UWHear: Through-wall Extraction and Separation of Audio Vibrations Using Wireless Signals

Ziqi Wang
University of California, Los Angeles
wangzq312@g.ucla.edu

Zhe Chen
Nanyang Technological University
chen.zhe@ntu.edu.sg

Akash Deep Singh
University of California, Los Angeles
akashdeepsingh@g.ucla.edu

Luis Garcia
USC Information Sciences Institute
lgarcia@isi.edu

Jun Luo
Nanyang Technological University
junluo@ntu.edu.sg

Mani B. Srivastava
University of California, Los Angeles
mbs@ucla.edu

ABSTRACT

An ability to detect, classify, and locate complex acoustic events can be a powerful tool to help smart systems build context-awareness, e.g., to make rich inferences about human behaviors in physical spaces. Conventional methods to measure acoustic signals employ microphones as sensors. As signals from multiple acoustic sources are blended during propagation to a sensor, such methods impose a dual challenge of separating the signal for an acoustic event from background noise and from other acoustic events of interest. Recent research has proposed using radio-frequency (RF) signals, e.g., Wi-Fi and millimeter-wave (mmWave), to sense sound directly from source vibrations. Whereas these works allow separating an acoustic event from background noise, they cannot monitor multiple sound sources simultaneously. In this paper, we present **UWHear**, a system that simultaneously recovers and separates sounds from multiple sources. Unlike previous works using continuous-wave RF, UWHear employs Impulse Radio Ultra-Wideband (IR-UWB) technology, in order to construct an enhanced audio sensing system tackling the above challenges. Further, IR-UWB radios can penetrate light building materials, which enables UWHear to operate in some non-line-of-sight (NLOS) conditions. In addition to providing a theoretical guarantee for audio recovery using RF pulses, we also implement an audio sensing prototype exploiting a commercial-off-the-shelf IR-UWB radar. Our experiments show that UWHear can effectively separate the content of two speakers that are placed only 25cm apart. UWHear can also capture and separate multiple sounds and vibrations of household appliances while being immune to non-target noise coming from other directions.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Computer systems organization** → **Sensors and actuators**.

KEYWORDS

RF sensing, IR-UWB Radar, Wireless Vibrometry, Audio Sensing.

SenSys '20, November 16–19, 2020, Virtual Event, Japan

ACM ISBN 978-1-4503-7590-0/20/11.
https://doi.org/10.1145/3384419.3430772

ACM Reference Format:
Ziqi Wang, Zhe Chen, Akash Deep Singh, Luis Garcia, Jun Luo, and Mani B. Srivastava. 2020. UWHear: Through-wall Extraction and Separation of Audio Vibrations Using Wireless Signals. In *The 18th ACM Conference on Embedded Networked Sensor Systems (SenSys '20), November 16–19, 2020, Virtual Event, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3384419.3430772

1 INTRODUCTION

Computational analysis of sound events and scenes has emerged as an important element of many sensing applications. Robust sound event detection and classification (SEDC) requires an accurate recognition and estimation of the onset and the endpoints of individual sound events in an audio signal [35]. SEDC is critical to exploiting acoustic modality for providing enhanced context-awareness in physical spaces [1], e.g., monitoring acoustic events in smart home and buildings [24, 38, 76], urban city acoustic surveillance and noise source identification [70], as well as audio-based activity recognition to provide care for elderly or disabled [13, 58, 61]. While effective analysis of acoustic scenes and events depend on a confluence of many different technologies, a particularly important role is played by the transducer hardware used to capture the acoustic energy and transform it into an electronic signal. Conventionally the transducer of choice is a microphone that converts mechanical sound pressure waves into electrical signals making use of electromagnetic, electromechanical, or capacitive phenomena [45].

As these microphones capture the overall sound pressure wave at their location, their output is a blend of the sounds arriving simultaneously from spatially separated sources in the environment, making it hard for the SEDC system to isolate any individual sound event. For example, if a vacuum cleaner and a blender are running simultaneously at different locations in a room, a microphone will output a signal that corresponds to a mixture of the two sounds in both time and frequency domains. More generally, besides the sound from the source of interest, a microphone captures all sorts of sounds present in the environment, including both background noises (e.g., traffic, airplane engine) and sounds from other independent sources of interest. This mixing of various sounds presents a dual challenge for the downstream sensor information processing tasks: in addition to filtering the background noises, separating the concurrently occurring acoustic events at multiple sources remains an issue, as depicted in the top portion of Figure 1. Previous microphone-based works have proposed using complicated infrastructure to solve the sound separation problem,

such as a large-scale or scattered microphone array [23, 43, 62]. Yet the background noise cannot be entirely eliminated [54]. Therefore, directly sensing sound from its source could provide a new perspective to tackle the challenges.

Following the line of sensing sound directly from its source, recent proposals have sought to augment audio sensing using wireless vibrometry [15, 65, 68, 76]. A wireless vibrometer sends out radio-frequency (RF) probe signals that will be reflected by the vibrating surface of a sound source. It subsequently analyzes these reflections to recover the source vibrations and to reconstruct the sound. To ensure satisfying performance in noisy environments, these RF audio sensing systems employ highly directional beams and attempt to isolate a source of vibration in the angular domain using laser beams [76] or antenna arrays for beamforming [65, 68]. However, this isolation directs the "attention" of a sensor to a single target, constraining the sensor data to a **one-dimensional** time series. While highly directional beams minimize the impact of background noises, they *do not scale to monitoring multiple sources* within reasonable system complexity. Moreover, some RF modalities (e.g., Wi-Fi and RFID) employed fail to achieve a fine-grained *ranging-resolution* due to their limited bandwidth. So they cannot differentiate acoustic events spatially close to each other [65, 69]. Increasing ranging-resolution requires an ultra-wide bandwidth that is often realized at a high-frequency range with very short wavelength [47]. However, signals at such a high-frequency suffer from a sharp propagation loss and poor penetration ability, resulting in a "visibility" issue: the mmWave and laser-based systems can only be operated in line-of-sight (LOS) conditions [49, 68, 76].

We propose UWHear, a fine-grained audio sensing system that is capable of identifying multiple sources simultaneously, resilient to background noise, and robust in non-line-of-sight (NLOS) scenarios. UWHear exploits the Impulse Radio Ultra-Wideband (IR-UWB) radar to enhance the process of sound recovery in challenging environments. Unlike other RF sensing systems that transmit continuous waves, IR-UWB radars send very short pulses in the time domain while occupying a wide frequency bandwidth; this wide bandwidth guarantees a fine ranging resolution. Essentially, for every transmitted probe pulse, the IR-UWB receiver may collect a number of reflected pulses. Consequently, sound sources can be well separated by accurately estimating the Time-of-Flight (ToF) of the reflected pulses. This procedure is repeated rapidly with constant intervals to produce **two-dimensional** data, i.e., multiple time series retrieved from different distance ranges, as depicted in the bottom portion of Figure 1. Working in a sub-10 GHz band, IR-UWB radars additionally possess the capability of penetrating light building materials. In other words, they strike a good balance between signal penetration ability and ranging resolution, which is particularly critical for identifying NLOS sound sources. Finally, UWB's transmission power is limited, which ensures co-existence with other communication schemes in the same frequency band, such as WiFi and Bluetooth. IR-UWB is well known for its low peak pulse output power, typically less than 6 dBm [7].

We implement UWHear using a commercial-off-the-shelf (COTS) IR-UWB radar, and test its capabilities both qualitatively and quantitatively. We evaluate UWHear with respect to the challenges faced by the current technologies that we highlighted earlier. UWHear is capable of (i) through-wall sensing of audio vibrations , (ii) recovering and separating the sounds from two sources placed as close as 25 cm in distance without any cross-interference, and (iii) retrieving the sound from a number of real-world household tools such as a vacuum cleaner and a hand drill. Our major contributions in this work are as follows:

- To the best of our knowledge, UWHear is the first work to investigate the possibility and the benefits of extracting audio from IR-UWB radar responses.
- We provide a theoretical analysis on performing audio sensing using non-continuous, impulse-based wireless signals.
- We implement UWHear using a COTS IR-UWB radar sensor with optimal driver settings, as well as a pure statistical signal processing pipeline.
- We demonstrate UWHear's capability to deal with multiple target sounds simultaneously in both qualitative and quantitative manners. We also test the limits of this system and show it can successfully perform through-wall audio sensing.

The rest of this paper is organized as follows. We discuss related literature in Section 2. In Section 3, we describe the theory behind using IR-UWB radar for audio sensing mathematically. In Section 4, we provide the detailed system design for UWHear. We evaluate UWHear's performance on audio sensing as well as sound source separation in Sections 5 and 6, respectively. Finally, we conduct some discussions in Section 7 along with a conclusion in Section 8.

2 RELATED WORK

In this section, we summarize a representative set of related works that focus on sound separation and denoising, wireless vibrometry, and applications of UWB devices.

2.1 Sound Separation and Denoising

In prior works researchers have devoted considerable efforts towards addressing the challenge of separating multiple target sounds and non-target noises that are fused in the same signal. Microphone arrays are frequently used to perform sound separation and localization. These systems typically apply time-difference-of-arrival (TDoA) beamforming and triangulation to identify the sound of interest and locate its source. Some of these works employ spatially dispersed microphone arrays. For example, [62] and [23] use a distributed wireless sensor network. These systems require the construction of infrastructures. Some other works use carefully designed geometric microphone array shapes, including circular [40], cubical [59] and three-ring [54]. The ARL PXI Tetrahedral acoustic microphone array [43] also has a complicated rigid skeleton. Some small microphone array implementations are now available, such as the Matrix Creator [30] and Amazon Echo [4]. However, they primarily address the scenario of a single dominant source. The authors of [53] propose using perpendicular cross-spectra fusion (PCSF) to reconcile the direction-of-arrival (DOA) estimation derived from different algorithms. Large array skeletons may also suffer from the inconvenience of the infrastructure. Meanwhile, smaller arrays assume the sound wave as a far-field signal and can only give an angular estimation of the sound sources. Therefore, they may experience problems dealing with multiple sounds from one direction.

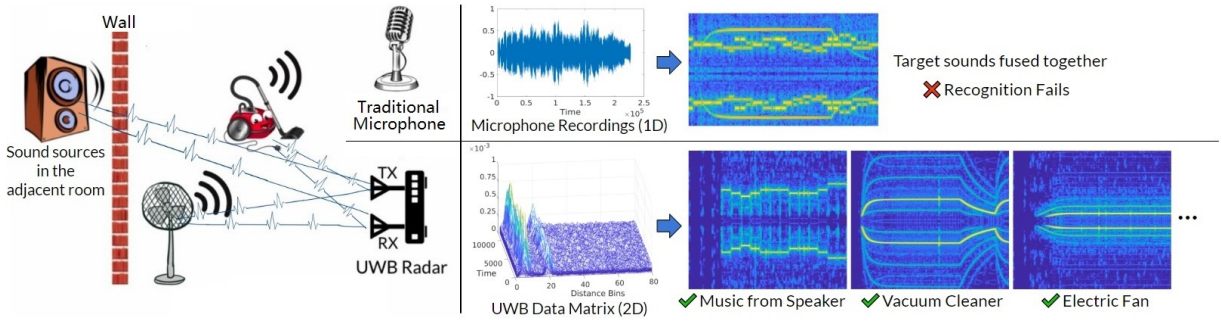

Figure 1: Overview of UWHear. Whereas even the up-to-date blind separation algorithms [66] fail to separate the one-dimensional acoustic audio recording, the two-dimensional IR-UWB sensing data enable UWHear to separate and recover sound events, leveraging its fine ranging resolution offered by the wide bandwidth of IR-UWB radars.

Another subset of prior research focuses on pure software-based solutions for audio denoising and separation [39, 50]. In particular, universal sound separation and denoising using deep learning has drawn significant attention in recent years. Exemplary works include [22, 27, 57] using convolutional neural networks (CNN), [44] using a WaveNet, and [20] employing a generative adversarial network (GAN). [36] uses recurrent neural network blocks to map the mixed sound into a latent space and can then separate five simultaneous speakers. However, machine learning models depend heavily on the completeness of training data, and it is virtually impossible to train a model with all the potential sound activities and noise in domestic environments.

2.2 Wireless Vibrometry

Wireless vibrometry refers to the technique of sensing vibration-related information using wireless signals. For example, Tagbeat [69] uses RFID tags to identify mechanical vibrations periods of spinning targets. It can troubleshoot automobile engines and can even monitor the shaking of blood samples in a high-speed centrifuge. [75] employs commercial WiFi signals to detect human breath status. [37] leverages frequency modulated continuous wave in ultrasound frequency to detect chest movements for sleep apnea assessments.

Mechanical vibrations whose frequencies lie in the auditory range create audible sounds. Several works have emerged in recent years, showing the ability to actively discover sound activities from the vibrations at the sound source using wireless vibrometry. [15] employs visible light to recover audio from vibrating objects (such as an empty potato chip bag) using a high-speed camera. Wi-Fi signals are also used for audio sensing. The channel state information in Wi-Fi carries hints of all kinds of movements, including fine-grained vibrations due to the micro-Doppler effect and the multi-path effect. [65] presents a through-wall eavesdropping system, where Wi-Fi signals generated by software-defined radio (SDR) are exploited to recover sound produced by loudspeakers.

Although some of the systems may raise privacy concerns, wireless vibrometry can also benefit our daily life. For example, VibroSight [76] employs lasers to detect the vibrations of household appliances. They attach retro-reflective tags on top of those appliances and shine a laser beam on those tags to retrieve target sound. The system may function as a central hub to document the usage of smart home appliances and help understand human behavior or save energy. In WaveEar [68], the authors create a Voice-User Interface (VUI) using mmWave radar. They use beamforming technology to focus on the throat of the target user and then use a U-shaped deep neural network to recover the voice from the signal. These methods, while capable of recovering sound activities in noisy environments, are not designed to deal with multiple activities or work in NLOS conditions.

2.3 UWB Devices and Their Applications

An UWB radio, by definition, is a radio whose operating frequency occupies a bandwidth more than 500 MHz. Today, UWB devices are widely used in lower-power communication systems [18, 29, 34].

In addition, UWB radars such as the Decawave DW1000 [21, 25] have increasingly been used for other tasks such as imaging [28], localization and tracking [3, 46, 74], material identification [17], and health monitoring [33, 60]. "Human presence sensor" created by Novelda [6] uses an IR-UWB radar to detect human presence and can be used to save energy in smart buildings. [42] deploys UWB beacons in augmented reality (AR) settings, and uses Time-of-Flight (ToF) ranging to provide localization for multi-user AR systems. V^{2}iFi [77] employs an IR-UWB radar to simultaneously monitor the vital signs of car drivers, including breath and heart rate.

3 AUDIO SENSING VIA IR-UWB

3.1 Intuitions for UWB Acoustic Sensing

IR-UWB radar operates by sending pulses and collecting responses. The super-short pulse duration of IR-UWB enables the use of time-of-flight for ranging tasks, which leads to the sound source separation capability. Previous modalities collect a **1D audio sample** from their target, and the only dimension here is time. IR-UWB radar, however, generates a **2D matrix**. For better understanding, we introduce the data structure of UWHear as follows.

The data collected from the departure of the probe pulse to the arrival of the last response is called a *frame*. All the frames are ordered chronologically, and Figure 2 shows an example of this data structure. The frames are placed along the Y-axis (the *slow time*). On the X-axis (*fast time*), we have reflective pulse responses with different time delays. Since the fast time denotes the round trip ToF of a pulse, we can convert the fast time into *distance bins*.

Suppose multiple targets are lying at different distances. We can separate them by fixating the fast time to a few particular values

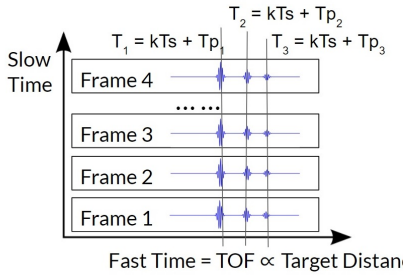

Figure 2: Illustration of the fast time and slow time.

representing the round-trip ToFs from the sensor to the sound sources. In other words, we find the distance bin corresponds to each sound source and take a slice from the 2D matrix to get a 1D time series as an estimation of that sound source. For example, in the lower half of Figure 1, we can see that two sound sources leave separate series of traces on the 2D matrix, and can be separated in the distance (*fast time*) domain.

However, it takes further processing to translate the time-domain slice into an audible sound. Even though IR-UWB provides a higher ranging resolution than other wireless modalities (e.g., Wi-Fi), it is still not possible to detect millimeter-level displacement caused by sound vibration directly with ToF estimation. Nonetheless, while non-trivial, it is possible to show that using the complex baseband equivalent processing in IR-UWB radar one can perform the sensing of sound-related vibrations. In the next section, our theory proves that the sound waveform values are proportional to the amplitude of the in-phase or quadrature part of the filtered sliced data.

3.2 IR-UWB Audio Sensing Theory.

The baseband equivalent representation of our IR-UWB radar system is shown in Figure 3. The notion of time t within one frame corresponds to the ToF of the signal pulse, which is also known as *fast time*. Meanwhile, IR-UWB sequentially transmits probe pulses with interval T_s, which can be treated as the sampling rate on the *slow time* (t_{slow}) axis. Usually, *fast time* is fine-grained in tens of picoseconds, while *slow time* has the scale of hundreds of microseconds.

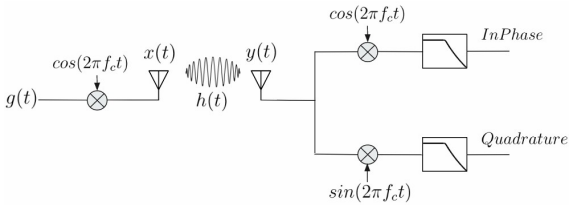

Figure 3: An IR-UWB radar system in equivalent baseband representation.

The IR-UWB radar transmits Gaussian pulses $g(t)$ modulated on a carrier frequency f_c, mathematically represented as

$$x(t) = g(t - kT_s)\cos\left(2\pi f_c(t - kT_s)\right), \tag{1}$$

where T_s is the pulse repetition rate, and the baseband Gaussian pulses $g(t)$ are given in [5] as

$$g(t) = V_{\text{tx}} \exp\left(-2\pi^2 f_B^2 log_{10}(e)\, t^2\right), \tag{2}$$

where f_B denotes the -10 dB bandwidth and V_{tx} is the maximum amplitude of the Gaussian pulse. The pulse sequence is sent out to interact with the objects in the environment and received by the receiving antenna. Note that in reality, the transmitting antenna and the receiving antenna are co-located. The channel frequency response in an indoor environment can be characterized as a summation of P paths with different time delays and attenuations:

$$h(t) = \sum_{p=1}^{P} \alpha_p \delta\left(t - T_p - T_p^D(t)\right), \tag{3}$$

where T_p is the round-trip ToF determined by the sound source distance. $T_p^D(t)$ is the time-varying delay caused by minute target movement, such as the cone being pushed back and forth by the coil in any speaker, but for static objects, $T_p^D(t) = 0$. Our goal in wireless audio sensing is to recover the $T_p^D(t)$, which can be translated into the sound. The movements of sound sources, if any, are much slower compared to its vibrations, and can be safely ignored in a short time window.

The received signal $y(t)$ can be modeled as a convolution of the transmitted signal and the channel impulse response, plus additive noise, i.e.,

$$y(t) = x(t) * h(t) + n(t) = \sum_{p=1}^{P} \alpha_p g\left(t - kT_s - T_p - T_p^D(t)\right) \times \cos\left(2\pi f_c\left(t - kT_s - T_p - T_p^D(t)\right)\right) + n(t).$$

On the receiver side, the received signal $y(t)$ is downconverted. Then, $y(t)$ is multiplied with the carrier frequency in a mixer, and passed through a low-pass filter. Here, we take the in-phase branch as an example. Looking at the cosine part only, we have:

$$\begin{aligned} m(t) &= \cos\left(2\pi f_c(t - kT_s - T_p - T_p^D(t))\right) \times \cos\left(2\pi f_c(t - kT_s)\right) \\ &= \frac{1}{2}\left[\cos\left(2\pi 2 f_c\left(t - kT_s - \frac{T_p}{2} - \frac{T_p^D(t)}{2}\right)\right)\right. \\ &+ \left.\cos\left(2\pi f_c\left(T_p + T_p^D(t)\right)\right)\right]. \end{aligned}$$

The $2f_c$ frequency term is filtered out, leaving the low-frequency $\frac{1}{2}\cos\left(2\pi\left(T_p + T_p^D(t)\right)\right)$ term only. Based on this, we can rewrite the in-phase baseband signal after down-conversion and filtering as:

$$\begin{aligned} y_{\text{in-phase}}(t) &= \text{LPF}\left[y(t)\cdot\cos\left(2\pi f_c(t - kT_s)\right)\right] \\ &= \frac{1}{2}\sum_{p=1}^{P} \alpha_p g\left(t - kT_s - T_p - T_p^D(t)\right) \times \\ &\quad \cos\left(2\pi f_c\left(T_p + T_p^D(t)\right)\right) + \tilde{n}(t). \end{aligned} \tag{4}$$

Similarly, we can have the quadrature component obtained via down-conversion and filtering as:

$$\begin{aligned} y_{\text{quad}}(t) &= \text{LPF}\left[y(t)\cdot\sin\left(2\pi f_c(t - kT_s)\right)\right] \\ &= \frac{1}{2}\sum_{p=1}^{P} \alpha_p g\left(t - kT_s - T_p - T_p^D(t)\right) \times \\ &\quad \sin\left(2\pi f_c\left(T_p + T_p^D(t)\right)\right) + \tilde{n}(t). \end{aligned} \tag{5}$$

The target ToF T_p can be translated into target distances when multiplied with the speed of light. As different targets have different T_p's, we can select out any sound source k by setting $t = kT_s + T_{p_k}$ to perform **sound source separation**. For those paths without audio-related movement whose $T_p^D(t) = 0$, the response $y(t = kT_s + T_p)$ ideally will not change over slow time. We can filter those static responses out by applying a static clutter suppression algorithm that will be introduced in Section 4.

Supposing the sound-related vibration is captured in path p_0, we can isolate the received signal from such a path by setting $t = t_p = kT_s + T_{p_0}$, which can be written as

$$y_{\text{in-phase}}(t_p) = \frac{1}{2}\alpha_{p_0} g\left(T_{p_0}^D(t_p)\right)\cos\left(2\pi f_c T_{p_0} + 2\pi f_c T_{p_0}^D(t_p)\right) + \tilde{n}(t_p),$$

$$y_{\text{quad}}(t_p) = \frac{1}{2}\alpha_{p_0} g\left(T_{p_0}^D(t_p)\right)\sin\left(2\pi f_c T_{p_0} + 2\pi f_c T_{p_0}^D(t_p)\right) + \tilde{n}(t_p).$$

We can estimate the scale of $T_{p_0}^D(t)$. Suppose the sound-related displacement is 2mm, and the UWB carrier frequency is 7.3GHz

$$\max\left(T_{p_0}^D\right) = \frac{d}{c} = \frac{2\times 10^{-3}}{3\times 10^{-8}} = 6.67\times 10^{-12}, \quad (6)$$

$$\max\left(2\pi f_c T_{p_0}^D\right) = 6.67\times 10^{-12}\times 2\times\pi\times 7.3\times 10^{9} = 0.305. \quad (7)$$

Because both values are minimal, we can use Maclaurin series to expand $g(t)$ around $g(0)$ and ignore the 2^{nd} and higher-order terms:

$$g(t) = g(0) + g'(0)t + o(t^2) = V_{\text{tx}} + 0\cdot t + o(t_2) = V_{\text{tx}} + o(t^2). \quad (8)$$

Plugging Equation (8) into $y_{\text{in-phase}}$ and y_{quad}, and ignoring high order infinitesimals as well as noise, we get the form of

$$y_{\text{in-phase}}(t_p) = \frac{1}{2}\alpha_{p_0} V_{\text{tx}}\cos\left(2\pi f_c T_{p_0} + 2\pi f_c T_{p_0}^D(t_p)\right), \quad (9)$$

$$y_{\text{quad}}(t_p) = \frac{1}{2}\alpha_{p_0} V_{\text{tx}}\sin\left(2\pi f_c T_{p_0} + 2\pi f_c T_{p_0}^D(t_p)\right). \quad (10)$$

By Taylor expansion, $f(t) = \sum_{n=0}^{\infty}\frac{f^{(n)}(t_0)}{n!}(t-t_0)^n$, where $f^{(n)}(t_0)$ is the n-th derivatives of $f(t)$ at t_0, we know that

$$\sin(t) = \sin(0) + \cos(0)t + o(t^2) \approx t, |t| < \epsilon$$

$$\cos(t) = \cos(\frac{\pi}{2}) - \sin(\frac{\pi}{2})t + o(t^2) \approx -t, |t-\frac{\pi}{2}| < \epsilon$$

$$\sin(t) = \sin(\pi) + \cos(\pi)t + o(t^2) \approx -t, |t-\pi| < \epsilon$$

$$\cos(t) = \cos(\frac{3\pi}{2}) - \sin(\frac{3\pi}{2})t + o(t^2) \approx t, |t-\frac{3\pi}{2}| < \epsilon$$

Here ϵ is a small value marking the vicinities around 0, $\frac{\pi}{2}$, π, and $\frac{3\pi}{2}$. In Equation (6), we already show that $2\pi f_c T_{p_0}^D(t_p)$ is a very small number. While the constant $2\pi f_c T_{p_0}$ is very large, mod $(2\pi f_c T_{p_0}, 2\pi)$ is going to put the component inside the sine or cosine of Equation (9) and (10) near one of the four vicinities above. Without the loss of generality, we assume mod $(2\pi f_c T_{p_0}, 2\pi) \approx 0$, then

$$y_{\text{quad}}(t_p) = \frac{1}{2}\alpha_{p_0} V_{\text{tx}} 2\pi f_c T_{p_0}^D(t_p) = \frac{1}{2}\alpha_{p_0} V_{\text{tx}}\frac{2}{c} d_{p_0}^D(t_p) 2\pi f_c$$
$$= \alpha_{p_0} V_{\text{tx}}\frac{2}{c}\pi f_c d_{p_0}^D(t_p),$$

where $d_{p_0}^D(t)$ is the sound source (e.g. speaker diaphragm) displacement, and c is the speed of light. It is clear that the amount of target micro displacement is linearly proportional to the amplitude of the quadratic part of the receiving signal. In other cases, it will be linearly proportional to the amplitude of the in-phase part. Note that $t_p = kT_s + T_{p_0}$ is a function of the frame number k, and $d_{p_0}^D$ changes over slow time. For example, if a sine wave single tone (f_{music}) sound is played, then the $d_{p_0}^D$ should be modeled as,

$$d_{p_0}^D(t_{\text{slow}}) = \max\left(d_{p_0}^D\right)\times\sin\left(2\pi f_{\text{music}} t_{\text{slow}}\right).$$

We can treat $d_{p_0}^D(t_p) = d_{p_0}^D(kT_s + T_p)$ as the speaker movement $d_{p_0}^D$ being sampled at interval T_s, i.e., sampled at the UWB frame rate. As this causes $y_{\text{quad}}(t_p)$ or $y_{\text{in-phase}}(t_p)$ to be proportional to $d_{p_0}^D(t_p)$, we now conclude that we can **recover the sound-related movement** from the amplitude of UWB in-phase or quadrature data, whichever gives a higher signal quality.

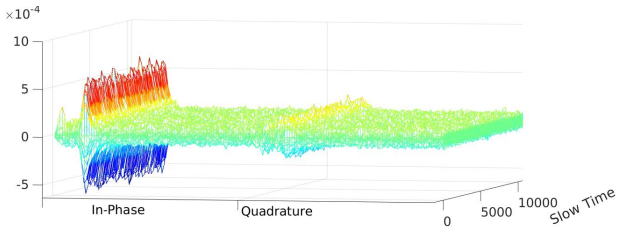

Figure 4: 3D visualisation of the baseband data amplitude collected with a speaker.

Figure 4 shows an example of real-world data collected using IR-UWB radar and a speaker, after all the processing to be introduced in Section 4. It is clear that most of the sound-related fluctuations appear in the amplitude of the in-phase data. Previous works in IR-UWB radar sensing like [77] often use the amplitude or phase of the I/Q data for further processing. However, in the audio sensing scenario, the old practice may incur some problems. In the most extreme case, where the sound-related variance only appears on the amplitude of the in-phase (or quadrature) data, calculating $\sqrt{y_{\text{in-phase}}^2 + y_{\text{quad}}^2}$ will introduce undesired $2f_{\text{music}}$ components and defeat the purpose of reliable audio sensing.

To summarize, we have shown that one can extract the sound-related vibration information by analyzing the amplitude of the in-phase or quadrature of UWB receiving signal, whichever gives a higher signal-to-noise-ratio. Since our work aims to extract the vibration related information from the IR-UWB radar sensor readings, the analysis in this section provides theoretical support to achieve this goal.

4 UWHEAR: DESIGN AND IMPLEMENTATION

4.1 System Overview

Having formulated a theoretical basis for using IR-UWB radar to recover sound, we now build a real-world system from a commercial-off-the-shelf IR-UWB radar board, and implement a data processing pipeline to put the theory into practice. Figure 5 gives an overview of UWHear, our UWB audio sensing system.

UWHear uses an IR-UWB radar that sends out pulses at a constant rate, collects the reflected impulses, and downconverts the radio frequency data to the baseband I/Q data. The I/Q data is then analyzed with our processing pipeline which consists of several algorithmic modules. First, to battle the phase variations caused

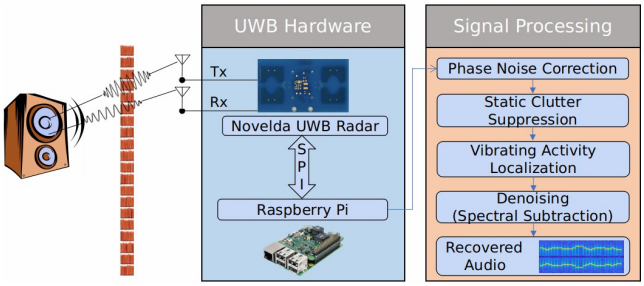

Figure 5: Design blocks of UWHear.

by sampling clock jitters, we employ the Phase Noise Correction algorithm. Then Static Clutter Suppression removes the reflections caused by static objects like walls and furniture. As we have analyzed in Section 3, the sound-related information will appear on the amplitude of the real or imaginary part of the I/Q data. Therefore, we juxtapose the in-phase part and the quadrature part. Since the reflected pulses have various ToFs corresponding to a wide distance range, it is crucial to locate the distance bins where vibrations happen using the Vibrating Target Localization module. Finally, we can obtain the recovered sound with further denoising, such as a spectral subtraction algorithm. Then we can have a recovered sound for further processing, e.g., sound classification or speech recognition.

4.2 Hardware and Drivers

Our system is implemented with Novelda Xethru X4M05 IR-UWB radar board combined with a Raspberry Pi 3B+. Figure 6(a) shows the system hardware stack. The blue board is the IR-UWB radar transceiver, and it is connected with the Raspberry Pi using a self-made connector board. The connection between the Pi and the radar is realized via an SPI interface.

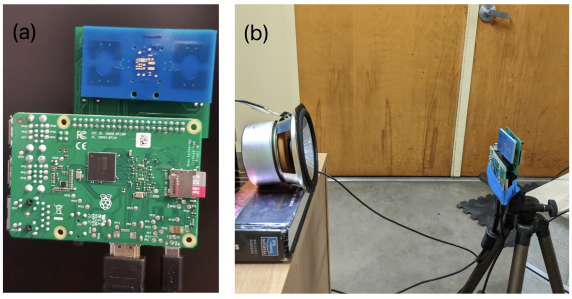

Figure 6: Hardware components of UWHear: (a) Hardware appearance (b) Proof-of-concept experiment setup.

4.2.1 UWB Data Collection. The X4M05 radar board consists of an X4A02 Antenna board and a Novelda X4 impulse radar transceiver System on Chip (SoC). According to its datasheet [7], the IR-UWB radar operates at a center frequency of 7.29GHz with a bandwidth of 1.4GHz.

In our IR-UWB radar transmitter hardware, the Gaussian pulses are modulated on a sub-10 GHz carrier frequency. At the receiver side, a digital down-conversion is performed on the received Radar Frame (RF) data inside the X4 SoC to retrieve the baseband pulses, making each data point a complex double representing in-phase and quadrature (I/Q) baseband data. This down-conversion stage will decimate the RF data by a factor of 8. We can then calculate the distance between adjacent distance bins in the baseband data as

$$bb\_interval = \frac{LightSpeed}{2 \times SamplingRate} = \frac{2.998 \times 10^8 m/s \times 8}{2 \times 23.328 \times 10^9 Hz} = 0.0514\text{m},$$

where 23.328 Giga-Samples/s is the rate with which the RF data are sampled. Since the maximum length of the received Radar Frame (RF) data before the downconversion has 1536 bins, the maximum range of such a radar system becomes

$$max\_dist = 1536 \div 8 \times 0.0514 = 9.874\text{m}.$$

As a summary, the collected data is going to be a complex matrix with dimension $fast\_time \times slow\_time$. The fast time dimension indicates the target distance while the slow time dimension indicates the elapsed time.

4.2.2 Driver Settings. The Xethru radar driver is implemented based on [8], with modifications to enable faster data transfer and to strike a balance between sampling rate and signal-to-noise-ratio(SNR). In this section, we describe the major changes in the driver settings to enable audio sensing.

SPI clock. The X4 radar SoC receives configuration and sends data to Raspberry Pi using Serial Peripheral Interface Bus (SPI). Once the X4 radar SoC finishes a data frame, it raised an SPI interrupt so that the controller (in our case, Raspberry Pi) can read the data. Owing to the fact that the radar SoC only caches the last frame it received, the clock of SPI should be set higher to ensure that the data can be transported in time. We set the clock to 32MHz, which is the highest sampling rate that the GPIO Interface library for the Raspberry Pi would allow.

Transmission Power. The radar transceiver can operate at three different transmitting power settings, which are low(0.48 pJ/pulse), medium(1.47 pJ/pulse), high(2.65 pJ/pulse). In our experiment, we test on both the medium level and the high level, and they are both capable of audio sensing. A higher power level can increase the sensitivity and effective range of the system. However, these settings should be performed carefully to comply with FCC regulations.

Effective Range. As analyzed in Section 4.2.1, the maximum range of the IR-UWB radar can be as far as 9.87m. The minimum and maximum detection distance is subject to change in the driver settings to focus on a specific range. For example, in our experiments, we set the starting point to 0.3m so that the first few bins are discarded, since they are usually overfilled by crosstalks between the transmitting and receiving antennas.

DAC Settings and Sampling Rate. According to [5], X4 uses a swept-threshold sampling method. Because the pulse duration is so short that a standard DAC will never be fast enough, the Swept-Threshold Sampling method is adopted to address this problem. The received signal frame is compared against a threshold to generate one-bit values for all data points in this frame. The threshold will increase by one step before the response of the next repeated pulse comes. Due to the extremely high pulse repetition rate, the vibrating target can be approximated as static in such a short period, which

means that the repeated frames can be treated the same as the previous ones. Then after a certain number of frames, we can have a multiple-bit digital representation of the original analog frame. The procedure is denoted as one *iteration*. It is also possible to average multiple *iterations*, or to average multiple pulses during one step (increase *pulse-per-step*) to improve SNR. However, if these two knobs are set too high, the sampling rate will be limited. This relationship can be mathematically described as

$$FPS = \frac{PulseRepetitionFrequency \times DutyCycle}{Iterations \times PulsePerStep \times (DAC_{max} - DAC_{min} + 1)}.$$

By default, the Pulse Repetition Frequency is set to be 15.1875 MHz, $DAC_{max} = 1100$, and $DAC_{max} = 949$. Heuristically, we pick $Iterations = 20, PulsePerStep = 2$, and $FPS = 1.5$ kHz. Currently, due to the limitations of SPI transfer speed, the sampling rate cannot exceed 1.6kHz; otherwise a packet loss will be inevitable. The sampling rate is also adjustable and future works allowing a higher data transfer rate are desired to allow higher FPS.

The data are cached locally in the Raspberry Pi and then transferred to a desktop computer with AMD Ryzen 7 2700X processor for processing. Figure 6(b) demonstrates a typical setting of our proof-of-concept experiment. The IR-UWB radar system is mounted on a tripod and placed at a distance from the speaker. The speaker is connected with a cell phone to play the test tones.

4.3 Signal Processing Pipeline

The collected data are analyzed with our processing algorithms shown in Figure 5 that consist of a few modules: Phase Noise Correction that removes sampling clock jitters, Static Clutter Suppression that suppresses the reflections caused by static objects, and Vibration Activity Localization that determines the distances of the vibrating targets. Finally, we can acquire recovered audio after denoising and normalizing. We will introduce these modules separately in the following parts of this section.

4.3.1 Phase Noise Correction. The basic idea behind our work is to measure the amplitude change over time of the in-phase or quadrature data caused by source vibrations. However, many factors will block us from retrieving the information related to sound vibration, and one of those factors is the phase noise. Phase noise is introduced due to the imperfection of the signal sampling clock. These imperfections may include crystal defects and phase lock loop (PLL) error. Ideally, if we select out the data from one distance bin and analyze the phase over time, the phase should remain virtually the same supposing there are no vibrations at the current bin. However, with phase noise, one may still observe a rapid change of phase back and forth, which will then lead to the system mistakenly believe in the existence of a vibration in this bin or will cause distortion in the recovered sound. Figure 7 shows an example of the phase noise between adjacent UWB frames.

We perform phase noise correction following the method proposed in [9]. The insight here is that the signal amplitude in the first few distance bins is always high, which is due to the "crosstalks" between the transmitting antenna (tx) and the receiving antenna (rx), i.e., direct signal leakage from the tx to the rx. Our idea is that this crosstalk can be leveraged as a baseline for phase calibration. We first calculate the mean phase of bin 1 and use it as a standard

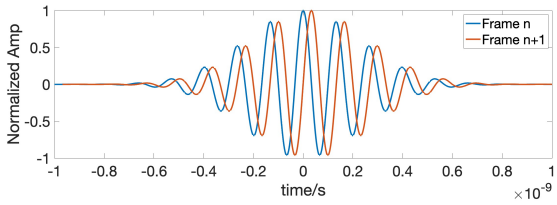

Figure 7: Example of phase noises between adjacent frames.

reference phase. For each frame i, we calculate the difference $\Delta\phi_i$ between the phase of its first element (i.e., bin 1 data) and the reference phase. Then we multiply all samples from the current frame with $e^{j\Delta\phi_i}$ to offset the phase error.

4.3.2 Static Clutter Suppression. While vibrations can create a unique pattern on the receiving data, static objects like walls and furniture will also reflect UWB pulses and create strong responses. As shown in Figure 8(a), the high peaks around bin 20 and bin 50 are the evidence of static clutters. The static responses are so strong that the useful signal is buried underneath. Luckily, the static clutter is usually time-invariant in a select bin. We apply a Butterworth finite impulse response filter (FIR) on each distance bin, with the stopping frequency at 20Hz and the passing frequency at 70Hz. To ensure zeros phase distortion at the beginning of the sequences, the FIR filtering is applied to input frame data in both the forward and reverse directions. The stop-band attenuation is set at -80dB.

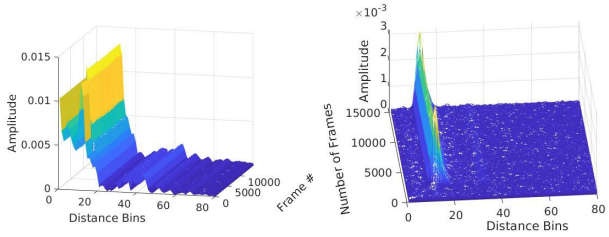

Figure 8: Results of static clutter removal. Left: (a) Raw data after phase noise correction. Right: (b) After static clutter removal.

Figure 8(b) shows the result after static clutter removal. The static peaks in Figure 8(a) are filtered out. Also, in the experiment shown in Figure 8, the sound lasts for about 12500 frames (8.3s), which is reflected in the peaks colored with green. From the filtered data, we can also see that the speaker is placed about 92.5 cm from the sensor (the ground truth is 100 cm) as we see time-varying patterns around bin 18. Our static clutter suppression filter is able to filter-out the low-frequency responses caused by human activities or chest motion related to breathing. In addition, due to the low-pass nature of UWB audio sensing (to be discussed in future sections), we also provide an option of doing pre-emphasis at this stage:

$$y(t) = x(t) - \alpha x(t-1),$$

where $\alpha \in (0.95, 1)$. This difference equation works as a high-pass filter to compensate for the signal loss in high-frequency ranges.

4.3.3 Vibrating Target Localization. UWB data contain multiple time series (columns) that correspond to different distance bins. As shown in previous cases, we may visually locate the vibrations in some cases. However, it is vital to select candidate bins with a high signal-to-noise-ratio (SNR). Since the signal is still pretty noisy in some channels, only doing thresholding or calculating variance in the time domain will not give satisfactory results.

We choose to solve this problem in the frequency domain. Our insight here is that, compared with noise, a channel (frames within a certain distance bin) with sound vibration information has a more concentrated spectrum than a noisy channel. For example, music will have basic notes and their higher-order harmonics. While human voice power is more widely distributed in the spectrum, we can still observe basic frequencies F_0 and their harmonics. Therefore, we firstly perform a Discrete Fourier Transform (DFT) over all channels to get their spectrums. Then the Herfindahl-Hirschman index (HHI) is used to calculate the concentration level of those spectrums. The Herfindahl-Hirschman index was introduced in economic fields as a measure of market concentration. It is calculated by squaring the "market share" of each frequency and then summing the resulting numbers. Here the "market share" is defined as the power of the current frequency divided by the overall power of the signal time series. The distance bins with the highest HHIs are selected as the candidates of bins containing vibration information. For slow moving objects like vacuum robots, the target is slowly moving between distance bins. We can perform vibrating target localization in short slow-time windows of 1s, and then stitch those time series of interest as the final output.

4.3.4 Denoising and Normalization. After locating the vibrating target, we can acquire an audio signal estimation by slicing that distance bin from the data. However, the recovered sound, while clearly audible, still contains non-negligible background noise which sounds like an air flow in traditional microphone recordings. This noise is the $n(t)$ ignored in Section 3. Our visual observation is that $n(t)$ is very close to an Additive White Gaussian Noise (AWGN).

For additive noise, a simple but powerful denoising solution is the spectral subtraction (SS). The underlying idea of SS is straightforward, and its typical flow chart is illustrated in Figure 9. Suppose the signal $x(t) = s(t) + d(t)$, where $s(t)$ is the signal part and $d(t)$ is the noise part. $x(t)$ is divided into overlapping frames. Then after fast Fourier transform, the spectrum of noise $\hat{D}(w)$ can be estimated and updated continuously using pure noise frames. What remains to do is to subtract the noise spectrum amplitude from the noisy signal, i.e.,

$$|\hat{X}(w)| = \sqrt{|X(w)|^2 - |\hat{D}(w)|^2}.$$

The spectrogram amplitude is then multiplied with the original phase to get an estimation of the clean signal $s(t)$, i.e., $\hat{x}(t)$.

Famous variants of the SS methods are linear SS [11], non-linear SS [10], and multi-band SS [26], whose implementations can be found in [71–73]. The spectral subtraction algorithm has some inherent problems, for example, music noise introduced by noise residuals. However, linear SS is sufficient for our case. The output of such a SS module is then normalized and output as a .wav file to generate the recovered sound. Also, we perform a Short Time Fourier Transform (STFT) to visualize the recovered sound.

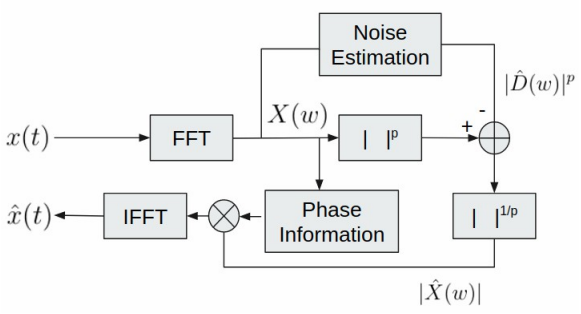

Figure 9: Flow chart of a typical spectral subtraction pipeline.

In the output of STFT, the X-axis stands for time while the Y-axis represents frequency.

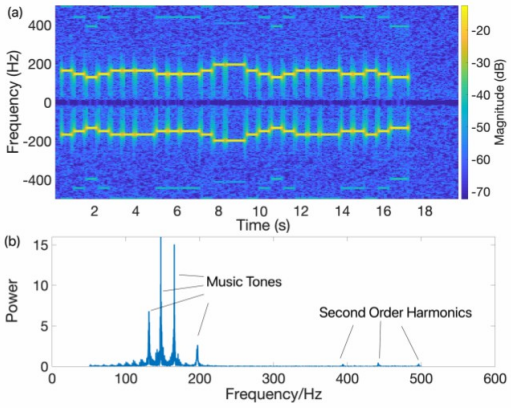

Figure 10: Results of the proof-of-concept experiments.

The results of the proof-of-concept experiments are shown in Figure 10, where we play a single tone song *Mary has a little lamb*. From the visualization, we can see that all the notes are recovered clearly. We also notice that in rare cases there is an interference of 60Hz and its multiples. This is probably due to the complicated power frequency electromagnetic field emitted by the circuit regulator or other devices. When this happens, an IIR comb filter can be applied to filter out the power frequency components. In the next section, we will explore and evaluate the capabilities of UWHear to perform audio sensing. Factors including distance, target frequency, sound source placement, and the through-wall propagation loss will be studied by field experiments.

5 UWHEAR SOUND SENSING PERFORMANCE

In this section, we aim to test the performance boundaries of UWHear as a sound sensor. The propagation of wireless signals is affected by a number of factors, including but not limited to distance, angle, and blockage. We evaluate the influence of four different factors using controlled experiments: i) distance between the sound source and the sensor, ii) through-wall penetration loss, iii) sound source placement angle, and iv) sound source frequency.

Experimental Setup. In order to better control the experimental variables and quantify the evaluation results, we use studio speakers as a sound source in this section, while in Section 6.4 we will test UWHear with a more natural setting using household appliances as sound sources. We play a single "C4" music tone whose frequency is 261.63 Hz. In terms of the evaluation metric, we use signal-to-noise-ratio (SNR) to measure the quality of the recovered sound, which is defined as

$$\text{SNR} = 10log_{10}(\frac{E_s}{E_n}),$$

where E_s and E_n are the energy of the signal and the noise, separately. In our experiments, we noticed that the recovered sound may have a slight frequency drift (0.5-3 Hz) from the test tone probably due to the sampling clock error. Thus, we estimate the power of the signal in the frequency domain by firstly localizing the peak in the spectrum near the target frequency (261.63 Hz) and then summing the energy in nearby frequency bins (within 5 Hz) as an estimation of signal energy, while using the remaining energy as an estimation of noise. Due to the fact that the noise may be time-varying, we employ a 1.5 second window with a 500 ms overlap to calculate the short-time SNR. Also, as we are using a single frequency probe signal for quantitative testing, using a filter-based denoising method may not be fair. In order to reflect the genuine noise characteristics of the hardware, the SNR data reported in the remaining part of this paper are acquired without the denoising stage of the the signal processing pipeline.

5.1 Distance Between Sound Source and Sensor

As we have analyzed in Section 4, the system suffers from additive noise close to Gaussian white noise. Meanwhile, it is a common knowledge that wireless signal strength will decay in space. The speaker volume is tuned to 79.3 dB/SPL at one meter distance measured by a microphone meter. The UWHear hardware is placed in front of the speaker at a distance starting from 50 cm and is evaluated at increments of 50 cm. At each distance we collect the data for 10 s, and analyse with a 1500 ms sliding window with a 500 ms overlap. Figure 11(a) shows a typical setting.

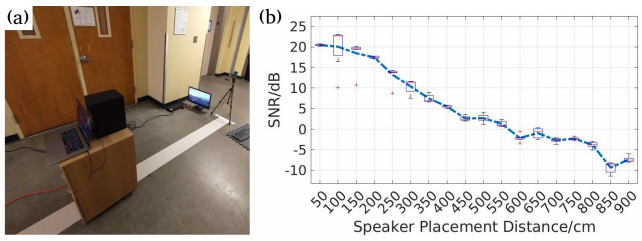

Figure 11: SNR vs speaker placement distance: (a) Experiment Setup, and (b) SNR plot across different distances.

Figure 11 shows the results. We notice that the SNR is decreasing almost linearly over distance following a typical wireless channel fading pattern. Medical research [31, 51] suggests that an audio with -5~0 dB SNR is still perceivable and understandable for the human ears. So we can see that the maximum detection range of UWHear is about 8 m. While it is slightly shorter than the range of Wi-Fi based work [65], the transmission power of UWHear (≤ 6 dBm [7]) is much lower than that of the software defined radio (≥ 20 dBm [14]) used in the previous work.

5.2 Through-wall Penetration Loss

We have previously hypothesized that the UWHear system can operate in non-line-of-sight (NLOS) scenarios, i.e., the hardware can recover the sound behind building materials. In this experiment, the speaker and the sensor are separated by a hollow wall (a wall made of wood and plaster between the bedroom and the living room) with a overall thickness of 11.5 cm. Similar to the previous experiment, we vary the distance of the speaker. The setting is shown in Figure 12(a), where the speaker is put inside the bedroom and the sensor is placed in the living room.

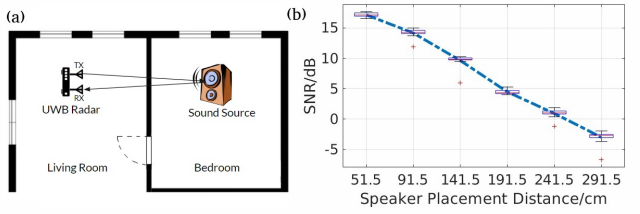

Figure 12: SNR vs through-wall speaker placement distance: (a) Experiment Setup, and (b) SNR plot across different distances.

Figure 12(b) displays the results of through-wall sound retrieving experiments. Generally, the SNR still follows a linear trend as distance is varied. Compared to that of free space, the through-wall results suffered a one-time loss of around 5 dB. Also, the slope of SNR dropping is slightly steeper than that of free space. For machine processing, the effective range may reduce since current models for audio perception are usually not very robust to noise [32]. Generally speaking, the system can operate through a wall within a range of 2.5 meters with reasonable performance degradation.

5.3 Sound Source Placement Angle

In reality, it is not practical to require the speaker or other sound sources to always be aligned with the sensor. Thus, it is necessary to understand the influence of the relative angle between the speaker diaphragm surface and the sensor Tx-Rx surface. We decompose this problem into two sub-problems. First, if the sound source is facing the sensor, but the sensor is pointing in another direction, then the recovered sound quality will be negatively affected. Currently, our sensor is equipped with a directional antenna whose 5dB main lobe is 50° both in elevation and azimuth. Thus, for this problem, we argue that this problem can be solved by aggregating multiple instances of UWHear, each covering a field of view.

Secondly, if the sensor beam is in the right direction, but the sound source is placed at a different angle, then the performance may vary. Intuitively, the incoming signal beam will experience diffuse reflection on the speaker cone, where a certain proportion of the signal will still be reflected back. Thus, we measure the effect of speaker placement angle, whose setting is shown in Figure 13(a). The distance is fixed at one meter and the speaker is rotated to a few certain angles. The speaker volume measured to be 74.5 dB/SPL at 1m distance.

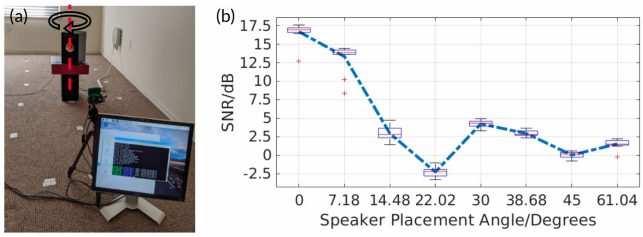

Figure 13: SNR vs speaker placement angle: (a) Experiment Setup, and (b) SNR plot across different angles.

From Figure 13(b), we can see that the SNR drops quickly after the speaker cone deviates over 10 degrees. The SNR fluctuates up and down and hovers at around 2.5dB. This observation coincides with our intuition: the fluctuation is a consequence of the speaker cone geometry. At some certain angle, the direct reflection will be stronger than other angles. The diffused reflection signal can still provide clues about the cone vibration. Also, it is worth noting that household appliances and tools typically have a more complicated and less directional geometric shape, which increases the percentage of power being reflected back to the radar receiver.

5.4 Sound Source Frequency

The final characteristic of the system that we want to test is the frequency response. Placing the speaker at 1m distance, we play test tones from 100Hz to 600Hz with increments of 100Hz. Our results in Figure 14 generally show a loss-pass trend, which provides a hint that we should use pre-emphasis in the signal processing to compensate for this low-pass nature.

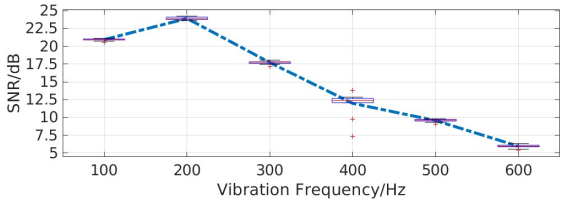

Figure 14: UWB audio sensing system frequency response.

6 UWHEAR SOUND SEPARATION PERFORMANCE

In the previous section, we explore the performance boundaries of UWHear. One of the greatest advantages of using IR-UWB lies in the fine ranging resolution brought by its ultra-wide signal bandwidth. With precise ToF estimation, it is capable of dealing with multiple targets at the same time. In this section, we will evaluate the system's sound separation performance in domestic environments.

6.1 Sound Source Distance Measurement

We have made hypothesis that, with fine ToF estimation, UWHear is capable to measure the distance from the sensor to the speaker precisely, and this capability will in turn support the sound separation functionality. To evaluate this capability, we aggregate the data from the first two experiments described in Sections 5.1 & 5.2 and estimate the speaker distance from the data. The empirical cumulative distribution function (CDF) plot of estimation error is shown in Figure 15.

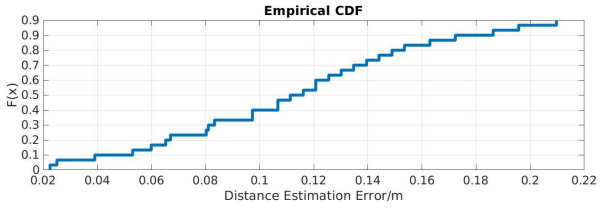

Figure 15: UWHear sound source distance estimation error: the empirical CDF curve.

The mean error is 11.19 cm, the median error is 11.37 cm, and the standard deviation is 4.88 cm. Thus, we can see that the our system can give an accurate distance estimation, and we can lock the sound sources within its two adjacent distance bins. These results demonstrate that UWB audio sensing is *distance-aware* in terms of estimating how far the sound sources are from the sensors.

6.2 Qualitative Sound Separation Test

Sound separation is an active research field. Once target sounds and noises are mixed in the microphone recordings, it is difficult to separate them apart as they are entangled both in time and frequency domain. Deep learning-based blind separation algorithms have been used to solve this problem [27, 66]. Our system proposes a new potential solution to this problem: separating the sound in the IR-UWB *fast time* domain. Our system is able to deal with multiple simultaneous sounds occurring at different distances, and separate them apart based on ToF ranging in *fast time* domain.

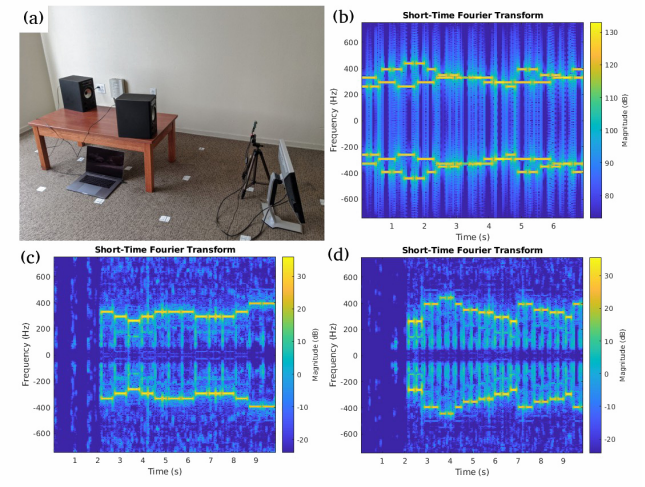

Figure 16: Sound separation using UWHear: (a) Experimental settings (b) Spectrogram of the microphone audio (c-d) Spectrogram of the recovered sound from the two speakers

Figure 16 demonstrates our qualitative experiments on sound separation. As shown in part (a), the two speakers are placed at different distances, one at 58 cm playing *Mary has a little lamb* and

another at 122 cm playing *Twinkle twinkle little star*. The spectrogram of the sound recorded by a microphone is shown in part(b), where the two songs are entangled and cannot be easily separated. We use the open-source Free Universal Sound Separation (FUSS) baseline separation model [66] trying to separate the microphone audio, but the model failed to give satisfying results. This is probably due to the discrepancy between our real-world testing case and its training data. Part (c) and (d) show the output of our system. By selecting different distance bins, we can separate the two songs without any residual. These results demonstrate that UWHear is capable of directly capturing multi-track audio for applications such as acoustic scene classification and sound event detection.

6.3 Quantitative Sound Separation Test

We also quantitatively measure the sound separation ability of UWHear. Figure 17(a) shows the experimental setting. We put two speakers in the UWB sensor's field of view in roughly the same direction. Speaker 1 is playing a C4 note (261.63 Hz), while Speaker 2 plays a A3 note (220.00 Hz). Speaker 1 is placed 1 meter from the sensor. Speaker 2 starts from the distance of 1.5m, and is moved towards the sensor in increments of 5cm up until a distance of 0.5m[1]. We use d to denote the distance between the two speakers, $d \in [-50, 50]$ cm, as shown in the figure. For the sound recovered from Speaker 1, we define the *sound purity* as the ratio of the C4 note energy to the A3 note energy in decibels, which is, in essence, the SNR ignoring background noise. The purity for Speaker 2 sound is defined similarly.

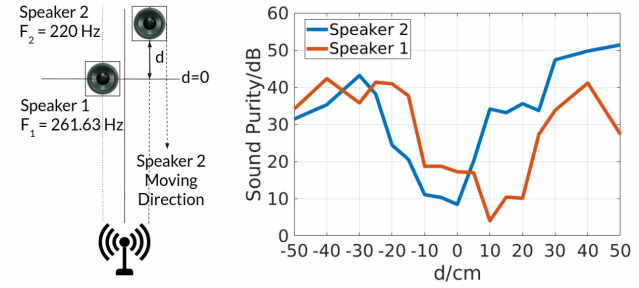

Figure 17: Quantitative analysis of sound separation using UWHear: (a) Experiment setup, and (b) Recovered sound purity against the distance between the two speakers.

Figure 17(b) demonstrates the results of our qualitative experiments on sound separation. Perceptually, if the sound purity is over 20dB, i.e., the target sound has 100x more energy than its counterpart, the effect of a non-target sound is not audible and can be ignored. The experiments show that the target sound may be spread across a few adjacent distance bins, probably due to the fact that the speaker case is also vibrating, as well as the fact that multipath reflections may also "leak" some information. Figure 17(b) demonstrates the feasibility of collecting two sound sources separately without cross-interference if the sound sources are placed **25 cm** apart. The two purity curves are slightly asymmetric due to manual placement errors in the speaker distances and angles.

[1]Through theoretical calculations, we know that the spatial resolution of IR-UWB radar is roughly 10 cm, which means, in theory, two sound sources can be separated apart if they are 10cm apart in distance.

6.4 UWHear in Household Settings

Thus far, we have only tested UWHear on speakers as the contents played in a speaker are easier to control and quantify. For a real-world audio sensing system, the system need to deal with the sound from heterogeneous sources. In this experiment, we test UWHear on some commonly seen sound sources in domestic environments.

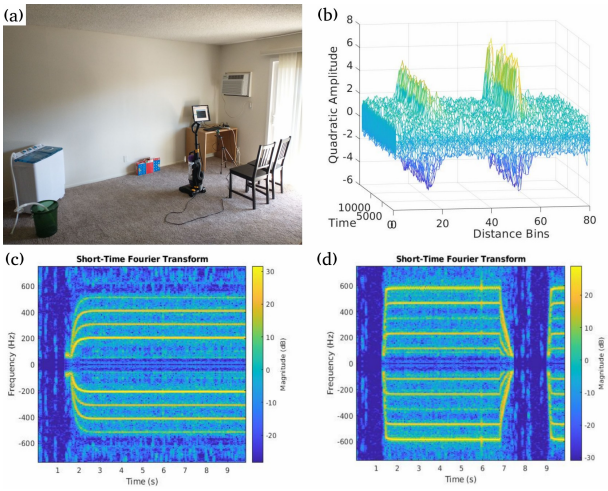

Figure 18: Sound source separation in Household Settings:(a) Experimental settings (b) Visualization of the quadrature part of the IR-UWB data after processing (c) Spectrogram of the recovered vacuum cleaner sound (d) Spectrogram of the recovered washing machine sound.

Figure 18(a) shows the experimental settings: A washing machine and a vacuum cleaner are placed a few meters in front of the sensor at a random angle, and are operating simultaneously. In the background, a wall AC unit is also operating to produce non-negligible noise. Figure 18(b) visualize the quadrature part of the IR-UWB data after processing. We notice two ridges at distance bin 18 and 53, standing for the washing machine and the vacuum cleaner separately. The spectrograms in Figure 18(c)(d) expose the different frequency characteristics of those sound sources and elucidate the starting and stopping phase of the motors. There are neither cross-interference between the two target sound, nor traces of the noise from the wall AC unit. Through this experiment, we hope to shed some light on UWHear's potential to collect multi-track audio in domestic environments, which can then benefit the sound event detection and classification systems.

7 DISCUSSION AND FUTURE WORK

Improving sampling rate. A major limitation of current UWHear prototype is the relatively low sampling rate. While IR-UWB radars have a fast sampling system on the fast time (collecting responses), the vibrations can only be recovered by analysing a series of frames. Thus, what matters is the granularity of the *slow time*, i.e., the frame rate. Currently, the frame rate cannot go beyond 1.6 kHz owing to the fact that the X4 IR-UWB radar chip only caches the last frame it receives, and that the data transmission speed is limited by the SPI interface. In order to perform recovery of human voice, we

need a sampling rate of at least 3kHz (the sampling rate of a landline telephone) to ensure a sufficient understanding of human speech, as the voiceless consonants that are critical in human speech usually have only high-frequency components. Currently, it is possible to understand the digits read by a human speaker with UWHear recovered audio. However, it may be difficult to understand a much larger corpus due to the sampling rate limitations. Future efforts should leverage superior hardware such as Quad Serial Peripheral Interface (QSPI), or substitute the Raspberry Pi with an FPGA to increase the data transmission rate.

Improving SNR. Even though Section 5 shows that UWHear is capable of operating in complicated environments, it still suffers from a drop in SNR under unfavorable conditions. In the worst cases, the recovered sound may degrade below the quality threshold for machine processing or the human auditory system. The performance of our system is limited by the COTS radar board and the antennas. We expect that an innovative design of the hardware components that increases the transmission power may help. In X4 driver settings, if the IR-UWB radar power setting is increased from "mid" to "high", the effective range of our system increases. Further adding a controllable low-noise amplifier and a power amplifier between the X4 SoC and the antenna might help to increase the system performance.

Increasing field of view. UWHear is currently using a directional antenna whose field of view (FOV) is about 50 degrees in azimuth and elevation, which implies that the sensor can only work if the sound source lies in the sensor's FOV. In the future, we hope that this directional problem can be solved by integrating multiple IR-UWB radar instances into a single board and stitch their data to cover all directions. Another potential solution is to use multiple instances of IR-UWB radars with omnidirectional antennas to perform trilateration [55].

Potential Applications of UWHear. UWHear provides a robust audio sensing interface for Sound Event Detection and Classification (SEDC), which can benefit a number of downstream applications. In domestic environments, collecting multi-track source-separated audio can help us detect appliance use and human physical activities in home environments with the accurate onset and offset time estimation [16, 56], helping save energy or making the environment more responsive. Additionally, in industrial settings, SEDC with UWHear provides the ability to monitor several machines and devices' functionality. UWHear, in this case, will have the ability to detect abnormal vibrations or early failures robustly. At an urban scale, we expect that UWHear can help urban sound tagging with spatiotemporal context. Currently, LiDARs are widely used to build 3D models around us [2, 41, 63]. For example, a fusion of LiDAR and UWHear can tag the object with vibration signatures that may be important, for example, in classifying vehicles and human activities. Finally, the spatial audio generated from UWHear may cooperate with RF activity sensing systems [19, 52, 64] and be used to understand complex events and human behaviors [48, 67]. We also envision that IR-UWB radar technologies can be incorporated on mobile platforms like smartphones to make rich inferences using audio-related vibrations like [12] did. We are glad to notice that Ultra-Wideband technology is now available on iPhone 11.

Other Limitations and Generalization. Different materials of the target interact with the IR-UWB radar wave differently. It could reflect, absorb, or be penetrated by the signal, or show a combination of the three in most cases. For example, our system may work fine on metal and polyester speaker diaphragms, but its sensitivity might drop for paper cones. Moreover, our tests reveal that the UWB-based system cannot recover voice directly from a human throat. Generally speaking, UWHear works better on objects with good reflections and noticeable vibrations. For example, it may generalize well on sensing machine vibrations but may encounter difficulties recovering plastic bags' sound.

Above are the intrinsic disadvantages of using a single wireless signal for vibrometry. To overcome them, one of the future directions can be constructing a comprehensive wireless vibrometry system that combines modalities such as mmWave, IR-UWB, and lasers. Operating at different frequency ranges, these technologies can compensate for each other and make a more robust audio sensing system. It might also be possible to merge the IR-UWB audio sensor with traditional microphones to provide additional spatial information. Currently, UWHear focuses mainly on indoor near-field environments. However, we believe the idea of using short-duration pulses and ToF to collect vibrations with spatial information is universal and can be expanded to other sensing technologies.

8 CONCLUSION

In this paper, we propose UWHear, an audio sensing system using impulse radio Ultra-wideband (IR-UWB) radar. We mathematically formulated the theory of recovering audio using non-continuous impulse-based wireless vibrometry. We also implement UWHear using a commercial-off-the-shelf (COTS) IR-UWB radar and a learning-free signal processing pipeline. Our results show that this system is able to retrieve the sound directly from multiple sound sources and also estimate the distances from each source to the sensor. Such characteristics allow us to acquire and separate multiple sounds of interest simultaneously in the presence of background noise. We also show that UWHear is capable of through-wall sensing of audio vibrations. We believe that this a promising step towards robust audio sensing in complicated environments and it will benefit a broad set of applications that involve computational analysis of sound events and scenes.

ACKNOWLEDGMENTS

The authors would like to thank the shepherd and the reviewers for their comments that helped improve this work, and also thank Tianyue Zheng of Nanyang Technological University, Singapore for helpful discussions. This research was sponsored in part by the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA, by the Army Research Laboratory (ARL) under Cooperative Agreement W911NF-17-2-0196, and by the National Science Foundation (NSF) under awards CNS-1329755 and CNS-1705135. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the funding agencies. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a better understanding of context and context-awareness. In *International symposium on handheld and ubiquitous computing*. Springer, 304–307.

[2] Fawad Ahmad, Hang Qiu, Ray Eells, Fan Bai, and Ramesh Govindan. 2020. CarMap: Fast 3D Feature Map Updates for Automobiles. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2020)*. 1063–1081.

[3] Amr Alanwar, Henrique Ferraz, Kevin Hsieh, Rohit Thazhath, Paul Martin, João Hespanha, and Mani Srivastava. 2017. D-SLATS: Distributed simultaneous localization and time synchronization. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 1–10.

[4] Amazon. 2020. *Amazon Alexa Premium Far-Field Voice Dev Kit.* https://developer.amazon.com/en-US/alexa/alexa-voice-service/dev-kits/amazon-premium-voice Accessed: 2020-10-18.

[5] Nikolaj Andersen, Kristian Granhaug, Jørgen Andreas Michaelsen, Sumit Bagga, Håkon A Hjortland, Mats Risopatron Knutsen, Tor Sverre Lande, and Dag T Wisland. 2017. A 118-mw pulse-based radar soc in 55-nm cmos for non-contact human vital signs detection. *IEEE Journal of Solid-State Circuits* 52, 12 (2017), 3421–3433.

[6] Novelda AS. 2020. Novelda Presence Sensor. https://novelda.com/novelda-presence-sensor.html. Accessed: 2020-07-01.

[7] Novelda AS. 2020. X4 Datasheet - Impulse Radar Transceiver SoC. https://novelda.com/fo/themes/default/img/contents/x4_datasheet_revF.pdf. Accessed: 2020-07-01.

[8] Novelda AS. 2020. Xethru Raspberry Driver Example. https://github.com/novelda/Legacy-SW/tree/master/Examples/X4Driver_RaspberryPi. Accessed: 2020-05-28.

[9] Novelda AS. 2020. XeThru X4 Phase Noise Correction. https://github.com/novelda/Legacy-Documentation/blob/master/Application-Notes/XTAN-14_XeThru_X4_Phase_Noise_Correction_rev_a.pdf. Accessed: 2020-05-28.

[10] Michael Berouti, Richard Schwartz, and John Makhoul. 1979. Enhancement of speech corrupted by acoustic noise. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4. IEEE, 208–211.

[11] S Boll. 1979. A spectral subtraction algorithm for suppression of acoustic noise in speech. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4. IEEE, 200–203.

[12] Chao Cai, Zhe Chen, Henglin Pu, Liyuan Ye, Menglan Hu, and Jun Luo. 2020. AcuTe: Acoustic Thermometer Empowered by a Single SmartPhone. In *Proc. of the 18th ACM SenSys*. 1–14. https://doi.org/10.1145/3384419.3430714.

[13] Jianfeng Chen, Alvin Harvey Kam, Jianmin Zhang, Ning Liu, and Louis Shue. 2005. Bathroom activity monitoring based on sound. In *International Conference on Pervasive Computing*. Springer, 47–61.

[14] Mango Communications. 2015. WARP v3 User Guide: RF Interfaces. https://warpproject.org/trac/wiki/HardwareUsersGuides/WARPv3/RF Accessed: 2020-10-18.

[15] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J Mysore, Frédo Durand, and William T Freeman. 2014. The visual microphone: passive recovery of sound from video. (2014).

[16] Gert Dekkers, Steven Lauwereins, Bart Thoen, Mulu Weldegebreal Adhana, Henk Brouckxon, Toon van Waterschoot, Bart Vanrumste, Marian Verhelst, and Peter Karsmakers. 2017. The SINS Database for Detection of Daily Activities in a Home Environment Using an Acoustic Sensor Network. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*. 32–36.

[17] Ashutosh Dhekne, Mahanth Gowda, Yixuan Zhao, Haitham Hassanieh, and Romit Roy Choudhury. 2018. Liquid: A wireless liquid identifier. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 442–454.

[18] Maria-Gabriella Di Benedetto. 2006. *UWB communication systems: a comprehensive overview*. Vol. 5. Hindawi Publishing Corporation.

[19] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. 2020. RF-Net: A Unified Meta-Learning Framework for RF-enabled One-Shot Human Activity Recognition. In *Proc. of the 18th ACM SenSys*. 1–14. https://doi.org/10.1145/3384419.3430735.

[20] Chris Donahue, Bo Li, and Rohit Prabhavalkar. 2018. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5024–5028.

[21] Igor Dotlic, Andrew Connell, Hang Ma, Jeff Clancy, and Michael McLaughlin. 2017. Angle of arrival estimation using decawave DW1000 integrated circuits. In *2017 14th Workshop on Positioning, Navigation and Communications (WPNC)*. IEEE, 1–6.

[22] Francois G Germain, Qifeng Chen, and Vladlen Koltun. 2018. Speech denoising with deep feature losses. *arXiv preprint arXiv:1806.10522* (2018).

[23] Anthony Griffin, Anastasios Alexandridis, Despoina Pavlidi, Yiannis Mastorakis, and Athanasios Mouchtaris. 2015. Localizing multiple audio sources in a wireless acoustic sensor network. *Signal Processing* 107 (2015), 54–67.

[24] Danilo Hollosi, Jens Schröder, Stefan Goetze, and Jens-E Appell. 2010. Voice activity detection driven acoustic event classification for monitoring in smart homes. In *2010 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL 2010)*. IEEE, 1–5.

[25] Antonio Ramón Jiménez and Fernando Seco. 2016. Comparing Decawave and Bespoon UWB location systems: Indoor/outdoor performance analysis. In *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 1–8.

[26] Sunil Kamath and Philipos Loizou. 2002. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise.. In *ICASSP*, Vol. 4. Citeseer, 44164–44164.

[27] Ilya Kavalerov, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin Wilson, Jonathan Le Roux, and John R Hershey. 2019. Universal sound separation. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 175–179.

[28] M Klemm, IJ Craddock, JA Leendertz, A Preece, DR Gibbins, M Shere, and R Benjamin. 2010. Clinical trials of a UWB imaging radar for breast cancer. In *Proceedings of the Fourth European Conference on Antennas and Propagation*. IEEE, 1–4.

[29] Rakhesh Singh Kshetrimayum. 2009. An introduction to UWB communication systems. *IEEE Potentials* 28, 2 (2009), 9–13.

[30] MATRIX labs. 2020. The IoT Development Board for Building Incredibly Smart Products. https://www.matrix.one/products/creator Accessed: 2020-10-18.

[31] Dawna Lewis, Kendra Schmid, Samantha O'Leary, Jody Spalding, Elizabeth Heinrichs-Graham, and Robin High. 2016. Effects of noise on speech recognition and listening effort in children with normal hearing and children with mild bilateral or unilateral hearing loss. *Journal of Speech, Language, and Hearing Research* 59, 5 (2016), 1218–1232.

[32] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach. 2014. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 4 (2014), 745–777.

[33] Xiaolin Liang, Jianqin Deng, Hao Zhang, and Thomas Aaron Gulliver. 2018. Ultra-wideband impulse radar through-wall detection of vital signs. *Scientific reports* 8, 1 (2018), 1–21.

[34] Liang Liu, Junyan Ren, Xuejing Wang, and Fan Ye. 2007. Design of low-power, 1GS/s throughput FFT processor for MIMO-OFDM UWB communication system. In *2007 IEEE International Symposium on Circuits and Systems*. IEEE, 2594–2597.

[35] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. TUT database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 1128–1132.

[36] Eliya Nachmani, Yossi Adi, and Lior Wolf. 2020. Voice Separation with an Unknown Number of Multiple Speakers. *arXiv preprint arXiv:2003.01531* (2020).

[37] Rajalakshmi Nandakumar, Shyamnath Gollakota, and Nathaniel Watson. 2015. Contactless sleep apnea detection on smartphones. In *Proceedings of the 13th annual international conference on mobile systems, applications, and services*. 45–57.

[38] Tuan Anh Nguyen and Marco Aiello. 2013. Energy intelligent buildings based on user activity: A survey. *Energy and buildings* 56 (2013), 244–257.

[39] Tsuyoki Nishikawa, Hiroshi Saruwatari, and Kiyohiro Shikano. 2003. Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 86, 4 (2003), 846–858.

[40] Despoina Pavlidi, Anthony Griffin, Matthieu Puigt, and Athanasios Mouchtaris. 2013. Real-time multiple sound source localization and counting using a circular microphone array. *IEEE Transactions on Audio, Speech, and Language Processing* 21, 10 (2013), 2193–2206.

[41] Hang Qiu, Fawad Ahmad, Fan Bai, Marco Gruteser, and Ramesh Govindan. 2018. Avr: Augmented vehicular reality. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 81–95.

[42] Niranjini Rajagopal, John Miller, Krishna Kumar Reghu Kumar, Anh Luong, and Anthony Rowe. 2018. Demo abstract: welcome to my world: demystifying multi-user AR with the cloud. In *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 146–147.

[43] Christian G Reiff. 2009. Acoustic source localization and cueing from an aerostat during the NATO SET-093 field experiment. In *Unattended Ground, Sea, and Air Sensor Technologies and Applications XI*, Vol. 7333. International Society for Optics and Photonics, 73330M.

[44] Dario Rethage, Jordi Pons, and Xavier Serra. 2018. A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5069–5073.

[45] Hugh Robjohns. 2001. A brief history of microphones. In *Microphone Data Book*. http://microphone-data.com/media/filestore/articles/History-10.pdf.

[46] Antonio Ramón Jiménez Ruiz and Fernando Seco Granja. 2017. Comparing ubisense, bespoon, and decawave uwb location systems: Indoor performance analysis. *IEEE Transactions on instrumentation and Measurement* 66, 8 (2017), 2106–2117.

[47] F. Sabath, E. L. Mokole, and S. N. Samaddar. 2005. Definition and classification of ultra-wideband signals and devices. *URSI Radio Science Bulletin* 2005, 313 (2005),

12–26.
[48] Swapnil Sayan Saha, Sandeep Singh Sandha, and Mani Srivastava. 2020. *Deep Convolutional Bidirectional LSTM for Complex Activity Recognition with Missing Data.* Springer.
[49] Sana Salous, Vittorio Degli Esposti, Franco Fuschini, Reiner S Thomae, Robert Mueller, Diego Dupleich, Katsuyuki Haneda, Jose-Maria Molina Garcia-Pardo, Juan Pascual Garcia, Davy P Gaillot, et al. 2016. Millimeter-Wave Propagation: Characterization and modeling toward fifth-generation systems.[Wireless Corner]. *IEEE Antennas and Propagation Magazine* 58, 6 (2016), 115–127.
[50] Hiroshi Sawada, Shoko Araki, Ryo Mukai, and Shoji Makino. 2006. Blind extraction of dominant target sources using ICA and time-frequency masking. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 6 (2006), 2165–2173.
[51] Elahe Shojaei, Hassan Ashayeri, Zahra Jafari, Mohammad Reza Zarrin Dast, and Koorosh Kamali. 2016. Effect of signal to noise ratio on the speech perception ability of older adults. *Medical journal of the Islamic Republic of Iran* 30 (2016), 342.
[52] Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani Srivastava. 2019. Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar. In *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*. 51–56.
[53] Nikolaos Stefanakis, Despoina Pavlidi, and Athanasios Mouchtaris. 2017. Perpendicular cross-spectra fusion for sound source localization with a planar microphone array. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 9 (2017), 1821–1835.
[54] Yuki Tamai, Yoko Sasaki, Satoshi Kagami, and Hiroshi Mizoguchi. 2005. Three ring microphone array for 3d sound localization and separation for mobile robot audition. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 4172–4177.
[55] Federico Thomas and Lluís Ros. 2005. Revisiting trilateration for robot localization. *IEEE Transactions on robotics* 21, 1 (2005), 93–101.
[56] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. 2019. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *Workshop on Detection and Classification of Acoustic Scenes and Events*. New York City, United States. https://hal.inria.fr/hal-02160855
[57] Efthymios Tzinis, Scott Wisdom, John R Hershey, Aren Jansen, and Daniel PW Ellis. 2020. Improving universal sound separation using sound classification. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 96–100.
[58] Michel Vacher, Dan Istrate, Laurent Besacier, Jean-François Serignat, and Eric Castelli. 2004. Sound detection and classification for medical telesurvey.
[59] J-M Valin, François Michaud, Jean Rouat, and Dominic Létourneau. 2003. Robust sound source localization using a microphone array on a mobile robot. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, Vol. 2. IEEE, 1228–1233.
[60] Swaroop Venkatesh, Christopher R Anderson, Natalia V Rivera, and R Michael Buehrer. 2005. Implementation and analysis of respiration-rate estimation using impulse-based UWB. In *MILCOM 2005-2005 IEEE Military Communications Conference*. IEEE, 3314–3320.
[61] Kavitha Viswanathan and Sharmila Sengupta. 2015. Blind navigation proposal using SONAR. In *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*. IEEE, 151–156.
[62] Hanbiao Wang, Jeremy Elson, Lewis Girod, Deborah Estrin, and Kung Yao. 2003. Target classification and localization in habitat monitoring. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, Vol. 4. IEEE, IV–844.
[63] Qiaosong Wang. 2019. Towards Real-time 3D Reconstruction using Consumer UAVs. *arXiv preprint arXiv:1902.09733* (2019).
[64] Ziqi Wang, Zhihao Gu, Junwei Yin, Zhe Chen, and Yuedong Xu. 2018. Syncope detection in toilet environments using Wi-Fi channel state information. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 287–290.
[65] Teng Wei, Shu Wang, Anfu Zhou, and Xinyu Zhang. 2015. Acoustic eavesdropping through wireless vibrometry. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. ACM, 130–141.
[66] Scott Wisdom, Hakan Erdogan, Daniel P. W. Ellis, Romain Serizel, Nicolas Turpault, Eduardo Fonseca, Justin Salamon, Prem Seetharaman, and John R. Hershe. 2020. What's All the FUSS About Free Universal Sound Separation Data? https://github.com/google-research/sound-separation/tree/master/models/dcase2020_fuss_baseline Accessed: 2020-10-18.
[67] Tianwei Xing, Marc Roig Vilamala, Luis Garcia, Federico Cerutti, Lance Kaplan, Alun Preece, and Mani Srivastava. 2019. Deepcep: Deep complex event processing using distributed multimodal information. In *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 87–92.
[68] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. 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*. ACM, 14–26.
[69] Lei Yang, Yao Li, Qiongzheng Lin, Xiang-Yang Li, and Yunhao Liu. 2016. Making sense of mechanical vibration period with sub-millisecond accuracy using backscatter signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 16–28.
[70] Jiaxing Ye, Takumi Kobayashi, and Masahiro Murakawa. 2017. Urban sound event classification based on local and global features aggregation. *Applied Acoustics* 117 (2017), 246–256.
[71] Esfandiar Zavarehei. 2020. Berouti Spectral Subtraction MATLAB implementation. https://www.mathworks.com/matlabcentral/fileexchange/7653-berouti-spectral-subtraction. Accessed: 2020-10-18.
[72] Esfandiar Zavarehei. 2020. Boll Spectral Subtraction MATLAB implementation. https://jp.mathworks.com/matlabcentral/fileexchange/7675-boll-spectral-subtraction. Accessed: 2020-10-18.
[73] Esfandiar Zavarehei. 2020. Multi-band Spectral Subtraction MATLAB implementation. https://www.mathworks.com/matlabcentral/fileexchange/7674-multi-band-spectral-subtraction. Accessed: 2020-10-18.
[74] Cemin Zhang, Michael Kuhn, Brandon Merkl, Aly E Fathy, and Mohamed Mahfouz. 2006. Accurate UWB indoor localization system utilizing time difference of arrival approach. In *2006 IEEE radio and wireless symposium*. IEEE, 515–518.
[75] Dongheng Zhang, Yang Hu, Yan Chen, and Bing Zeng. 2019. BreathTrack: Tracking indoor human breath status via commodity WiFi. *IEEE Internet of Things Journal* 6, 2 (2019), 3899–3911.
[76] Yang Zhang, Gierad Laput, and Chris Harrison. 2018. Vibrosight: Long-Range Vibrometry for Smart Environment Sensing. In *The 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, 225–236.
[77] Tianyue Zheng, Zhe Chen, Chao Cai, Jun Luo, and Xu Zhang. 2020. V2iFi: in-Vehicle Vital Sign Monitoring via Compact RF Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol* 4, 2 (Jun 2020). https://doi.org/10.1145/33973211