RadEar: A Self-Supervised RF Backscatter System for Voice Eavesdropping and Separation

Qijun Wang*, Peihao Yan*, Chunqi Qian†, and Huacheng Zeng*
*Department of Computer Science and Engineering, Michigan State University
†Department of Radiology, Michigan State University

***Abstract*—Eavesdropping on voice conversations presents a growing threat to personal privacy and information security. In this paper, we present RadEar, a novel RF backscatter-based system designed to enable covert voice eavesdropping through walls. RadEar consists of two key components: (i) a batteryless RF backscatter tag covertly deployed inside the target space, and (ii) an RF reader located outside the room that performs signal demodulation, voice separation, and denoising. The tag features a compact, dual-resonator design that achieves energy-efficient frequency modulation for *continuous* voice eavesdropping while mitigating self-interference by separating excitation and reflection frequencies. To overcome the challenges of weak signal reception and overlapping speech, the RF reader employs self-supervised learning models for voice separation and denoising, trained using a remix-based objective without requiring ground-truth labels. We fabricate and evaluate RadEar in real-world scenarios, demonstrating its ability to recover and separate human speech with high fidelity under practical constraints.**

***Index Terms*—Audio security, eavesdropping, RF backscatter communication, self-supervised learning, information privacy**

I. Introduction

Eavesdropping poses a critical threat to our society as it undermines the fundamental right to privacy and compromises the confidentiality of personal and business conversations. In an era where voice-enabled technologies are deeply embedded in everyday life, unauthorized access to spoken conversations can lead to identity theft, intellectual property loss, corporate espionage, and even national security breaches. Unlike traditional data breaches, voice eavesdropping is particularly insidious because it can occur passively and covertly, without leaving digital traces. As communication technologies evolve, so do the capabilities of adversaries to exploit wireless technologies and physical environments to capture sensitive information from a distance. This escalating threat demands a deeper understanding of possible eavesdropping technologies and their attack surfaces, enabling researchers and policymakers to develop countermeasures that safeguard voice privacy.

Since voice-based acoustic signals are easily blocked by soundproof materials, direct acoustic eavesdropping from outside a private space is often infeasible. As a result, Radio Frequency (RF)-based communication and sensing technologies have been explored in increasingly sophisticated forms for eavesdropping applications. Existing efforts include mmWave-based voice detection [1]–[6], RFID-based voice detection [7]–[9], and UWB-based voice detection [10]. While these systems demonstrate the potential of RF signals to recover speech, their practical applications are constrained to controlled environments and lack robustness in real-world scenarios. For example, mmWave signals poorly penetrate walls due to high-frequency attenuation [11], [12], RFID-based methods require physical proximity to the speaker, and UWB approaches face challenges with multipath interference and synchronization.

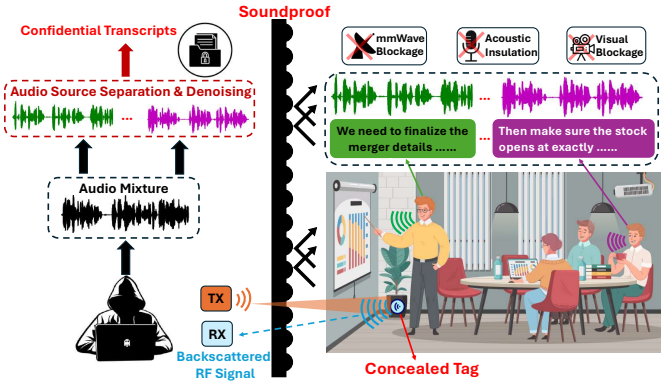

Fig. 1: Threat model and system configuration.

In this paper, we consider a threat model illustrated in Fig. 1, where a confidential conversation involving one or multiple people takes place in a private room protected by soundproof walls. An adversary, located outside the target space, aims to eavesdrop on the conversation without physical access. Moreover, the adversary seeks to separate individual speakers' voices to enable automatic speech transcription, paving the way for advanced semantic analysis such as speaker identification, intent recognition, or topic extraction. To execute this attack, we present an RF backscatter system called Radio Ear, or RadEar for short. RadEar comprises two main components: (i) a passive RF backscatter tag covertly placed within the target space, and (ii) an RF reader positioned outside the room to capture and decode the RF signals modulated by the ambient acoustic vibrations of human speech.

To support a reliable eavesdropping task, the design of the RF backscatter tag must address several key challenges. *First*, the tag must be batteryless and operate solely on harvested energy for computation and communication, enabling permanent deployment within the target space. Moreover, the tag must be small (e.g., less than one inch) and unobtrusive to allow for easy concealment (e.g., covertly attached under a table or chair). *Second*, the tag must support *continuous* voice streaming. While prior batteryless RF backscatter tags have demonstrated voice transmission capabilities (e.g., WISP [13], Battery-Free Phone [14], MARS [15]), they are incapable of supporting *continuous* voice streaming, as they rely on long-time energy harvesting for short-time operation. *Third*,

the tag must reliably modulate the RF signal to maximize the eavesdropping range. A notorious issue in many RF backscatter systems (e.g., RFID) is self-interference: when the tag modulates the voice signal on the same frequency as the excitation signal, the RF reader experiences strong self-interference, significantly degrading its detection range.

To address these challenges, we introduce a novel design for the RF backscatter tag. Our proposed tag comprises four elements: a piezoelectric sensor, a voltage-sensing resonator (VSR), a parametric resonator (PR), and a dipole antenna. The piezoelectric sensor transduces voice-induced pressure fluctuations into voltage signals, which modulate the VSR's resonance frequency. The VSR is magnetically coupled to the PR, which serves as an energy pump and introduces spectral separation via voltage-tunable dual-mode resonance. The modulated signal is radiated through a dipole antenna toward the RF reader, boosting the eavesdropping range.

This new design offers four key advantages over existing backscatter tags. First, it achieves significant spectral separation between excitation and reflection, fundamentally mitigating the issue of *self-interference* at the RF reader. Second, voice signals are modulated via analog frequency shifts, enabling *continuous* voice streaming without digitization. Third, the tag supports frequency modulation (FM), establishing a linear relationship between shifts in the tag's resonance frequency and the amplitude of the voice signal. This greatly simplifies voice recovery at the RF reader. Fourth, the entire design is compact enough to fit on a one-inch PCB, allowing for easy concealment within the target space. Additionally, the tag operates at a low frequency (e.g., 915 MHz), which offers strong wall penetration and favorable propagation characteristics for through-wall eavesdropping.

To increase the eavesdropping range and enable voice separation, the RF reader plays a critical role but faces two key challenges. First, the received signal is extremely weak due to passive modulation and wall attenuation, complicating voice recovery. Second, voice separation is inherently challenging because there are no ground-truth labels for individual speakers' voice data. The low signal-to-noise ratio (SNR) further exacerbates this challenge, making it difficult to isolate and recover clean voice signals.

To address these issues, we propose a learning-based RF reader design that incorporates two innovative components: a voice *separation* model and a voice *denoising* model. Both models are trained in a self-supervised manner to improve their performance without relying on labeled data. Our design leverages the statistical regularities present in human voice mixtures, which allow models to learn meaningful patterns without explicit supervision. Inspired by remixing-based approaches such as those in [16], [17], we train our models to decompose voice mixtures into multiple components and use a reconstruction loss to guide training. Specifically, the separated components are recombined to reconstruct the original mixture, and the loss measures the fidelity of this reconstruction. This remixing-based training provides an implicit form of supervision, encouraging the models to learn useful representations of the underlying sources, thereby improving both voice separation and noise suppression in weak RF signals.

TABLE I: Comparison of Eavesdropping Attack Methodologies. "GT Req." indicates whether ground truth data is required for training. Symbol legend: ○= Low, ◐= Medium, ●= High.

Previous Work	Hardware	Max. Dist.	Thru-Wall	GT Req.	Preset Input?	Mobility	Generalizability	
							User	Env.
WaveEar [18]	mmWave Radar	2 m	✗	✓	✗	○	◐	◐
AccelEve [19]	Accelerometer	0 m	✗	✓	✓	-	◐	◐
UWHear [20]	IR-UWB Radar	8 m	✓	✗	✗	-	◐	◐
Tag-Bug [21]	RFID Tag	2 m	✓	✓	✓	-	◐	◐
AccEar [22]	Accelerometer	0 m	✗	✓	✗	◐	◐	◐
MILLIEAR [23]	mmWave Radar	5 m	✓	✓	✗	-	◐	◐
mmSpy [24]	mmWave Radar	1.8 m	✗	✓	✓	○	◐	◐
mmPhone [25]	mmWave Radar	5 m	✓	✗	✓	-	◐	◐
Wavesdropper [1]	mmWave Radar	5 m	✓	✓	✓	◐	○	◐
mmEve [5]	mmWave Radar	8 m	✗	✓	✗	◐	◐	◐
RADIOMIC [4]	mmWave Radar	4 m	✓	✗	✗	◐	◐	◐
AmbiEar [3]	mmWave Radar	2.5 m	✗	✓	✗	◐	◐	◐
Radio2Text [2]	mmWave Radar	1.5 m	✓	✓	✗	-	◐	◐
Shi *et al.* [6]	mmWave Radar	11 m	✓	✓	✓	◐	◐	◐
mmEav [26]	mmWave Radar	3 m	✗	✓	✓	-	◐	○
RFSpy [7]	RFID Tag	3.5 m	✓	✓	✗	◐	◐	◐
VibSpeech [10]	mmWave Radar	5 m	✓	✗	✗	○	◐	◐
mmEar [27]	mmWave Radar	2 m	✗	✓	✗	○	◐	◐
RF-Parrot [8]	RF tag	1 m	✓	✓	✓	-	◐	◐
Onishi *et al.* [28]	RF Transceiver	2 m	✓	✗	✗	-	◐	◐
This work	**RF transceiver**	**8 m**	✓	✗	✗	●	●	●

We have fabricated the RF backscatter tag and built the RF reader, and evaluated the performance of `RadEar` in real-world environments. Extensive experiments demonstrate that `RadEar` provides a satisfactory voice recovery and separation under various realistic conditions. Table I summarizes the key achievements of `RadEar` against existing approaches.

This work advances the state-of-the-art as follows.

- It introduces a novel design for an RF backscatter tag capable of supporting *continuous* voice streaming via *frequency modulation*.
- It presents a new RF reader design that incorporates *self-supervised* modules for voice separation and denoising.
- Extensive experiments validate the practicality of the proposed eavesdropping system and demonstrate its superior performance.

II. Threat Model and System Overview

A. Threat Model

We consider a threat model as illustrated in Fig. 1, where a confidential conversation involving one or multiple individuals takes place inside a private room protected by soundproof walls. An adversary, positioned outside the target area, aims to *continuously* eavesdrop on the conversation without any physical access to the room. Furthermore, the adversary seeks to separate individual speakers' voices to enable automatic speech transcription, thereby facilitating advanced semantic analysis such as speaker identification, intent recognition, and topic extraction.

To execute this attack, the adversary covertly deploys an RF backscatter tag inside the target room. This tag is battery-less and operates solely by harvesting energy from an external RF reader located outside the target area. This tag is required to operate continuously for voice streaming with a full duty cycle. Additionally, the tag must be small and unobtrusive, allowing it to be discreetly placed within the environment.

The adversary has no access to clean ground-truth audio data from individual speakers, no control over their positions, and no ability to physically tamper with or alter the environment beyond the initial placement of the tag. However,

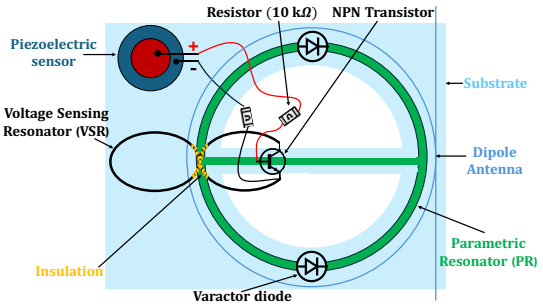

Fig. 2: Diagram of our RF backscatter tag.

the adversary can access a location outside the target space (e.g., behind a wall) to deploy an RF reader. This reader is responsible for RF signal transmission, signal processing, and learning-based computation, with the goal of recovering and separating the voice streams of individual speakers.

B. Proposed Eavesdropping System

To execute this attack, we present `RadEar`, an RF backscatter-based eavesdropping system consisting of two main components: (i) a passive RF backscatter tag covertly placed within the target space, and (ii) an RF reader positioned outside the room to perform signal demodulation, recovery, and voice separation. The tag features a novel architecture with two coupled resonators (VSR and PR), which enable separation of the excitation and reflection signal frequencies, thereby mitigating self-interference for signal demodulation at the RF reader. This architecture also supports energy-efficient frequency modulation for voice signals, enabling *continuous* voice streaming on the tag. The RF reader incorporates two learning-based modules for voice separation and denoising, both of which can be trained in a self-supervised manner to enhance performance in real-world, unlabeled environments.

The following sections detail the RF tag and reader.

III. RF Backscatter Tag Design

A. Overview

The design of a backscatter tag for continuous voice streaming faces two key challenges: *(i) Frequency Separation:* Traditional RF backscatter tags, like RFID, reflect signals at the same frequency as their excitation source. Since the excitation signal is much stronger, it overwhelms the weak reflected signal, causing severe self-interference at the RF reader. This degrades sensitivity, limits range, and leads to high packet error rates (PER). While such degradation may be tolerable for RFID, it is unacceptable for voice streaming, which requires a reliable, continuous data link. Thus, effective separation of excitation and reflection frequencies is essential. *(ii) Efficient Voice Signal Modulation:* RF backscatter tags harvest only minimal power, insufficient to support conventional digital voice modulation. While analog modulation is more power-efficient, existing techniques still consume too much energy for continuous streaming on batteryless tags. These limitations call for novel, low-power modulation methods designed to enable real-time voice communication over RF backscatter.

To address these two challenges, we propose an RF backscatter tag. This RF tag is designed to operate passively, harvesting energy from an external RF reader while modulating acoustic signals onto the reflected carrier. As shown in Fig. 2, the tag consists of four primary components: a piezoelectric sensor, a voltage-sensing resonator (VSR), a parametric resonator (PR), and a dipole antenna. The piezoelectric sensor captures pressure fluctuations induced by nearby voice activity and converts them into low-voltage electrical signals. These signals directly modulate the impedance characteristics of the VSR, whose resonant frequency shifts accordingly. The modulated VSR is coupled to the PR, which acts as an energy pump to enhance the backscatter efficiency. Additionally, the PR separates the excitation and backscatter/reflection signals in frequency due to the dual-mode resonance structure. Finally, the composite signal is radiated via the dipole antenna toward the RF reader, which demodulates the acoustic content embedded in the carrier waveform.

The layout of the VSR and PR resonators, along with their coupling mechanism, is co-designed and jointly optimized to ensure high signal integrity under low-SNR and wide-angle acoustic detection conditions, enabling long-range eavesdropping capabilities. In what follows, we describe the design and operation of the VSR and PR in detail.

B. Voltage-Sensing Resonator (VSR)

The VSR serves as the core transduction unit that maps acoustic pressure into modulated RF backscatter. To support high-sensitivity operation under passive constraints, we implement a novel BJT-based resonator structure that achieves frequency modulation through piezo-induced voltage variations.

The resonator consists of an ∞-shaped metallic conductor with a central symmetry axis, terminated by a bipolar junction transistor (NPN-type) across its open ends. The base and emitter terminals of the transistor are connected to the differential output of a piezoelectric diaphragm, while the collector remains electrically floating. This configuration introduces two PN junctions in series, forming a voltage-variable capacitive network embedded within the resonant loop.

When an incident acoustic waveform induces pressure $p(t)$ on the piezo element, it generates a voltage $v_{\text{pz}}(t) \propto p(t)$ across the base-emitter junction. This voltage modulates the depletion width $w(t)$ of each PN junction, thereby altering its junction capacitance:

$$C_{\text{PN}}(t) = \frac{\varepsilon A}{w(t)} \approx \frac{C_0}{\sqrt{1 + v_{\text{pz}}(t)/\phi_T}}, \tag{1}$$

where C_0 is the junction capacitance at equilibrium, ϕ_T is the thermal voltage, and εA represents the junction permittivity-area product. The total capacitance of the resonator becomes a function of the input sound pressure. Denote $f_{\text{res}}(t)$ as the instantaneous resonance frequency of VSR. Then, based on Eqn (1), we have:

$$f_{\text{res}}(t) = \frac{1}{2\pi\sqrt{LC_{\text{PN}}(t)}} \approx \frac{1}{2\pi\sqrt{LC_0}}\left(1 + v_{\text{pz}}(t)/\phi_T\right)^{\frac{1}{4}}. \tag{2}$$

In practice, the voice-induced piezo voltage is weak and thus we have $v_{\text{pz}}(t)/\phi_T \ll 1$. Therefore, based on the Taylor series expansion, Eqn (2) can be approximated by:

$$f_{\text{res}}(t) \approx \frac{1}{2\pi\sqrt{LC_0}}\left(1 + \frac{v_{\text{pz}}(t)}{4\phi_T}\right). \tag{3}$$

Eqn (3) shows the linear relationship between VSR's resonance frequency $f_{\text{res}}(t)$ and the input sound pressure $v_{\text{pz}}(t)$. This is Frequency Modulation (FM). It encodes the temporal variations of voice signals directly into the tag's resonant frequency, which is later extracted by the RF reader.

In our design, the bipolar transistor's dual-junction structure enables symmetric capacitance tuning, which improves linearity and reduces phase noise in VSR's frequency modulation. Additionally, the ∞-shaped conductor suppresses common-mode interference, spatially varying environmental noise, and multipath artifacts. Together, these features enhance the robustness and stability of the VSR's frequency modulation across diverse indoor deployment scenarios.

C. Parametric Resonator (PR)

Separation of Excitation and Reflection Frequencies: Traditional backscatter systems suffer from self-interference due to spectral overlap between excitation and reflection signals, making it difficult for the RF reader to recover weak modulated signals. To address this, we propose a dual-mode PR, as illustrated in Fig. 2, which supports concurrent circular and butterfly resonance modes. Let f_c and f_b denote the resonance frequencies of the circular and butterfly modes, respectively. When the tag is excited by an external signal at frequency $f_{\text{ex}} = f_c + f_b$, both modes are simultaneously activated, generating signals at f_c and f_b. We designate f_c as the carrier frequency of the reflected signal and enhance its radiation with a dipole antenna. This separation of excitation and reflection frequencies allows the RF reader to effectively isolate the backscattered signal and suppress self-interference.

Magnetic Coupling Between PR and VSR: The VSR is magnetically coupled to the PR's circular resonance mode, allowing voice-induced voltage shifts to modulate the reflection frequency via frequency modulation. The butterfly mode in the PR facilitates energy pumping and oscillation without interfering with voice encoding. This dual-mode coupling scheme enables passive voice modulation and amplification while maintaining spectral isolation. The resonant coupling further enables efficient transfer of modulated energy to the antenna, enhancing backscatter signal strength while preserving the tag's passive nature.

To improve voice-induced frequency modulation efficiency, the PR is designed as a lumped-element LC circuit incorporating a voltage-tunable varactor diode. The resonance frequency f_c of its circular mode is tuned to closely match the nominal resonance frequency of the VSR. When the two resonators are placed in close proximity, they are electromagnetically coupled via mutual inductance, enabling energy exchange.

By carefully tuning the varactor in the PR, we bias its resonance slightly off the carrier frequency to maximize sensitivity to the frequency-shifted sidebands induced by voice signals. This asymmetric tuning allows the PR to function as a passive, frequency-selective amplifier, concentrating energy from the voice-modulated subcarriers into a narrow spectral band aligned with the reader's receiving frequency.

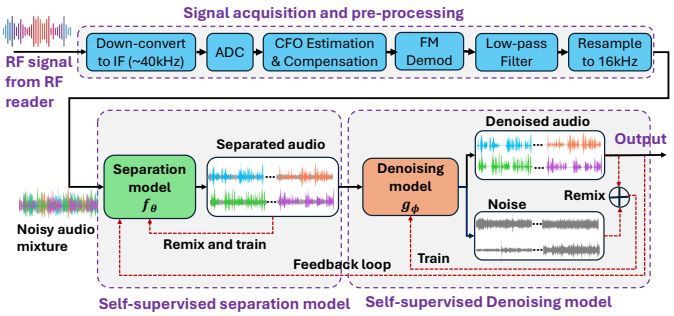

Fig. 3: Diagram of our RF reader design.

Dipole Antenna: A dipole antenna is wrapped around the edge of the PR's circular conductor to enhance inductive coupling. Resonant amplification allows small variations in the VSR's frequency to produce detectable amplitude and phase shifts in the backscattered signal, enabling robust acoustic signal recovery even under low-SNR indoor conditions.

Backscattered Signal: In principle, the PR and the dipole antenna together function as an amplifier, enabling the VSR to more effectively radiate its modulated signal. Coarsely, the instantaneous frequency of the tag's backscattered signal can be approximated as:

$$f_c(t) \approx \gamma_1\gamma_2 f_{\text{res}}(t) \approx \frac{\gamma_1\gamma_2}{2\pi\sqrt{LC_0}}\left(1 + \frac{v_{\text{pz}}(t)}{4\phi_T}\right), \tag{4}$$

where γ_1 and γ_2 are the gains from the PR and dipole antenna, respectively. Overall, the coupled VSR–PR system acts as a compact, passive frequency modulation transducer that enhances acoustic detection sensitivity, forming the physical foundation for long-range voice detection.

IV. RF Reader Design: Audio Recovery and Separation

The RF reader plays a critical role in demodulating, separating, and denoising the voice signal received from the backscatter tag. Fig. 3 illustrates our RF reader design, which comprises three key components: (i) signal acquisition and pre-processing, (ii) a self-supervised model for voice separation, and (iii) a self-supervised model for voice denoising. In the following sections, we describe each component in detail.

A. Signal Acquisition and Pre-Processing

To detect the acoustic signal, the RF reader performs continuous-wave (CW) transmission at a fixed carrier frequency f_{ex}. Meanwhile, it receives the backscattered signal from the tag at a different frequency f_{c}. The isolation of the transmitting and receiving signal frequencies avoids the notorious self-interference issues and thus significantly improves the voice detection range. The backscattered RF signal is first down-converted to an intermediate frequency (IF) (e.g., 40 kHz) and then sampled for digital signal processing. The reason why we convert the radio signal to IF rather than

baseband (zero-IF) is twofold. First, voice signals contain rich low-frequency components (typically below 100 Hz), and the RF front end often suffers from disturbances near the DC component. Therefore, using IF will avoid interference at low frequencies. Second, the radio frequency suffers from CFO. Since the subsequent process needs to compensate the CFO, using IF will not incur an additional processing burden.

The digitalized signal samples are then processed by a sequence of blocks for CFO compensation. In this process, the CFO is tracked over time and compensated for individual segmented signal frames. The signal after CFO compensation can be expressed as:

$$r(t) = A(t) \cdot \exp\left(j2\pi \cdot f_{\text{bb}}(t) \cdot t\right) + n(t), \tag{5}$$

where $A(t)$ represents time-varying signal amplitude and $f_{\text{bb}}(t)$ represents the instantaneous frequency deviation introduced by the acoustic signal, and $n(t)$ captures RF noise. We demodulate the voice signal by differentiating the phase of $r(t)$ over time. The demodulated voice signal, denoted as $x(t)$, can be written as:

$$x(t) = \frac{f_{\text{s}}}{2\pi\Delta}[\angle r(t+\Delta) - \angle r(t)], \tag{6}$$

where f_{s} is the sampling rate of $r(t)$ and Δ is a fixed small time step. We note that the demodulated voice signal $x(t)$ in Eqn (6) is an estimate of $v_{\text{pz}}(t)$ in Eqn (1). The demodulated signal $x(t)$ serves as the input to the downstream voice separation and denoising pipeline.

B. Self-supervised Audio Separation Model

The demodulated signal contains a mixture of speech from multiple speakers, along with background noise, environmental interference, and hardware imperfections. Since the number of speakers, denoted by N, is unknown, the signal can be modeled as: $x(t) = \sum_{i=1}^{N} s_i(t) + n(t)$, where $s_i(t)$ represents the voice signal from voice source i and $n(t)$ represents the noise. The objective of this model is to estimate the number of voice sources and separate them. i.e., $[\hat{s}_1(t), \hat{s}_2(t), \ldots, \hat{s}_{\hat{N}}(t)] = f_\theta(x(t))$, where $f_\theta(\cdot)$ represents this model parameterized by θ, $\hat{N}$ is the number of estimated voice sources, and $\hat{s}_i(t)$ is the estimated voice signal from source i.

Why Self-Supervised Learning? Mixed voice signals can be separated through self-supervised learning by leveraging the inherent statistical regularities in audio mixtures. Human speech tends to be sparse in the time-frequency domain, meaning different speakers often dominate different spectrogram regions. This sparsity allows a model to infer which components belong together based on co-occurrence and consistency across mixtures. In self-supervised training, models can learn from "mixtures of mixtures" without needing ground-truth sources. The model separates each input into components and uses a reconstruction loss to check whether these components can be recombined to recover the original mixtures. This remixing constraint acts as a form of supervision, guiding the model to discover meaningful representations of individual sources. Over time, the model learns to identify and disentangle recurring speech patterns, enabling effective source separation purely from the structure of the input data.

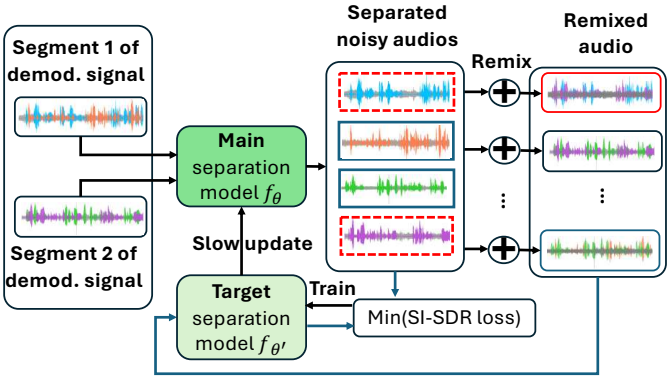

Fig. 4: Self-supervised training of audio separation model.

Main Idea: Our design was inspired by the Mixture Invariant Training (MixIT) [16], which enables speaker separation using only the demodulated audio stream. Training samples are constructed by adding two segments of the audio to form a "mixture of mixtures" (MoM). A neural network is trained to separate the MoM into a set of latent sources, which are then remixed to approximate the original input mixtures. The model is optimized using a signal-level loss that promotes source separation without requiring any ground-truth voice data. Importantly, this method supports a variable number of audio sources, making it well-suited for our application.

Our Design: We use Conv-TasNet [29] as the backbone network for both the main and target models, as it has been widely adopted for audio processing tasks. Fig. 4 illustrates the self-supervised training process. Two segments of the demodulated audio stream are selected and separately fed into the main model, which produces two sets of pseudo-source signal estimates. We then randomly select one signal from each set and apply random weights to generate a remixed signal. The remixed signal is then fed into the target model. We compute the scale-invariant signal-to-distortion ratio (SI-SDR) loss between the selected pseudo-source signals generated by the main model and the outputs of the target model, and use this loss to update the target model. The main and target models share the same architecture, and the parameters of the target model are slowly copied to the main model over time.

Mathematically, denote $x_1(t)$ and $x_2(t)$ as two demodulated signal segments as shown in Fig. 4. Let $\hat{\mathcal{S}}^{(1)} = f_{\boldsymbol{\theta}}(x_1)$ and $\hat{\mathcal{S}}^{(2)} = f_{\boldsymbol{\theta}}(x_2)$ represent the pseudo-source signal sets generated by the main model. We randomly select one signal from each set, denoted as $\hat{s}_a \in \hat{\mathcal{S}}^{(1)}$ and $\hat{s}_b \in \hat{\mathcal{S}}^{(2)}$, and construct a remixed signal $\tilde{x}(t) = \hat{s}_a(t) + \hat{s}_b(t)$. The target model then produces an estimated source set $\hat{\mathcal{S}}' = f_{\boldsymbol{\theta}'}(\tilde{x})$. The loss function is defined as:

$$\mathcal{L}_{\text{sep}} = \min_{\pi \in \mathcal{P}} \sum_j \ell\left(\hat{s}'_j(t), \hat{s}_{\pi(j)}(t)\right), \tag{7}$$

where $\ell(\cdot,\cdot)$ denotes the scale-invariant signal-to-distortion ratio (SI-SDR) loss, and $\mathcal{P}$ is the set of all permutations over the selected pseudo-source signals. This loss trains the target network with parameters $\boldsymbol{\theta'}$, which periodically update the

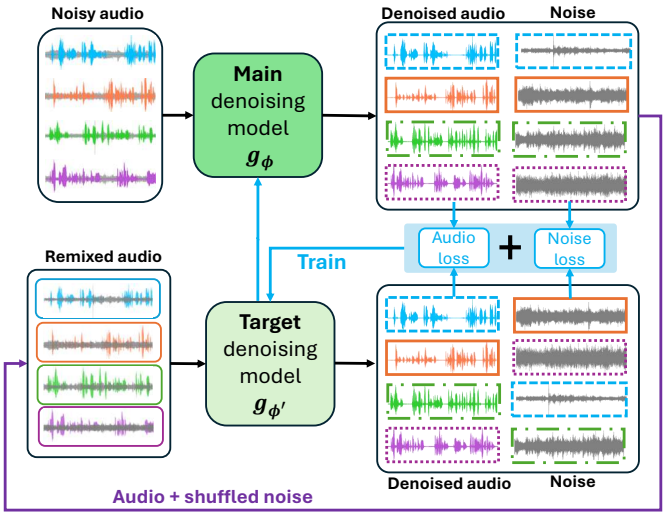

Fig. 5: Self-supervised training of audio denoising model.

main network via $\boldsymbol{\theta} \leftarrow \lambda \cdot \boldsymbol{\theta} + (1-\lambda) \cdot \boldsymbol{\theta'}$, where λ is an empirical updating rate.

For this design, we have the following remarks. *First*, we use a main and target model pair to improve training stability. Direct updates to the main model led to unstable convergence in our experiments. This issue is well known in reinforcement learning, where a target network is used to mitigate instability. *Second*, to accelerate training, we pre-train the main model on out-of-domain (OOD) speech datasets LibriMix [30], and initialize the target model randomly. The main model remains frozen and receives periodic EMA updates with high momentum to preserve stability. *Third*, since the loss is computed over all permutations of separated signals, the computational cost grows rapidly with the number of speakers. We therefore limit the number of speakers to two in our experiments.

C. Self-Supervised Denoising Model

Main Idea: Noise and voice signals exhibit distinct characteristics in the temporal domain. Speech typically follows structured, periodic patterns corresponding to phonemes and syllables, while noise is often unstructured, irregular, or statistically random. These differences enable a model to learn discriminative features that separate speech from noise. By training on diverse mixtures, the model can recognize and exploit these temporal patterns to effectively isolate voice signals from noise. Our design drew inspiration from the self-training speech enhancement approach proposed in [17]. The core idea is to remix the signal estimates with permuted noise estimates to generate pseudo-labels, which are then used for the self-supervised training of the denoising model.

Our Design: Fig. 5 shows the self-supervised training process for our denoising model. Similar to the separation task, we adopt a main-target model pair for signal denoising. The main and target models share the same backbone network architecture, e.g., SuDoRM-RF [31] in our experiments. The main model is used to separate a batch of noisy signals, producing corresponding signal and noise estimates. Next, we *shuffle* the noise estimates and remix them with the signal estimates to generate new inputs for the target model. The target model then produces new signal and noise estimates. We compute the loss by comparing the signal and noise estimates from the main and target models and use this loss to train the target model. The parameters of the target model are then used to slowly update the main model parameters. To accelerate the training process, we pre-train the main model using the OOD speech dataset DNS [32], while initializing the target model with randomized parameters. To prevent degradation from noisy updates, we adopt a high-momentum EMA to slowly refine the main model parameters over time.

Mathematically, denote $\hat{s}_i(t)$ as the estimated signal from a single source (i.e., output of the separation model, where $i = 1, 2, \ldots, N$. Let $\hat{s}_i(t) = \hat{c}_i(t) + \hat{n}_i(t)$, where $\hat{c}_i(t)$ and $\hat{n}_i(t)$ are the signal and noise estimates generated by the main model. Denote $p(\cdot)$ as a permutation function for the integers in $\{1, 2, \ldots, N\}$. Then, we shuffle the noise estimates and mix them with the signal estimates: $\tilde{s}_i(t) = \hat{c}_i(t) + \hat{n}_{p(i)}(t)$, which is fed to the target model. Denote $\tilde{c}_i(t)$ and $\tilde{n}_i(t)$ as the signal and noise estimates from the target model. Then, we calculate the loss function via:

$$\mathcal{L}_{\text{den}} = \sum_{i=1}^{N} \left[\ell\big(\hat{c}_i(t), \tilde{c}_i(t)\big) + \ell\big(\hat{n}_{p(i)}(t), \tilde{n}_i(t)\big) \right]. \tag{8}$$

This loss function is used to update the target model. Periodically, we use the target network parameters to update the main network parameters by: $\boldsymbol{\phi} \leftarrow \lambda' \cdot \boldsymbol{\phi} + (1-\lambda') \cdot \boldsymbol{\phi'}$, where λ' is an empirical updating rate.

Feedback of Denoised Signal: As shown in Fig. 3, the denoised signal can be fed back to the separation model to generate improved signal mixtures for continual training. After the denoising model estimates enhanced speech streams from noisy inputs, these outputs are used to construct new synthetic mixtures for the separation model. This feedback mechanism enables the separation model to train on cleaner and more structured input signals, improving its ability to identify and disentangle overlapping sources.

This design is grounded in the concept of distributional alignment [33]. Noisy pseudo-labels from early separation outputs introduce inconsistencies that may hinder learning. By applying denoising prior to remixing, the feedback loop provides higher-quality training targets that better reflect the underlying source structure. The resulting mixtures exhibit clearer speech patterns with increased sparsity in the time-frequency domain, which promotes convergence and facilitates more accurate source decomposition.

Moreover, this feedback encourages mutual co-adaptation between the denoising and separation models. As the denoiser improves the quality of its output, the separator receives progressively cleaner mixtures, allowing both models to refine their parameters in a collaborative and self-reinforcing manner.

V. Experimental Evaluation

To assess the performance of `RadEar`, we conduct a comprehensive evaluation that spans both hardware and software

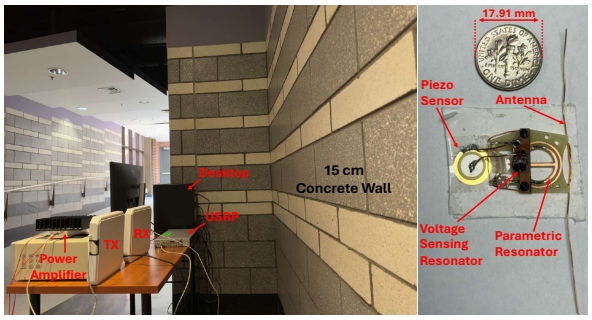

Fig. 6: RF reader and backscatter tag.

components. Our evaluation seeks to answer the following key questions:

- **Effectiveness:** Can `RadEar` reliably recover intelligible speech from raw RF-demodulated signals in multi-speaker acoustic environments?
- **Robustness:** How does `RadEar` perform under varying deployment conditions, including different distances, orientations, and environmental noise?
- **Adaptability:** Does the system maintain performance when the tag is deployed at diverse locations or when speakers move dynamically within the space?

A. Implementation

RF Tag. Fig. 6 shows our custom-designed RF tag comprising four components: a Voltage Sensing Resonator (VSR), a Parametric Resonator (PR), a piezoelectric sensor, and a half-wave dipole antenna. The VSR uses an ∞-shaped coil (32-G copper wire, 5 turns and 1 turn on two rods spaced 1.8 mm apart), terminated by a bipolar junction transistor (BJT) whose base and emitter connect to the piezoelectric sensor. Operating in the active region, the BJT serves as a voltage-variable capacitor, enabling frequency modulation of the VSR via piezo-induced voltage. The PR is a circular planar resonator on a 0.8-mm G10 substrate (13.5/14.5 mm inner/outer diameters), incorporating varactor diodes (9.1 pF) across split gaps to support dual-mode resonance (circular and butterfly modes). A half-wave dipole antenna wraps around the PR's perimeter for loose inductive coupling, boosting energy transfer without a direct electrical connection. Together, the PR acts as a frequency-selective amplifier that upconverts the VSR's modulated signal into a detectable backscatter waveform.

RF Reader. The RF reader setup, as shown in Fig. 6, consists of a USRP N310, a power amplifier (PA), two directional antennas, and a host PC. The reader continuously transmits an excitation signal at 915 MHz with a power output of 30 dBm. In response, the tag generates a frequency-modulated backscatter signal centered at 515 MHz, corresponding to the circular-mode resonance frequency of the PR.

Separation Model. We adopt Conv-TasNet [29] as the backbone network for source separation. The main model is pre-trained on LibriMix [30] (5s segments), using STFT (512-point window, 128 hop, Hann window) with mixture-consistent masks. After in-domain data collection, we employ a main-target model setup: the main model generates pseudo-labels via cross-mixture remixing, and the target model is trained with permutation-invariant negative SNR loss (max 25 dB). Training uses Adam (lr = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$), batch size 128, gradient clipping, and early stopping.

TABLE II: Ablation study.

Configuration	Choice			
w/o PT	✔	✘	✘	✘
w/ PT	✘	✔	✔	✔
EMA Update	✔	✘	✔	✔
Feedback Loop	✔	✘	✘	✔
SI-SDR	7.56	8.75	9.42	10.87
LLR	0.45	0.37	0.34	0.29
STOI	0.69	0.78	0.81	0.85
PESQ	2.89	3.04	3.09	3.27

Denoising Model. We adopt SuDoRM-RF [31] as the backbone network for denoising. The main model is pre-trained on DNS [32] to estimate speech/noise components. These are shuffled and remixed to train a randomly initialized target model with negative SI-SDR loss. We apply EMA updates from the target to the main model after each step for stability. Training uses 16 kHz audio, Adam optimizer (initial lr = 10^{-3}, halved every 6 epochs), batch size 2, and standardized inputs.

B. Experimental Setup and Evaluation Metrics

As illustrated in Fig. 6, we position the RF reader outside the target room, separated by a 15 cm thick concrete wall that is impermeable to mmWave signals. The RF tag is discreetly deployed inside the room and placed in inconspicuous locations to avoid detection. To evaluate the quality of recovered voice signals, we employ a combination of signal-level and perceptual-level metrics. Signal-level metrics offer objective measures of reconstruction fidelity, while perceptual metrics reflect intelligibility and quality as perceived by human listeners. Specifically, we use the following metrics for evaluation.

- **SI-SDR (Signal-to-Distortion Ratio):** A signal-level metric that quantifies the fidelity of the reconstructed signal relative to the distortion, invariant to global scaling.
- **LLR (Log-Likelihood Ratio):** A signal-level metric that measures spectral distortion between the clean and enhanced signals using linear predictive coding coefficients.
- **STOI (Short-Time Objective Intelligibility):** A perceptual-level metric ranging from 0 to 1, which assesses speech intelligibility by comparing the short-time spectral envelopes of the clean and processed signals.
- **PESQ (Perceptual Evaluation of Speech Quality):** A perceptual-level metric ranging from 1 to 4.5, designed to quantify the overall perceptual quality of speech based on comparisons with a reference signal.

For LLR, lower values indicate better performance. For SI-SDR, STOI, and PESQ, higher values correspond to improved recovery quality.

C. Ablation Study

To quantify the contribution of each software component in `RadEar`, we conduct a comprehensive ablation study focusing on three key design elements: (i) pre-training, (ii) exponential moving average (EMA) updates, and (iii) the

TABLE III: RadEar's performance at different distances.

Distance	SI-SDR↑	LLR↓	STOI↑	PESQ↑
50 cm	13.75	0.23	0.92	3.46
100 cm	12.17	0.28	0.89	3.32
150 cm	10.24	0.31	0.84	3.19
200 cm	7.93	0.47	0.71	2.92
250 cm	7.06	0.51	0.67	2.83
300 cm	5.22	0.69	0.56	2.51
350 cm	4.49	0.77	0.53	2.35
400 cm	4.15	0.85	0.50	2.31
450 cm	2.77	0.96	0.39	1.76
500 cm	2.13	1.15	0.34	1.62
550 cm	1.84	1.37	0.31	1.48
600 cm	1.13	1.44	0.25	1.17

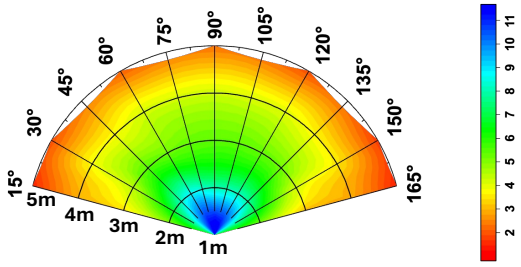

Fig. 7: Measured SI-SDR Across RadEar Orientations.

feedback loop between the separation and denoising modules. All experiments are performed under the same dataset and conditions used in the main evaluation.

As shown in Table II, pre-training provides a strong initialization that significantly boosts performance across all metrics. In the absence of pre-training, we observe a substantial drop in SI-SDR (from 10.87 dB to 7.56 dB), along with corresponding declines in intelligibility and perceptual quality. In addition, the feedback loop, which enables continual co-adaptation between the denoiser and separator, further enhances the overall system performance.

D. Impact of Victim-to-Tag Distance

To assess how speaker-to-tag distance impacts RadEar's ability to recover intelligible speech, we systematically vary the distance from 50 cm to 600 cm in 50 cm increments, while keeping the RF reader and tag positions fixed. The tag is concealed beneath a table to emulate a realistic covert deployment. As shown in Table III, RadEar maintains strong performance when the speaker is within 2.5 m of the tag, achieving median SI-SDR above 7.06 dB, LLR below 0.51, STOI above 0.67, and PESQ exceeding 2.83. Under these conditions, the perceptual quality of the recovered speech remains nearly indistinguishable from the original. As the distance increases beyond 2.5 m, all metrics exhibit a gradual decline, reflecting diminished acoustic excitation at the tag and reduced signal fidelity.

E. Impact of Reader-to-Tag Orientation

A key design element of RadEar is its circular parametric resonator, which efficiently harvests RF energy across a wide range of incident angles. While energy transfer is theoretically maximized when the reader is perpendicular to the resonator, indoor multipath reflections help compensate for angular misalignment.

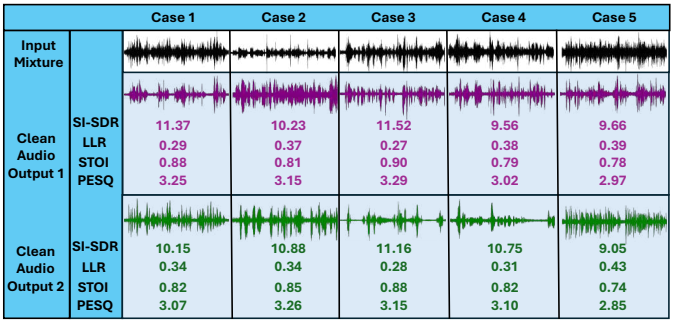

Fig. 8: Impact of noise on separation and denoising.

To evaluate orientation sensitivity, we vary the angle between the RF reader and tag from 15° to 165° in 15° steps, across multiple reader-to-tag distances (1–5 m). Recovered audio is evaluated using the SI-SDR metric.

As shown in Fig. 7, RadEar consistently recovers speech at all tested angles. Although SI-SDR slightly varies with angles, the degradation is far less than that caused by increased distance. These results show that the resonator's circular geometry and indoor reflections enable reliable energy harvesting and voice recovery without precise reader alignment.

F. Impact of Environmental Noise

To assess RadEar 's resilience in noisy environments, we introduce controlled background disturbances during voice capture and evaluate their impact on audio recovery quality. Specifically, we simulate five realistic acoustic scenarios in which two participants speak simultaneously while different environmental noise sources are active. The noise types include: (i) background music played through a speaker, (ii) keystrokes from a mechanical keyboard, (iii) periodic smartphone notification alerts, (iv) prerecorded urban street noise, and (v) speech from online videos played on a mobile phone.

As illustrated in Fig. 8, each case corresponds to one of the above noise conditions. RadEar maintains strong performance under background music and smartphone alerts, with minimal degradation in speech separation. In the keystroke scenario, the system effectively removes the pulse-like tapping sounds while preserving speech intelligibility. Performance declines moderately in the presence of urban noise and online video playback, primarily due to the introduction of competing speech-like signals that overlap with target voices. Nevertheless, even in these more challenging settings, RadEar continues to recover intelligible speech, demonstrating strong resilience to diverse acoustic interference.

G. Impact of Tag Placement

To examine how tag placement affects RadEar 's performance, we test three typical concealment scenarios: (i) beneath a desk, (ii) inside an inconspicuous box placed on a shelf, and (iii) mounted on the backside of a furniture block. As illustrated in Fig. 9, Deployment1 through 3 correspond to the above cases, respectively. Across all settings, RadEar consistently achieves intelligible speech recovery, demonstrating strong robustness to placement variation.

Among the three, enclosing the tag in a box slightly reduces performance, likely due to attenuation of acoustic and RF

Fig. 9: Mel Spec. across tag locations and speaker mobility.

signals. Mounting the tag on the backside of a furniture block performs better than placing it under a desk, as furniture blocks are more stationary and less affected by human interaction, while tables may experience frequent movement or vibration, introducing interference that degrades signal quality.

H. Impact of Target Movement

To evaluate the robustness of `RadEar` under target mobility, we design two experimental scenarios involving speaker movement during voice capture: *(i) Single moving participant.* Two participants remain stationary and speak simultaneously, while a third individual walks continuously within the room, occasionally passing near the tag. *(ii) Two moving speakers.* Both speaking participants walk around the room while conversing. Their movement paths include varying distances, angles, and intermittent proximity to the tag.

As illustrated in Fig. 9, Mobility1 and Mobility2 correspond to the two scenarios described above. In Mobility1, where only a non-speaking participant moves, `RadEar` maintains strong performance with no noticeable degradation. In contrast, Mobility2 introduces significant degradation, likely due to increased variation in speaker position and orientation, which leads to greater signal overlap and spectral ambiguity. Nevertheless, `RadEar` recovers intelligible speech, demonstrating resilience under complex real-world mobility.

VI. Related Work

A. Acoustic eavesdropping

Recent advances have shown that speech can be recovered without traditional microphones, using various wireless and sensing modalities. These systems differ in sensing targets, signal processing, and hardware platforms.

A large body of work employs millimeter-wave (mmWave) radar to detect voice-induced vibrations from human tissue or nearby surfaces. For example, Wavesdropper [1] and Radio2Text [2] utilize commercial mmWave sensors to extract vibrations from the throat or environmental surfaces. AmbiEar [3] and RADIOMIC [4] extend this to passive objects, while mmEve [5] targets earpiece emissions. Shi et al. [6] introduce a phased MIMO array that improves directionality and enables limited through-wall sensing, though it still relies on precise object localization and calibration.

Backscatter-based systems like RFSpy [7] and RF-Parrot [8] utilize passive tags or conductive paths to capture vibrations, yet depend on resonant structures or wired access. UWB systems such as VibSpeech [10] offer multipath resilience but are constrained by transmission power and limited range.

Some systems adopt unconventional sensing mechanisms. Sound of Interference [28] exploits electromagnetic leakage from digital microphones via near-field probes. Meanwhile, mmEve [5] enhances mmWave sensing via generative denoising and IQ compensation to mitigate motion artifacts.

Despite their diversity, existing approaches face three core limitations: (i) mmWave and UWB signals are heavily attenuated by walls; (ii) many systems require powered sensors, precise alignment, or object-specific targeting; and (iii) most rely on supervised learning with clean audio pairs. In contrast, `RadEar` operates in sub-GHz bands for stronger wall penetration, supports passive and batteryless eavesdropping without alignment, and uses a self-supervised learning framework that removes the need for labeled data.

B. Audio Separation and Enhancement.

Traditional supervised models like Conv-TasNet [29] and SepFormer [34] achieve strong performance but rely on clean, labeled data. To relax this requirement, self-supervised methods have emerged. MixIT [16] trains on mixtures of mixtures but may over-separate. Follow-ups like TS-MixIT [35] and MixCycle [36] improve training stability. Denoising methods like RemixIT [17] and Self-Remixing [37] iteratively bootstrap pseudo-labels without clean references. Most prior works address either separation or denoising in isolation. `RadEar` unifies both through a remixing feedback loop, where the denoised outputs are fed back to refine separation, and the separation results in turn guide further enhancement. This self-reinforcing process enhances stability, mitigates over-separation, and enables robust recovery from degraded RF-demodulated signals, without requiring any labeled audio.

VII. Conclusion

In this paper, we presented `RadEar`, a novel batteryless RF backscatter-based eavesdropping system that enables voice recovery from within a soundproof room. Unlike existing backscatter tags, `RadEar` supports *continuous* voice streaming while operating solely on harvested energy. Our design enables reliable audio eavesdropping through frequency-domain self-interference mitigation and ultra-low-power analog FM voice encoding on a compact batteryless tag, with an RF reader using self-supervised models for voice denoising and separation. Extensive experiments demonstrate the system's ability to recover and separate human speech with high fidelity under realistic conditions. This work introduces a new class of passive, long-range voice eavesdropping attacks and underscores the need for renewed attention to RF-based privacy threats.

Acknowledgment

This work was supported in part by NSF Grant ECCS-2225337 (Q. Wang, P. Yan, and H. Zeng) and in part by NSF Grant ECCS-2144138 and NIH Grant RF1NS128611 (C. Qian).

REFERENCES

[1] C. Wang, F. Lin, Z. Ba, F. Zhang, W. Xu, and K. Ren, "Wavesdropper: Through-wall word detection of human speech via commercial mmwave devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–26, 2022.

[2] R. Zhao, J. Yu, H. Zhao, and E. C. Ngai, "Radio2text: Streaming speech recognition using mmwave radio signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 3, pp. 1–28, 2023.

[3] J. Zhang, Y. Zhou, R. Xi, S. Li, J. Guo, and Y. He, "Ambiear: mmwave based voice recognition in nlos scenarios," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–25, 2022.

[4] M. Z. Ozturk, C. Wu, B. Wang, and K. R. Liu, "Radiomic: Sound sensing via radio signals," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 4431–4448, 2022.

[5] C. Wang, F. Lin, T. Liu, K. Zheng, Z. Wang, Z. Li, M.-C. Huang, W. Xu, and K. Ren, "mmeve: eavesdropping on smartphone's earpiece via cots mmwave device," in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022, pp. 338–351.

[6] C. Shi, T. Zhang, Z. Xu, S. Li, D. Gao, C. Li, A. Petropulu, C.-T. M. Wu, and Y. Chen, "Privacy leakage via speech-induced vibrations on room objects through remote sensing based on phased-mimo," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 75–89.

[7] Y. Chen, J. Yu, Y. Chen, L. Kong, Y. Zhu, and Y.-C. Chen, "Rfspy: Eavesdropping on online conversations with out-of-vocabulary words by sensing metal coil vibration of headsets leveraging rfid," in *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*, 2024, pp. 169–182.

[8] Y. Yang, G. Wang, Z. An, G. Zhang, X. Cheng, and P. Hu, "Rf-parrot: Wireless eavesdropping on wired audio," in *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*. IEEE, 2024, pp. 701–710.

[9] Q. Wang, C. Qian, P. Yan, S. Zhang, and H. Zeng, "A batteryless wireless microphone using rf backscatter," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 9, no. 4, pp. 1–18, 2025.

[10] C. Wang, F. Lin, H. Yan, T. Wu, W. Xu, and K. Ren, "{VibSpeech}: Exploring practical wideband eavesdropping via bandlimited signal of vibration-based side channel," in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 3997–4014.

[11] S. Zhang, Q. Wang, M. Gan, Z. Cao, and H. Zeng, "Radsee: See your handwriting through walls using fmcw radar," in *Proceedings of Network and Distributed System Security Symposium (NDSS)*, 2025.

[12] S. Zhang, Q. Wang, K. Song, Q. Yan, and H. Zeng, "Radeye: Tracking eye motion using fmcw radar," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–13.

[13] R. Menon, R. Gujarathi, A. Saffari, and J. R. Smith, "Wireless identification and sensing platform version 6.0," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 899–905.

[14] V. Talla, B. Kellogg, S. Gollakota, and J. R. Smith, "Battery-free cellphone," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 2, pp. 1–20, 2017.

[15] N. Arora, A. Mirzazadeh, I. Moon, C. Ramey, Y. Zhao, D. C. Rodriguez, G. D. Abowd, and T. Starner, "Mars: Nano-power battery-free wireless interfaces for touch, swipe and speech input," in *The 34th Annual ACM Symposium on User Interface Software and Technology*, 2021, pp. 1305–1325.

[16] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," *Advances in neural information processing systems*, vol. 33, pp. 3846–3857, 2020.

[17] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, P. Smaragdis, and A. Kumar, "Remixit: Continual self-training of speech enhancement models via bootstrapped remixing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1329–1341, 2022.

[18] C. Xu, Z. Li, H. Zhang, A. S. Rathore, H. Li, C. Song, K. Wang, and W. Xu, "Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019, pp. 14–26.

[19] Z. Ba, T. Zheng, X. Zhang, Z. Qin, B. Li, X. Liu, and K. Ren, "Learning-based practical smartphone eavesdropping with built-in accelerometer." in *NDSS*, vol. 2020, 2020, pp. 1–18.

[20] Z. Wang, Z. Chen, A. D. Singh, L. Garcia, J. Luo, and M. B. Srivastava, "Uwhear: Through-wall extraction and separation of audio vibrations using wireless signals," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 1–14.

[21] C. Wang, L. Xie, Y. Lin, W. Wang, Y. Chen, Y. Bu, K. Zhang, and S. Lu, "Thru-the-wall eavesdropping on loudspeakers via rfid by capturing sub-mm level vibration," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–25, 2021.

[22] P. Hu, H. Zhuang, P. S. Santhalingam, R. Spolaor, P. Pathak, G. Zhang, and X. Cheng, "Accear: Accelerometer acoustic eavesdropping with unconstrained vocabulary," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1757–1773.

[23] P. Hu, Y. Ma, P. S. Santhalingam, P. H. Pathak, and X. Cheng, "Milliear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 11–20.

[24] S. Basak and M. Gowda, "mmspy: Spying phone calls using mmwave radars," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1211–1228.

[25] C. Wang, F. Lin, T. Liu, Z. Liu, Y. Shen, Z. Ba, L. Lu, W. Xu, and K. Ren, "mmphone: Acoustic eavesdropping on loudspeakers via mmwave-characterized piezoelectric effect," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 820–829.

[26] Y. Feng, K. Zhang, C. Wang, L. Xie, J. Ning, and S. Chen, "mmeavesdropper: Signal augmentation-based directional eavesdropping with mmwave radar," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 2023, pp. 1–10.

[27] X. Xu, Y. Chen, Z. Ling, L. Lu, J. Luo, and X. Fu, "mmear: Push the limit of cots mmwave eavesdropping on headphones," in *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*. IEEE, 2024, pp. 351–360.

[28] A. Onishi, S. H. Bhupathiraju, R. Bhatt, S. Rampazzi, and T. Sugawara, "Sound of interference: Electromagnetic eavesdropping attack on digital microphones using pulse density modulation," in *34th USENIX Security Symposium (USENIX Security 25)*, 2025, pp. 3865–3884.

[29] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[30] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[31] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2020, pp. 1–6.

[32] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *arXiv preprint arXiv:2005.13981*, 2020.

[33] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 687–10 698.

[34] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[35] J. Zhang, C. Zorila, R. Doddipatla, and J. Barker, "Teacher-student mixit for unsupervised and semi-supervised speech separation," *arXiv preprint arXiv:2106.07843*, 2021.

[36] E. Karamatlı and S. Kırbız, "Mixcycle: Unsupervised speech separation via cyclic mixture permutation invariant training," *IEEE Signal Processing Letters*, vol. 29, pp. 2637–2641, 2022.

[37] K. Saijo and T. Ogawa, "Self-remixing: Unsupervised speech separation via separation and remixing," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.