to 4 *m*. In contrast, the microphone recorder can only recover around 50% even below 2 *m* distance. Beyond that, the microphone cannot record any sound from the loudspeaker. For the purpose of experimental contrast, we cap the accuracy to 10% by assuming all testers randomly guess one out of ten numbers. For the sound-proof room, the experiment results are more interesting: while the microphone cannot capture any sound at all, ART can easily penetrate the sound-proof obstruction and still achieve high recognition accuracy.

To sum up, *an ART-enabled eavesdropper can indeed penetrate conventional sound isolators like walls and sound-proof windows.* Naturally, the attack may fail if the isolator (*e.g.*, metal walls) blocks radio signals completely. However, the above experiment already demonstrates alarming threat in the real-world.

From our extensive experiments, we also establish empirical relation between PSNR and recognition accuracy. When PSNR is above 15dB, testers typically can recognize words with close to 100% accuracy. When PSNR$>$ 10dB, most words can be recognized ($>$60% accuracy). The testers fail to recognize any words when PSNR falls below 6dB. In what follows, we continue with a more microscopic examination of other factors that may affect the threat level.

Eavesdropping using different carrier frequencies: It is known that lower-frequency radio signals experience less propagation loss and penetration loss through obstacles. We now run the basic ART algorithm over the 626-638 MHz UHF band and compare it with the 2.4 GHz WiFi band. For a fair comparison, the transmit power and RF gains of corresponding radios are calibrated such that, for

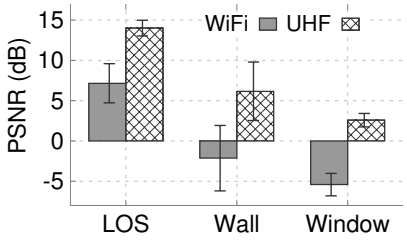

Figure 18: Eavesdropping quality using UHF and WiFi band signals, respectively.

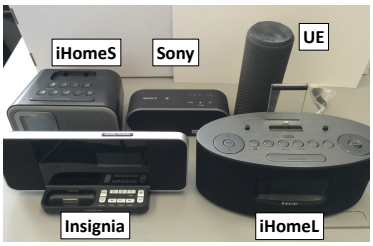

Figure 19: Testing 5 loudspeakers of different size and shape.

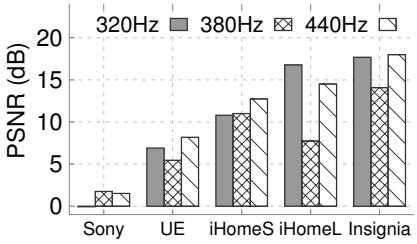

Figure 20: PSNR of different loudspeakers, with increasing size from left to right.

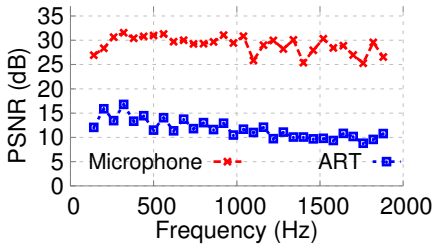

Figure 21: Frequency selectivity of ART compared to microphone.

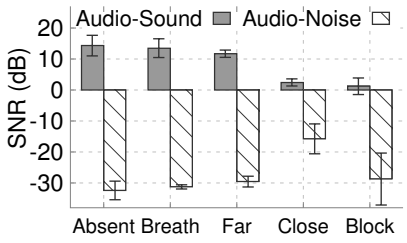

Figure 22: Impact of human present and activity on recovered audio quality.

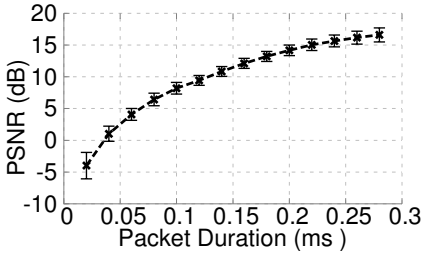

Figure 23: Audio PSNR under different packet lengths.

the same link distance, the received signal power levels' difference follows the free-space model. The results in Figure 18 show that *low-frequency UHF band presents a much higher level of PSNR in sound recovery*. The gain ranges from 7.5dB in LOS and 7 to 9 dB with wall or sound-proof window in between.

Impact of physical properties of loudspeaker: We have also evaluated ART's performance among 5 different commodity loudspeakers (Figure 19): Insignia, iHomeL, iHomeS, Sony and UE. The loudspeakers' volume is set to their largest level. Figure 20 shows that ART has high PSNR of recovered sound across all loudspeakers, and those with larger size tend to have higher PSNR due to stronger vibration.

Frequency fidelity of recovered audio: A common metric for evaluating the fidelity of sound acquisition system is the frequency response. We now test ART against this metric, in comparison with the microphone of a smartphone. The eavesdropper and microphone are both placed 1 m away, and within LOS of the loudspeaker. Figure 21 plots the PSNR across 140 Hz to 1900 Hz, covering the frequency range of human voice. ART's frequency response is relatively flat (*std.* 1.96 dB) and comparable to the microphone recorder (*std.* 1.77dB), which shows that the ART can achieve high-fidelity sound recovery, and does not distort audio differently across different frequencies.

Impact of environment dynamics: We next evaluate how nearby human activities affect ART. We test 5 different cases: *(i) Human Absence*: no human activity. *(ii) Nearby Breath*: a human stands 40cm away from the Tx (without blocking) and breaths normally. *(iii) Walk Far and (iv) Close*: a human randomly walks at a distance around 4m and 40cm to the Tx. *(v) Human Block*: a human blocks the LOS path between Tx and the loudspeaker. In all cases, the Tx is 1m away from the loudspeaker which plays a 400Hz sound. Figure 22 plots the recovered sound PSNR and audio noise floor.

We see that Human Breath and Walk Far almost have no impact on the PSNR or audio noise floor. Walk Close and Human Block decrease the audio PSNR by around 10dB. However, the root causes for the decrease differ in the two cases. For Walk Close case, human motion acts like a strong noise source to loudspeaker vibration, thus lowering the PSNR. In the Human Block case, the noise floor remains similar to the Human Absence, but the human body blockage significantly reduces the reflection signal strength.

From the test result, we conclude that ART can tolerate minor human movement, like breathing, or large motion, but not being too close to the loudspeaker or eavesdropper.

8.2 Validating Emissive Eavesdropping

8.2.1 Micro-benchmark Test

In the following experiments, the target is a smartphone (Moto X XT1053) with internal loudspeaker playing audio. Meanwhile, it sends packets through its WiFi interface to a nearby commercial access point Belkin N150. We do not modify any hardware or software on either the smartphone or the access point. Thus, their communication completely follows the IEEE 802.11g protocol. The adversary SDR works on the sniffer mode, which only captures raw signal samples overheard from the smartphone. We decode the WiFi packets on a PC using our software 802.11g decoder and then reconstruct the audio signals. Since the adversary has no control of the target smartphone, we evaluate the eavesdropping performance when the smartphone generates different traffic patterns. Note that due to 802.11's CSMA contention mechanism, the transmission time of packets is non-uniform, and we handle this problem using the re-interpolation designed in Sec. 5.

First, the smartphone initiates a TCP file transfer at 10 Mbps traffic rate, but with different packet lengths. Meanwhile, it plays a 400 Hz audio. Figure 23 shows that the decoded audio PSNR is logarithmically proportional to packet length. This is consistent with our asymptotic analysis in Sec. 5, which showed that the *RSS amplification algorithm can effectively improve PSNR by accumulating energy across each packet duration*.

We have a similar observation for the impact of packet rate (Figure 24): the PSNR improves by about 8dB as the packet rate increases from 1100 pkt/s to 1900 pkt/s.

The experiments clearly reveal alarming threats in today's wireless communications systems that strive for speed. Modern wireless

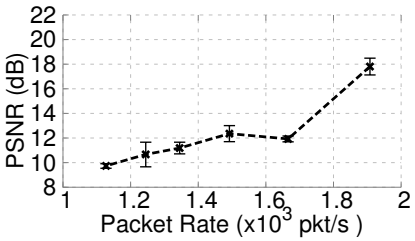

Figure 24: PSNR over different Packet Rate.

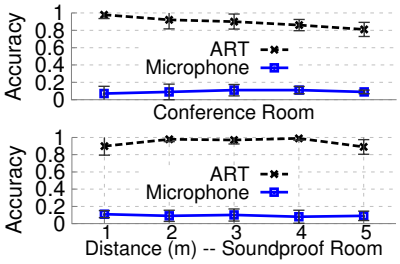

Figure 25: Accuracy of ART compared with a microphone.

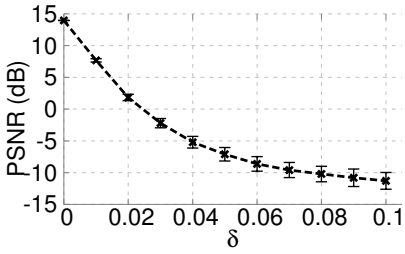

Figure 26: TPR's effect on Audio PSNR.

standards like 802.11n/ac widely adopt packet aggregation to improve transmission efficiency (*e.g.*, 802.11ac allows for up to 5.5ms of packet duration). Meanwhile, packet rate is escalating owing to high PHY-layer bit-rate (*e.g.*, 802.11ad enables up to hundred-Mega per second of packet rate). Riding on these trends, an adversary with wireless vibrometry can easily launch eavesdropping with ultra-high PSNR.

8.2.2 Overall Threat Analysis

Threats from through-wall emissive eavesdropping: To evaluate the practical threat, we use a similar blind-test setup as in the reflective vibrometry, except that the target is a smartphone playing the number sounds while running a file uploading application (with a mean packet duration of 0.25 ms and 1900 pkt/s). Figure 25 plots the blind-test results. We can see that emissive ART achieves almost 100% accuracy when eavesdropper is less than than 3 *m* away from the target with drywall or sound-proof window in between. In contrast, a microphone can barely recognize anything at all distances. Interestingly, the eavesdropping has good performance in spite of the low volume. The is attributed to the more "pure" vibration in the emissive case, where the radio signals come exclusively from the smartphone, whereas the irrelevant reflections in the reflective case contaminate radio signals.

To sum up, *emissive vibrometry constitutes a higher level of threat due to strong signals directly emitted by the target*. Considering the wide penetration of mobile devices that co-locate loudspeaker with radios, emissive ART can easily impinge on private communications over such devices.

Threats under real traffic patterns: Above we have tested emissive ART when the target is running file-uploading applications through WiFi. We have also tested other upload-traffic-dominated applications including FTP and Skype video call. We found FTP results in similar eavesdropping audio quality as above. Yet Skype call has a packet rate below 200 pkts/second, which is insufficient to provide usable audio sampling rate (Sec. 5). However, as camera resolution on smartphone increases along with WiFi bit-rate, we expect the packet rate of such applications will eventually reach the threat level. Besides, we also tested the emissive ART under environment dynamics, with similar observations as the reflective ART. We thus omit the details for the sake of space.

8.3 Validating TPR as a Counter Measure

In this section, we validate the TPR approach against emissive eavesdropping (Sec. 6). Since off-the-shelf smartphone's WiFi chipset cannot be modified to vary transmit power on a fine-grained way, we conduct trace-driven simulation instead. We collect the packet RSS traces following similar setup as in the micro-benchmark test, with default packet length 400 bytes and rate around 1900 pkt/s. Then, we enforce the RSS randomization mechanism on each of the collected packet, and inject the resulting packet trace into ART decoder. Figure 26 plots the decoded audio PSNR. We observe that PSNR is logarithmically proportional to randomization intensity δ, *i.e.*, it drops quickly as δ increases. Even with a small value $\delta = 0.05$, TPR reduces the PSNR from 14 dB to -7 dB (more than 2 orders of magnitude reduction), rendering the decoded sound inaudible. This δ value translates into a variation of transmit power by only 5%, which is unlikely to affect normal wireless communications. Therefore, once TPR is deployed on WiFi firmware, it can effectively counteract the emissive ART.

9. RELATED WORK

Our work is most closely related with prior art in the following domains:

Remote vibration detection. Microwave-based sound recovery was reported early in [25]. We differ from [25] in multiple aspects. First, our system builds on closed-from analysis, together with testbed experiments to demonstrate that both RSS and phase can be effectively used for audio-radio transformation. Second, whereas [25] works in LOS with directional antennas, we leverage the widely available MIMO radios for NLOS sound recovery. Third, we propose a novel emissive attack model that imposes emergent threats on ubiquitous WiFi devices.

Our work has also been inspired by the LADAR concept [5], conventionally used for inspecting structure safety (*e.g.*, bridge shaking under strong wind) and verifying rotational speed of mechanical vibration systems. A LADAR based laser microphone [26] can "hear" sound by measuring the vibration of a window. Recently, Davis *et al.* [9] showed that LADAR can be realized by directly monitoring and computing the vibrating spectrum of target using a high-speed camera. Limitations of such laser or vision based vibrometry systems are discussed in Section 1. WiFi-based vibrometry is less likely to be suspected by a victim, compared with laser or high-speed camera. It can exist in a tampered WiFi device near the victim, or even outside the sight. It is, however, more challenging, as WiFi signals are much less sensitive due to longer wavelength, omni-directional nature, and vulnerability to fading effects.

Radar based activity sensing. The reflective vibrometry follows a similar paradigm in radar sensing. Conventional military radar systems are used for ranging and moving target tracking [27]. Periodic human body activities, such as heart-beat, breathing, and walking, can cause an outstanding Doppler shift in the reflected signal [28, 29], which can be discerned by a radar receiver. But such detection systems require high frequency resolution, with multiple GHz of receiver sampling rate. Recently, substantial work has focused on realizing radar-like functions for mobile applications simply using WiFi signals. WiSee [7] can discriminate the fine Doppler patterns of 9 gestures by transforming WiFi signals into ultra-narrow-band pulses with high frequency resolution. WiVi [30] tracks human walking and gestures by creating a virtual antenna

array, and improving spatial resolution using the inverse synthetic aperture radar (ISAR) technology. WiHear [8] aims to detect human speech by analyzing radio reflections from mouth movements. It builds on a supervised learning framework, requires individual user to train the system extensively, and can recognize only a limited number of words (6 words) with high accuracy.

Acoustic eavesdropping. Auditory surveillance on human conversations is a common practice in espionage, mostly realized in practice using high-sensitivity microphones with wireless transmission capability. Modern acoustic eavesdropping techniques begin targeting electronic apparatus. For example, it is well established that keys on a keyboard can be distinguished by their sound, due to minute differences in mechanical properties such as their position on a slightly vibrating circuit board [2, 31]. Vibrations from key presses can also leak keystroke identities [32]. Remarkably, human speech can also disturb a nearby gyroscope's readings [33], which can in turn infer coarse information about the speaker (*e.g.*, gender).

10. CONCLUSION

We have presented ART, a new acoustic eavesdropping method that penetrates conventional sound-proof isolators using reflective or emissive wireless signals. The key principle and challenge lies in recovering and strengthening the loudspeaker's subtle vibration from the radio signal strength variation. Through an analytical framework and extensive experiments, we distill the key factors that enable highly sensitive ART, and enhance it with diversity-harnessing mechanisms that requires no training preambles from the information source (*i.e.*, the loudspeaker). We implement the ART eavesdropper on a software-radio platform and demonstrate its effectiveness in decoding high-quality audio even through sound-proof walls, showing its severe threat in practice. We have also introduced pragmatic countermeasures in response to this new threat.

Acknowledgement

The work reported in this paper was supported in part by the NSF under Grant CNS-1318292, CNS-1343363, CNS-1350039 and CNS-1404613. It was also partly supported by National Natural Science Foundation of China No. 61202410.

11. REFERENCES

[1] Z. C. Taysi, M. A. Guvensan, and T. Melodia, "TinyEARS: Spying on House Appliances with Audio Sensor Nodes," in *Proc. of ACM BuildSys*, 2010.

[2] L. Zhuang, F. Zhou, and J. D. Tygar, "Keyboard Acoustic Emanations Revisited," *ACM Transactions on Information System Security*, vol. 13, no. 1, 2009.

[3] M. Backes, M. Dürmuth, S. Gerling, M. Pinkal, and C. Sporleder, "Acoustic Side-channel Attacks on Printers," in *Proc. of USENIX Security*, 2010.

[4] D. Genkin, A. Shamir, and E. Tromer, "RSA Key Extraction via Low-Bandwidth Acoustic Cryptanalysis," 2014.

[5] P. Castellini, M. Martarelli, and E. Tomasini, "Laser Doppler Vibrometry: Development of Advanced Solutions Answering to Technology's Needs," *Mechanical Systems and Signal Processing*, vol. 20, no. 6, 2006.

[6] S. Sen, B. Radunovic, R. R. Choudhury, and T. Minka, "You Are Facing the Mona Lisa: Spot Localization Using PHY Layer Information," in *Proc. of ACM MobiSys*, 2012.

[7] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home Gesture Recognition using Wireless Signals," in *ACM MobiCom*, 2013.

[8] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We Can Hear You with Wi-Fi!" in *Proc. of ACM MobiCom*, 2014.

[9] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The Visual Microphone: Passive Recovery of Sound from Video," *ACM Trans. Graph.*, vol. 33, no. 4, 2014.

[10] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[11] W. Klippel and J. Schlechter, "Measurement and Visualization of Loudspeaker Cone Vibration," 2006. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=13716

[12] J. R. Stuart, "Noise: methods for estimating detectability and threshold," *Journal of the Audio Engineering Society*, 1994.

[13] M. Vorlander, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. Springer, 2008.

[14] X. Zhang and K. G. Shin, "E-Mili: Energy-Minimizing Idle Listening in Wireless Networks," *Proc. of ACM MobiCom*, 2011.

[15] S. Sur, T. Wei, and X. Zhang, "Bringing Multi-Antenna Gain to Energy-Constrained Wireless Devices," in *Proc. of ACM/IEEE IPSN*, 2015.

[16] J. Selva, "Functionally weighted Lagrange interpolation of band-limited signals from nonuniform samples," *IEEE Transactions on Signal Processing*, 2009.

[17] "ARP Request Replay Attack," http://www.aircrack-ng.org/doku.php?id=arp-request_reinjection, 2010.

[18] A. Musa and J. Eriksson, "Tracking Unmodified Smartphones Using Wi-Fi Monitors." Proc. of ACM SenSys, 2012.

[19] "Introduction to Wi-Fi Wireless Antennas," http://compnetworking.about.com/od/homenetworkhardware/a/introduction-to-wifi-wireless-antennas.htm, 2015.

[20] "WiFi Signal Attenuation," http://www.liveport.com/wifi-signal-attenuation, 2015.

[21] V. Shrivastava, D. Agrawal, A. Mishra, S. Banerjee, and T. Nadeem, "On the (in)feasibility of Fine Grained Power Control," *Mobile Computing and Communications Review*, vol. 11, no. 2, 2007.

[22] Rice University, "Wireless Open-Access Research Platform," http://warp.rice.edu/trac/wiki, 2013.

[23] Volo Wireless LLC., "Wideband UHF Daughter Card (WURC)," 2014.

[24] "WARPLab 7_3_0 Benchmarks," http://warpproject.org/trac/wiki/WARPLab/Benchmarks/WARPLAB_7_3_0, 2014.

[25] W. McGrath, "Technique and Device for Through-the-Wall Audio Surveillance," 2005, US Patent App. 11/095,122.

[26] R. P. Muscatell, "Laser Microphone," 1984, US Patent 4,479,265.

[27] J. Nanzer, *Microwave and Millimeter-wave Remote Sensing for Security Applications*. Artech House, 2012.

[28] V. Chen, F. Li, S.-S. Ho, and H. Wechsler, "Analysis of Micro-Doppler Signatures," *IEE Proceedings on Radar, Sonar and Navigation*, vol. 150, no. 4, 2003.

[29] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller, "3D Tracking via Body Radio Reflections," in *Proc. of USENIX NSDI*, 2014.

[30] F. Adib and D. Katabi, "See Through Walls with WiFi!" in *Proc. of ACM SIGCOMM*, 2013.

[31] S. Narain, A. Sanatinia, and G. Noubir, "Single-stroke Language-agnostic Keylogging Using Stereo-microphones and Domain Specific Machine Learning," in *Proc. of ACM WiSec*, 2014.

[32] P. Marquardt, A. Verma, H. Carter, and P. Traynor, "(Sp)iPhone: Decoding Vibrations from Nearby Keyboards Using Mobile Phone Accelerometers," in *Proc. of ACM CCS*, 2011.

[33] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing Speech From Gyroscope Signals," in *Proc. of USENIX Security Symposium*, 2014.