Experience: Practical Problems for Acoustic Sensing

Dong Li*, Shirui Cao*, Sunghoon Ivan Lee, Jie Xiong
University of Massachusetts Amherst
{dli, shiruicao, silee, jxiong}@cs.umass.edu

ABSTRACT

Acoustic sensing shows great potential to transform billions of consumer-grade electronic devices that people interact with on a daily basis into ubiquitous sensing platforms. In this paper, we share our experience and findings during the process of developing and deploying acoustic sensing systems for real-world usage. We identify multiple practical problems that were not paid attention to in the research community, and propose the corresponding solutions. The challenges include: (i) there exists annoying audible sound leakage caused by acoustic sensing; (ii) acoustic sensing actually affects music play and voice call; (iii) acoustic sensing consumes a significant amount of power, degrading the battery life; (iv) real-world device mobility can fail acoustic sensing. We hope the shared experience can benefit not only the future development of sensing algorithms but also the hardware design, pushing acoustic sensing one step further towards real-life adoption.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in ubiquitous and mobile computing**.

KEYWORDS

acoustic sensing, practical problems, audible leakage, real-world adoption, coexistence of sensing and music play

ACM Reference Format:
Dong Li*, Shirui Cao*, Sunghoon Ivan Lee, Jie Xiong. 2022. Experience: Practical Problems for Acoustic Sensing. In *The 28th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '22), October 17–21, 2022, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3495243.3560527

1 INTRODUCTION

Acoustic sensing has been extensively studied over the past few years. The research community has successfully exploited acoustic signals emitted from commodity devices to sense the contexts of human targets (e.g., hand gestures [39]) and environment (e.g., temperature changes [4]). Compared to other sensing modalities such as WiFi sensing, acoustic sensing can sense much finer-grained activities such as eye blink [23] and heartbeat [43], owing to the low propagation speed in the air (340 m/s). Furthermore, while WiFi sensing requires dedicated wireless network cards such as Intel 5300 that are not generally available in commodity WiFi access points or smartphones, the wide availability of speakers and microphones in consumer-grade electronic devices makes acoustic sensing one of the most promising candidates for real-life adoption.

However, during the process of pushing acoustic sensing from the laboratory to real world, we find several practical problems that were not paid attention to in the research community. Prior studies on acoustic sensing mainly devoted effort to improving the sensing accuracy/granularity [38, 39, 42], increasing the sensing range [20, 21, 24, 27, 28], and enabling new applications [10, 23, 36, 45, 46]. However, multiple critical practical problems still exist that can greatly hinder the real-world adoption of acoustic sensing if not properly addressed. This paper shares our experience and findings when we move one step further to push acoustic sensing for real-world adoption. We present the identified practical problems below.

The first practical problem is that *annoying audible sound leakage can ruin the user experience of acoustic sensing*. Specifically, to make the sensing process unobtrusive for human beings, acoustic sensing systems usually adopt inaudible acoustic signals whose frequency is above the human hearing range as the sensing signals [1]. However, according to our experiments, audible sound leakage generally exists, indicating that people can still hear audible sounds during the process of acoustic sensing. For some devices, the intensity level of the audible sound leakage can be higher than 65 dB, which is close to the noise generated by a vacuum cleaner [18].

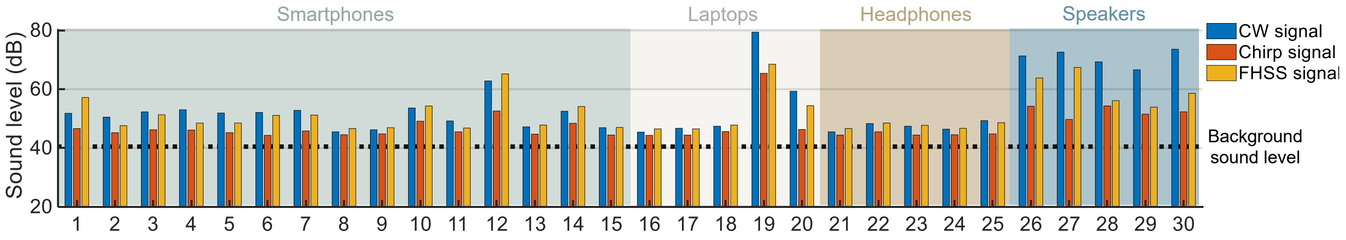
The second practical problem is that *acoustic sensing can significantly affect music play and voice call.* Acoustic sensing systems [5] usually assume that no other applications are using the speaker during the process of acoustic sensing. In real life, a user expects sensing and other applications such as music play to co-exist, which is critical for the wide adoption of acoustic sensing. However, based on our experiments, we find that acoustic sensing negatively impacts music play on a large variety of devices that run Android and Windows operating systems. Specifically, the quality of music play gets degraded if sensing and music play happen simultaneously.

The third practical problem is that *the large power consumption of acoustic sensing can greatly reduce the battery life of battery-powered devices.* During the sensing process, acoustic sensing systems unremittingly transmit and receive sensing signals at a high rate [5]. This process consumes a large amount of power especially for long-term sensing applications, which is non-negligible for battery-powered devices. Through our experiments, we observe that, when acoustic sensing is enabled for 2 hours, there is a 22% battery drop on a smartphone, a 78% battery drop on a smartwatch, and a 66% battery drop on a wireless earphone, respectively.

The fourth practical problem is that *device mobility can severely degrade the performance of acoustic sensing*. Existing sensing systems assume that the sensing device is stationary during the sensing

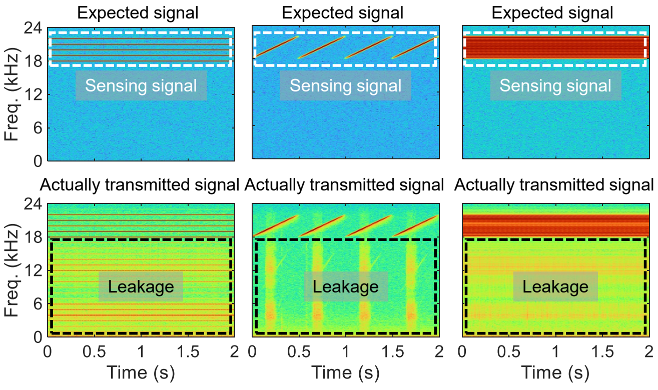
*Both authors contributed equally to the paper.

ACM MobiCom '22, October 17–21, 2022, Sydney, NSW, Australia
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9181-8/22/10…$15.00
https://doi.org/10.1145/3495243.3560527



1. Xiaomi 10S
2. Samsung Galaxy Note10+
3. Huawei P30 pro
4. Samsung Galaxy S22 Ultra
5. Samsung S9+
6. Sony Xperia
7. Samsung Galaxy Note20 Ultra
8. Huawei Mate 10 pro
9. Apple iPhone 13 Mini
10. Google Pixel 4
11. Apple iPhone 12
12. Apple iPhone 11 pro
13. Apple iPhone 7
14. Apple iPhone 6
15. Apple iPhone 5c
16. Dell G15 Series
17. Huawei Matebook X Pro
18. Thinkpad X1 Extreme
19. Macbook Pro Late 2013 13”
20. Macbook Pro 2019 16”
21. Huawei AM115
22. HyperX Cloud Alpha Pro
23. Denon AH-C700
24. Sony MH410C
25. Beats solo 2
26. ARVICKA Speaker
27. Logitech S120 Speaker
28. AL-202 Speaker
29. Sunln Speaker
30. Moloroll Speaker

Figure 1: The sound intensity level of the audible leakage for 30 commodity devices.



(a) CW signal. **(b) Chirp signal.** **(c) FHSS signal.**

Figure 2: The illustration of audible sound leakage for three types of sensing signals on a Google Pixel 4.

process [5], e.g., the device is placed on a table. This assumption is not always true, and we encounter quite a few real-world scenarios where the device is moving. For example, when a user wears the device for social distance measurement, the device moves a lot along with the human body. To extract useful information for social distance measurement, we need to first differentiate between human target and static objects such as pillars. This is an easy task when the device is stationary as we can use the Doppler information, i.e., Doppler value is zero to identify static objects such as pillars. However, when the device is moving, differentiation between human target and static objects becomes challenging because even signals reflected from static objects exhibit non-zero Doppler values. We believe this is a critical issue that needs to be tackled before acoustic sensing can be adopted for real-world usage.

2 AUDIBLE SOUND LEAKAGE

This section presents the first practical issue, i.e., audible sound leakage, that negatively influences the user experience of acoustic sensing. To illustrate the issue, we transmit three types of commonly-used ultrasound sensing signals using a Google Pixel 4 smartphone and use Sony PCM-D100 audio recorder [7] to record the actually transmitted signal. As shown in Figure 2, we can observe that, besides the inaudible sensing signals at high frequencies, Google Pixel 4 leaks audible sound noises at low frequencies for all the three signal types.[1] The audible leakage is extremely annoying which can cause poor user experience or even discomfort during the process of long-term acoustic sensing. In the following, we first conduct measurements to show that the leakage occurs on a variety of commodity devices. Then we dig deep to identify the root cause of the leakage. At last, we share our experience and present the solution to alleviate the audible leakage.

2.1 Measurements on Commodity Devices

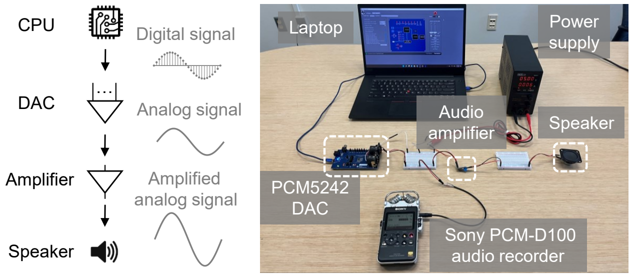
We conduct measurements to test whether audible sound leakage generally exists on commodity devices. To this end, we transmit three types of sensing signals in the inaudible frequency band (i.e., $18 - 22\ kHz$) from a variety of commodity devices and measure the sound level of the audible leakage.

The three chosen sensing signals are commonly adopted for sensing [5], including continuous wave (CW) signal, chirp signal, and Frequency Hopping Spectrum Spread (FHSS) signal[2], as shown in Figure 2. The commodity devices involve 15 smartphones, five laptops, five headphones, and five speakers. The sound intensity level is measured by the VLIKE sound level meter [37]. The distance between the sensing device and the sound level meter is varied based on the real-life usage of the devices, i.e., 10 cm for smartphones, 25 cm for laptops, 0 cm for headphones, and 60 cm for speakers. The volume is set to 80% of the maximum volume of each device. It is worth noting that, due to fast attenuation over distance of acoustic signals [28], most sensing systems adopt a high volume or even the maximum volume in order to achieve a larger sensing range and better sensing performance [31].

Figure 1 illustrates the measured sound level of the audible leakage when inaudible sensing signals are transmitted from the speaker for sensing. We can observe that the audible sound leakage generally exists on most commodity devices for all the three types of sensing signals. For some devices such as iPhone 11 Pro and MacBook Pro Late 2013, the sound levels of the leakage during the

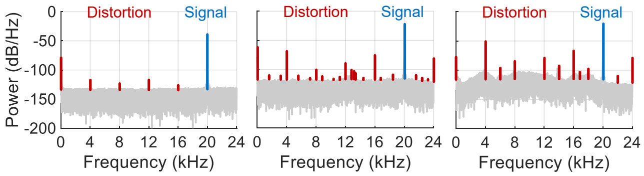
[1] The demo audio can be found at https://youtu.be/IcaulzTU_Ks.

[2] The GSM training sequence is usually adopted to control frequency hopping.



(a) System architecture. **(b) Test platform.**

Figure 3: The illustration of the system architecture and the test platform for acoustics signal generation analysis.



(a) DAC. **(b) Amplifier.** **(c) Speaker.**

Figure 4: The non-linear distortion after a 20 *kHz* continuous wave passes through each electronic component.

sensing process are higher than 65 *dB*, which is close to the noise generated by a vacuum cleaner [18]. Note that, even though the sound level of leakage is just 50 *dB*, due to its harsh and jarring nature, the leakage still causes severe sound pollution and makes users uncomfortable during long-term acoustic sensing.

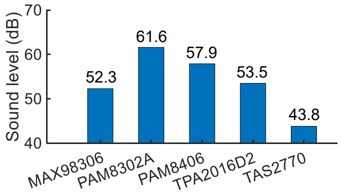
2.2 Acoustic Signal Generation Analysis System

To study why the audible sound leakage occurs, we build a test system to analyze the signal output from each electronic component at the speaker side. Figure 3a and Figure 3b illustrate the system architecture and the test platform, respectively. Specifically, the digital sensing signals are first generated by the CPU and then fed into the audio Digital-to-Analog Converter (DAC) to output the analog sensing signals. The analog sensing signals are further amplified by the audio amplifier and finally converted to acoustic waves by the speaker.

Next, we "hack" the signal transmission process by capturing the signal after it passes through each electronic component using the Sony PCM-D100 audio recorder [7] which has extremely low distortions. Specifically, we connect the output of the PCM5242 DAC [16] and PAM8302 amplifier [17] with the line-in jack of the audio recorder via a 3.5 *mm* audio cable. Furthermore, the output of the speaker [22] is recorded by the built-in electret condenser microphones on the audio recorder. The recorded signals are transferred to a laptop for visualization and analysis.

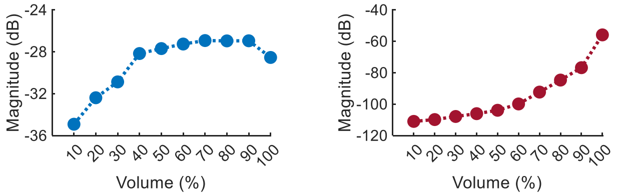
2.3 Findings and Solutions

This section shares our findings and solutions on how to alleviate audible sound leakage.



Amplifier name	Price per unit
MAX 98306	$1.182
PAM8302A	$0.265
PAM8406	$0.374
TPA2016D2	$0.846
TAS2770	$1.262

Figure 5: The comparison among different amplifiers.



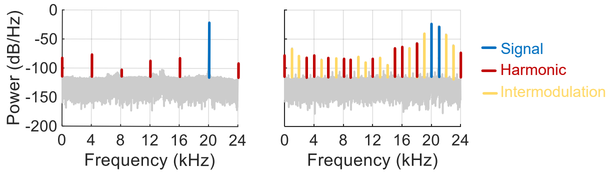
(a) Sensing signal. **(b) Distortion.**

Figure 6: The magnitude changes of the sensing signal and distortion as the amplifier volume increases.

Finding 1. *The non-linear distortion of the audio amplifier is the chief culprit of the audible sound leakage.* Figure 4 illustrates the Power Spectral Density of the signals output by each component when playing a 20 *kHz* continuous wave. We observe that, there exists non-linearity distortion for all three electronic components, where the audio amplifier introduces most of the distortion. Specifically, we can hardly hear the leakage from the audio that was recorded directly after the DAC. However, since the power of the distortion is increased by 43.59 *dB* (i.e., from −112.63 *dB* to −69.04 *dB*), the leakage after the amplifier becomes audible. In contrast, the power of the distortion is only increased by 15.7 *dB* (i.e., from −69.04 *dB* to −53.34 *dB*) after the speaker, which is much smaller compared with the distortion caused by the audio amplifier.

We dig deeper to investigate the audible sound leakage caused by different audio amplifiers. Five amplifiers are tested, including Adafruit MAX98306 [12], Adafruit PAM8302A [17], Diodes PAM8406 [13], Texas Instruments TPA2016D2 [15], and Texas Instruments TAS2770 [14]. To make a fair comparison, we adjust the output power of each amplifier to be the same. Figure 5 illustrates the sound intensity levels of the audible leakage and the corresponding prices for each amplifier. We can observe that the audible sound leakage varies dramatically among amplifiers. The leakage can be significantly reduced (i.e., from 61.6 *dB* to 43.8 *dB*) if a proper audio amplifier is adopted with a slightly higher price ($1.262 for TAS2770).

Finding 2. *The maximum volume does not necessarily bring the maximum sensing range due to the audible sound leakage.* One critical issue associated with acoustic sensing is the small sensing range [28]. Researchers and developers usually adopt the maximum volume to achieve a larger sensing range. Figure 6 shows the magnitude changes of the sensing signal and distortion as the amplifier volume increases. We can observe that, when the volume is below 40%, most of the increased volume contributes to the power increase of the sensing signal. However, as we continue to



(a) 20 *kHz*. **(b)** 20 *kHz* **and** 21 *kHz*.

Figure 7: The non-linear distortion comparison between single-frequency and multi-frequency continuous waves.

increase the volume, the non-linear distortion increases rapidly and consumes most of the increased power, resulting in a much slower power increase of the sensing signal. Interestingly, when the volume reaches 100%, the power of the sensing signal even decreases due to large distortion. The results indicate that we should not turn up the volume to the maximum for sensing since higher volume might cause more distortion, resulting in a decreasing power for the sensing signal.

Finding 3. *The severity of the audible sound leakage is dependent on the signal type.* From Figure 1, we can observe that, among the three signal types commonly used for sensing, chirp signal induces the least amount of leakage. To figure out the reason why the audible leakage varies with signal type, we transmit two signals through our test platform, i.e., single-frequency continuous wave (20 *kHz*) and multi-frequency continuous waves (20 *kHz* and 21 *kHz*), respectively. Figure 7 shows the Power Spectral Density of the signals output by the audio amplifier. We can observe that, when playing a 20 *kHz* continuous wave, the amplifier generates harmonics [6] at integer multiples of the original signal frequency:

$$f_h = n \cdot f_1, \tag{1}$$

where $n \in \mathbb{Z}^+$, and f_1 denotes the original frequency. Given that $f_1 = 20\ kHz$, the additional harmonic components generated are at $f_h = 40\ kHz$, $60\ kHz$, $80\ kHz$, $100\ kHz$, $120\ kHz$, etc. Due to the limited sampling rate $f_s = 48\ kHz$ of the DAC, the harmonic components result in aliasing signals at

$$f_a = |kf_s - f_h|, \tag{2}$$

where $k \in \mathbb{Z}^+$. If we substitute k, f_s and f_h into Equation (2), we can obtain that f_a equals to 8 *kHz*, 12 *kHz*, 16 *kHz*, 4 *kHz*, and 24 *kHz*. However, if we introduce another 21 *kHz* signal, there also exists the intermodulation distortion among different frequency components [33], which generates a lot of extra distortions. Given that the original frequency $f_1 = 20\ kHz$ and $f_2 = 21\ kHz$, the intermodulation distortion results in additional signal components at the sum and difference frequencies of the original frequencies, e.g., $f_2 - f_1 = 1\ kHz$ and $f_1 + f_2 = 41\ kHz$. According to Equation (2), the 41 *kHz* frequency component results in an aliasing signal at 7 *kHz*. The above-mentioned analysis explains why the chirp signal produces the least amount of audible sound leakage because there is just one single frequency at each timestamp.

Finding 4. *The audible sound leakage issue can be mitigated by masking it with white noise or music.*[3] The audible sound leakage is very harsh and jarring, making it extremely annoying. We find that, the issue of leakage can be mitigated by adding a masking sound, i.e., either white noise or soft background music. To demonstrate its effectiveness, we play a continuous wave signal alone and also play it with the two masks, i.e., white noise and soft music. To make a fair comparison, we keep the leakage power constant by fixing the power of the continuous wave signal. Figure 8 illustrates the rating score of the comfortableness from 10 volunteers after listening to the three audios where each of them lasts for one minute. We can observe that, after applying the masking sound, the score can be improved from 1.6 to 3.1 (white noise mask) and 4.6 (music mask) on a scale of 5, respectively. The comfortableness was evaluated using a 5-point Likert scale [19] ranging from 1 to 5 according to the perceived feelings from the study participants: 1 denotes “Extremely annoying”, 2 denotes “Very annoying”, 3 denotes “Annonying”, 4 denotes “A little bit annonying”, and 5 is “I'm fine with it”.

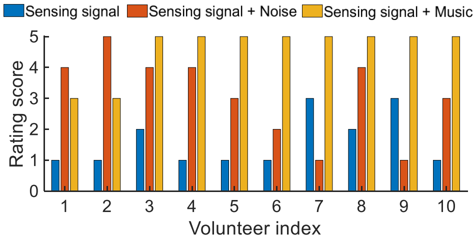


Figure 8: The comfortableness rating with and without sound masking for 10 volunteers.

Furthermore, to evaluate the effect of the sensing signal on the masking audio, we adopt the approach widely used in the field of speech recognition. Specifically, we chose 100 sentences from the public AISHELL-1 dataset [3] as the reference speech and added our sensing signal into the reference speech as the masked audio. Then we computed the quality of the masked audio using Perceptual Evaluation of Speech Quality (PESQ) whose score ranges from -0.5 (bad quality) to 4.5 (good quality) [46]. Note that the original speech audio has a score of 4.5. The average PESQ metric of the masked audio is 4.25, indicating that the sensing signal has little impact on the original speech. Furthermore, we also computed the intelligibility of the masked audio using Short-Time Objective Intelligibility (STOI) [36] whose score ranges from 0 (bad intelligibility) to 1 (good intelligibility). Note that the original speech audio has a score of 1.0. The average STOI score of the masked audio is 0.99, indicating that the effect of masking operation on audio intelligibility is negligible.

3 THE NEGATIVE IMPACT OF SENSING ON MUSIC PLAY/VOICE CALL

Previous acoustic sensing systems [5] assume that, during the sensing process, there should be no other types of audio like music played at the same time by the speaker, which constrains the wide adoption of acoustic sensing in real-world settings. For example, a user might want to perform hand gestures to switch songs via acoustic sensing when listening to music. Therefore, it is preferable if sensing and music play/voice call can happen at the same time. In this section, we report our experience and findings to achieve the above-mentioned objective.

[3] The demo audio can be found at https://www.youtube.com/watch?v=UIeMnARFBuw.

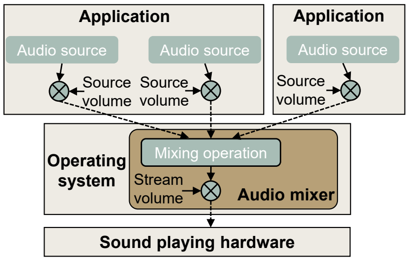


Figure 9: The general logic of audio mixing in existing operating systems.

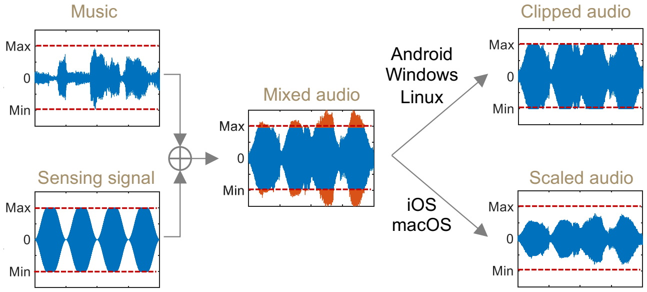
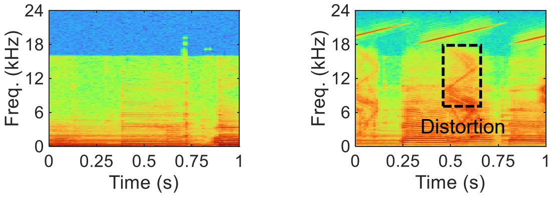


Figure 10: The "unfriendly" mixing strategies for audio mixer in existing operating systems.



(a) Music only. **(b) Sensing signals + music.**

Figure 11: Illustration of the degraded music quality when playing the sensing signal along with music on Sony Xperia.

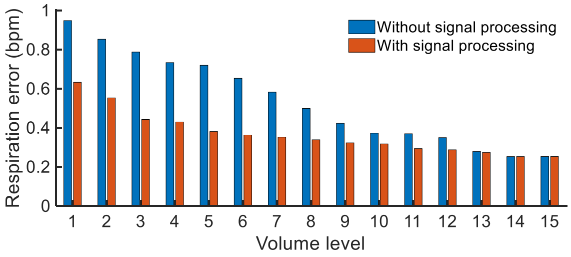


Figure 12: The respiration errors at different volumes with and without applying the advanced signal processing scheme.

3.1 Preliminary for Audio Mixing

Multiple audio sources created by either one application or several applications need to be mixed into one audio stream before being passed to the DAC as shown in Figure 3a. Figure 9 illustrates the general logic of audio mixing. Specifically, audio samples from multiple sources are multiplied by their corresponding source volume factors. Then they are passed to a software component called *Audio Mixer* [25]. After operations such as resampling, scaling, and adding at the Audio Mixer, multi-source audio samples are converted as one audio stream and then multiplied by the stream volume factor to control the total output volume.

3.2 Mixing the Sensing Signal with Music

To study the impact of acoustic sensing on applications such as music play, we conducted experiments on a diverse range of commodity devices running various operating systems, including Android, iOS, Windows, Linux, and macOS. Specifically, we play a song named "Twinkle, Twinkle, Little Star" and a sensing signal (i.e., chirp) sweeping from 18 kHz to 22 kHz together using two separate applications.

Finding 5. *Although existing operating systems support simultaneous playing of sensing signal and music/voice, the quality of music/voice does get affected.*[4] To address the issue of small sensing range, acoustic sensing usually adopts an extremely high volume (i.e., $\geqslant$ 80% of the maximum volume) to transmit the sensing signal. However, due to the limited digital representation range of the DAC,[5] the value of the mixed (added) audio is very likely to exceed the maximum value that the DAC can support. According to our experiments, existing audio mixers adopt "unfriendly" strategies to tackle this issue. As shown in Figure 10, for Android, Windows and Linux, the operating systems simply clip off the exceeding portion and set its value to the maximum allowable value of the DAC. Figure 11 demonstrates the negative impact of this operation, i.e., the degraded music quality. In contrast, iOS and macOS choose to scale down the volume of the audio sources, as shown in Figure 10. To eliminate the impact of the sensing signals on music play, we propose a simple yet effective scheme that adaptively tunes the volume of the sensing signals based on the volume of the music play, which makes sure the total volume does not exceed the maximum value. For example, if the music is using 60% of the maximum volume, we limit the volume of the sensing signal below 40% of the maximum volume. The above-mentioned strategy ensures that audio clipping and volume scaling down do not happen.

The sensing performance would be affected due to the decreasing power of the sensing signals when sensing and other audio applications co-exist. Advanced signal processing schemes [21] can be applied to increase the sensing range. We perform respiration monitoring using the Huawei P30 Pro smartphone to demonstrate the effectiveness of the proposed schemes. The smartphone is placed on a table, whose distance is 50 cm from the user. We play the chirp signal at different volume levels varying from level 1 to level 15 at a step size of 1 level. The chirp signal sweeps from 18 kHz to 22 kHz. Figure 12 presents the median respiration rate errors without and with applying the advanced signal processing schemes. We can observe that, even with only 33.3% (level 5) of the maximum volume, we can achieve a low respiration rate error of 0.36 breaths per minute (bpm), which is similar to the respiration error using 80%

[4] The audio demo can be found at https://youtu.be/lg23Bfdm4B0.
[5] A 16-bit DAC only outputs the value from -32768 (-2^{15}) to 32767 ($2^{15}-1$).

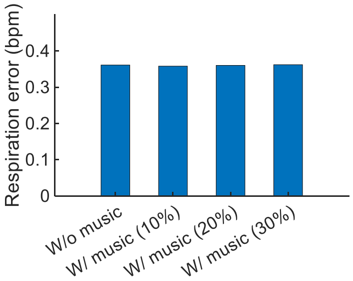


Figure 13: Impact of music play on respiration monitoring.

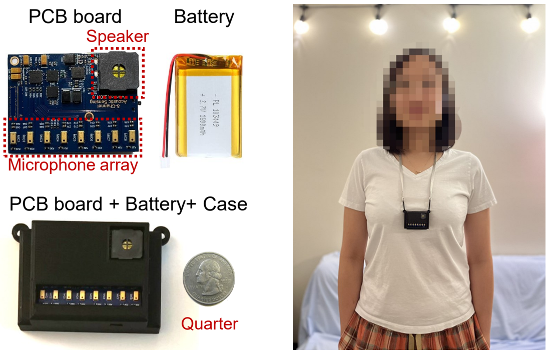
(level 12) of the maximum volume without applying the advanced signal processing scheme.

To evaluate the impact of music play on acoustic sensing, we asked one participant to sit at 50 *cm* in front of the smartphone and introduced music of three different volume levels when monitoring the participant's respiration. Specifically, we set the the volume of the sensing signal to 80% of the maximum volume. And the the volume of the music is set to 10%, 20% and 30% of the maximum volume, respectively. Note that for a music volume of 30%, clip happens and the quality of the music gets affected. As shown in Figure 13, the respiration errors for "W/o music" and "W/ music"at different volume levels are 0.361 *bpm*, 0.358 *bpm*, 0.360 *bpm* and 0.362 *bpm*, respectively. We do not observe any obvious difference among the four cases. This is because the frequency band adopted for sensing (> 18 *kHz*) is higher than that of music (< 16 *kHz*). We can easily remove the impact of the music by a bandpass filter. When clip happens, the sensing signal power is slightly reduced. However, as long as the sensing signal power is still above a threshold, the sensing performance is not affected.

4 LARGE POWER CONSUMPTION

Due to the wide availability of acoustic-enabled devices, acoustic sensing provides a unique opportunity to achieve 24/7 monitoring for human beings and their surrounding environments, e.g., smartphones can be applied for monitoring overnight sleep quality [32]. For long-term monitoring, the battery life is non-negligible for battery-powered devices. According to our experiments, after continuously sending out chirp signals at the maximum volume for two hours, we observe a 22% battery drop on a smartphone (Samsung S9+), a 78% battery drop on a smartwatch (Samsung Galaxy Watch 3), and a 66% battery drop on a wireless earphone (Apple Airpod Pro). This section shares our findings on how to reduce the power consumption for acoustic sensing based on our experience in deploying a self-designed device for social distance measurement during the COVID-19 pandemic. Note that the proposed solutions are general and can be applied to other acoustic sensing platforms.

Finding 6. *The power consumption of acoustic sensing can be significantly reduced if we introduce power control schemes.* We design and develop a lightweight, miniaturized wearable device that can be worn on the neck to measure social distance, as illustrated in Figure 14. The wearable device contains a self-designed printed circuit board (PCB) with a speaker and an array of 8 microphones. The PCB board is connected with a rechargeable lithium battery and held by a 3D-printed plastic case. To make sure the device can be worn comfortably for a long time, we miniature the PCB board through our meticulous design using 6-layer third-order high density interconnect (HDI) technology and the size of the PCB board is only 4.3 *cm* × 3.3 *cm*.



(a) Wearable Device. **(b) User case.**

Figure 14: The self-designed wearable device using acoustic sensing for the social distance measurement.

Just like existing acoustic sensing systems [20, 28], we initially transmit and receive the sensing signals unremittingly. However, according to our experiments, we observe that there is no human target in front of the wearer for more than 80% of the time when the wearer walks around the campus, indicating that most of the power is wasted during the sensing process. To address this problem, we apply a power control mechanism on wireless sensing. Specifically, we introduce the idle state where the sensing signals are transmitted and received at a much lower rate. Once a target is detected, the device enters into the active state where the sensing signals are transmitted and received at a normal rate.

Furthermore, existing acoustic sensing systems [5] adopt a fixed signal transmission power during the sensing process. However, the minimum required power for sensing a target at different distances varies a lot. The signal power required for sensing a close target is much smaller than that for sensing a far-away target. We, therefore, propose a scheme to adaptively tune the transmission power based on the distance between the target and the device. Specifically, we compute the distance information between the target and device using chirp-based acoustic ranging [20]. The chirp signal sweeps from 18 *kHz* to 22 *kHz* with a duration of 100 *ms* at a sampling rate of 48 *kHz*. Our self-designed wearable device can support 15 different levels of transmission power. We empirically determine the transmission power for each distance. It is worth noting that different applications require different levels of signal power. For example, as the reflection area is much smaller, hand tracking requires higher transmission power than human trajectory tracking. Therefore, power tuning strategy should be designed in consideration of the application type.

After applying the two proposed schemes, we can significantly reduce power consumption. Figure 15 illustrates a snapshot of the power consumption when a target gets close to the wearer, chats with the wearer, and then leaves. We can observe that, when there is no target around, the device is in an idle state. During the idle state,

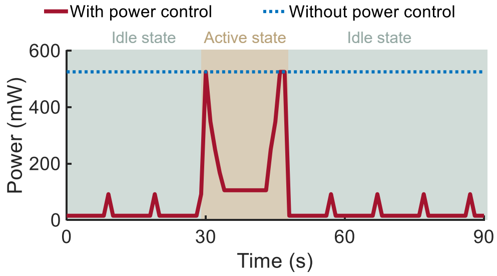


Figure 15: The illustration of power consumption when a human target gets close to the wearer, chats with the wearer, and then leaves.

Table 1: The average power consumption without and with applying power control schemes (unit: mW).

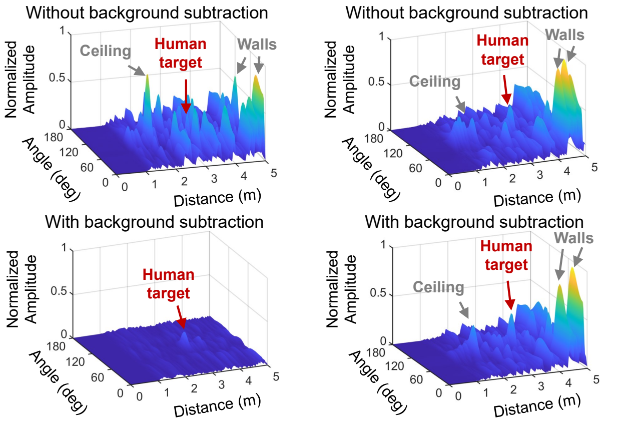
With/Without	Transmission	Reception	Processing	Total
Without	484.45	10.42	30.48	525.35
With	178.05	8.24	29.13	215.42

the power consumption is 15.58 mW when there are no sensing signals transmitted and 92.05 mW when the sensing signals are transmitted. When the device detects the target, it switches to the active state. With our adaptive power control scheme, the average power consumption is reduced from 525.35 mW to 215.42 mW. We further break down the power consumption into three parts, i.e., power consumption for signal transmission, power consumption for signal reception, and power consumption for signal processing. As shown in Table 1, our power control schemes mainly benefit from the power saving at signal transmission that is the most power-hungry part of acoustic sensing.

To demonstrate how the proposed power control schemes perform in real-life, we asked one participant to wear the device and walk around the campus for five hours. Compared with transmitting at the maximum power, the power consumption is reduced from 65.5% to 7.7% after applying the proposed schemes, which significantly extends the battery life. We also applied the proposed power control schemes when a user performs hand tracking using a Samsung S9+ smartphone for two hours. We repeat the experiment five times and the average power consumption is reduced from 22% to 10% after applying the proposed power control schemes. We believe power control is critical for the adoption of acoustic sensing in real life, and we move a step forward in this project.

5 HARMFUL IMPACT GENERATED BY DEVICE MOBILITY

Existing sensing systems assume that the sensing device keeps stationary during the sensing process [5]. For example, the sensing device is put on a table or mounted on a tripod. This is not always true, and we encounter quite a few real-world scenarios where the device is moving. For example, for social distance measurement, the device is worn by a wearer and moves when the wearer walks around and interacts with others. However, the sensing performance severely degrades when the device is not stationary. We believe this is a critical issue that needs to be tackled before we can push acoustic sensing to real-world settings. In this section, we share our findings of how to remove the harmful impact of device motions based on our experience on social distance measurement.



(a) Device on the table.

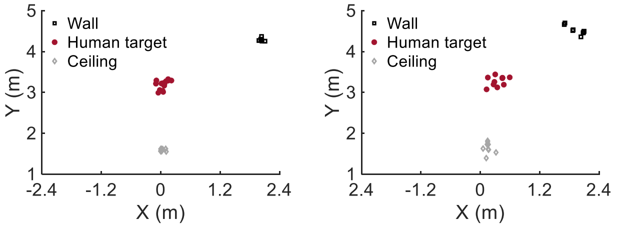
(b) Device on the wearer.

Figure 16: The background subtraction can remove the reflections from static objects when the device is stationary, while does not work when the device is moving.

Finding 7. *The movements of sensing device significantly degrade the performance of acoustic sensing.* Most existing research on acoustic sensing focuses on improving the accuracy [38, 39, 42], where the state-of-the-art studies have pushed the accuracy to sub-millimeter level [20]. However, we find that centimeter-level accuracy is good enough for a lot of real-world applications including social distance measurement. The real practical challenge we encounter in social distance measurement is that, in the presence of device movements, we cannot even differentiate between a human target and a static object such as a pillar. When the device is stationary, we can use the Doppler information, i.e., Doppler value is zero to identify static objects. However, when the device is moving, even signals reflected from static objects exhibit non-zero Doppler values.

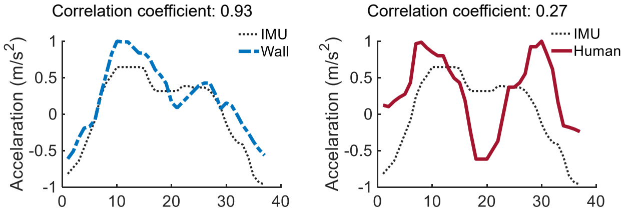
Two approaches were adopted by prior studies to eliminate the impact of static objects on target sensing. The first approach is to apply background subtraction [20, 28] to remove the reflection from the static objects using acoustic signals received at two adjacent timestamps. Figure 16a illustrates the range-angle profile when the sensing device is on the table, and a human target stands at 3 m in front of the device in a typical indoor environment. We can observe that background subtraction works when the device is stationary since the reflections from the static objects remain constant over time. However, when the device is moving, i.e., the device is worn by a user, the reflections from the static objects are not constant anymore, failing the background subtraction approach, as shown in Figure 16b.

The second approach is to differentiate the moving human target from static objects based on the variation of the estimated positions [21, 31, 43]. Figure 17 illustrates the extracted positions for the human target and the static objects when the device is stationary and moving, respectively. We can observe that, when the device is



(a) Device on the table. **(b) Device on the wearer.**

Figure 17: The variation of the extracted positions can differentiate human target from the static objects when the device is static, while becomes invalid when the device is moving.



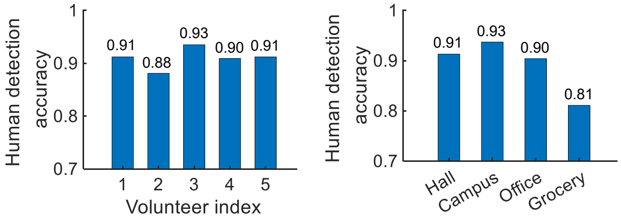
(a) Wall. **(b) Human target.**

Figure 18: Comparison between the acceleration data collected from the IMU sensor and that computed from acoustic signals reflected from human target and wall.

stationary, it is straightforward to differentiate the human target from other static objects. As shown in Figure 17a, the distances between the device and static objects (i.e., wall and ceiling) are more stable, while the distances between the device and the human target change a lot when the device is stationary. When the device is worn by a user, even if the wearer just stands still, the involuntary movement of the body [40] changes the position of the device, resulting in large distance changes for both the human target and the static objects, as shown in Figure 17b.

We involve the inertial measurement unit (IMU) sensor to address the device movement issue. Owing to its cheapness, miniaturization and low power consumption, IMU sensors widely exist in acoustic-enabled devices, including smartphones, smart watches and even earphones. Therefore, besides speaker and microphones, we also equip our custom-designed wearable device with IMU sensors. During the sensing process, both acoustic data and IMU data are collected and stored for post-processing. Since the sampling rates of the microphones and IMU sensors are different, we resample the IMU data using the cubic spline interpolation algorithm [29]. We adopt the timestamps provided by the high-accuracy Real-Time Clock (RTC) module inside the device to synchronize the acoustic data and IMU data.

To differentiate human target from static objects, we compare the acceleration data collected from IMU sensors with that computed using acoustic signals. As shown in Figure 18a, since the variations of the IMU sensor data and acoustic signals reflected from static object are only caused by device movement, they are highly correlated, i.e., the correlation coefficient is 0.93. In contrast, the variation of the acoustic signals reflected from the human target is caused not only by the device movement but also by the target movement. Therefore, it has a low correlation with the variation of the IMU sensor data, i.e., the correlation coefficient is only 0.27, as shown in Figure 18b.



(a) Effect of volunteer. **(b) Effect of environment.**

Figure 19: The human detection accuracy for different volunteers and environments, respectively.

By applying the above-mentioned method, we can accurately differentiate human target from static objects. We recruited five volunteers to conduct experiments at four different environments, including dining hall, campus, office, and grocery store. They were asked to wear the device shown in Figure 14a and walk around in each of the environments for 30 minutes. Furthermore, we asked one volunteer to follow the wearer to record the timestamps when the wearer interacts with other humans as the ground truths. Figure 19a and Figure 19b illustrate the human detection accuracy for five volunteers and four different environments, respectively. By applying the proposed method, we can achieve higher than 90% human detection accuracy in most scenarios. We observe a slightly lower human detection accuracy when the wearer walks in the grocery store due to much more complicated environment (e.g., crowded aisles).

6 RELATED WORK

This section summarizes the prior studies related to the four practical problems in the real-world settings.

Audible Sound Leakage. A lot of efforts [6, 11, 33, 34, 44] have been devoted to studying the non-linearity of the microphone hardware, i.e., the ultrasound signal can create a "shadow" signal in the audible frequency range after passing the microphone hardware. This phenomenon can enable applications such as jamming spy microphone, live watermarking of music, etc. However, none of them pay their attention to the non-linearity in the speaker hardware, which generates the audible sound leakage when playing inaudible sensing signals on commodity devices.

Coexistence of Sensing and Music Play/Voice Call. Prior studies [21, 30] demonstrate the feasibility of playing the sensing signals and music simultaneously. However, the multiple audio sources, i.e., music and sensing signals, are synthesized as one audio source before they are fed into the audio mixer on the operating system, which is impractical in real-world settings. Furthermore, we broadly explore the behaviors of the audio mixer in different operating systems and provide feasible solutions to reduce the impact of the sensing signals on music.

Large Power Consumption. The power control mechanisms have been extensively studied in the field of wireless communication [2] and backscatter communication [26]. We explore the feasibility of applying the power control schemes to extend the battery life for sensing applications on battery-powered devices.

Harmful Impacts Caused by Device Mobility. There are some previous studies discussing the device mobility issue for acoustic sensing. SpiroSonic [35] adaptively removes the distortion in the I/Q signals caused by hand movement when a user holds a smartphone to monitor lung functions. BreathListener [41] extracts the respiration signals based on the Energy Spectrum Density of signals and removes the interference from the driving environments using the background subtraction and Ensemble Empirical Mode Decomposition schemes. Different from the previous studies, the device displacement caused by human body movement is much larger, which cannot be removed by simply applying their methods.

7 DISCUSSION

The audible sound leakage is mainly a business challenge. We have showed that the non-linear distortion of the audio amplifier is the chief culprit of the audible leakage. A better amplifier can significantly reduce the amount of leakage. Therefore, the smartphone manufacturers can choose a better amplifier to address the leakage issue. The current speakers are designed for voice call and music play rather than acoustic sensing. If acoustic sensing becomes a mainstream function of future speakers, this issue can easily be fixed by smartphone manufacturers.

The negative impact of sensing signal on music play and voice call is a technical challenge. We have shown that the audio mixers in the existing operating systems adopt "unfriendly" strategies to tackle the signal overflow issue. We propose a simple yet effective scheme that tunes the volume of the sensing signals based on the maximum volume of the music play/voice call, which makes sure the total volume does not exceed the maximum value. We believe that it is possible to adjust the sensing signal power in a more fine-grained manner based on the instantaneous power of the voice/music. However, it requires real-time prediction of the voice/music power variation, which is very challenging. Investigation of fine-grained power tuning remains an important future research direction.

The large power consumption is a technical challenge. We proposed some straightforward schemes to reduce the power consumption for battery-powered devices. While the proposed challenges are general, the solutions to different applications need to vary to cope with the unique characteristics of each individual application. For example, the proposed distance-based power control scheme cannot be applied to sensing applications [8, 9] using earphones since the distance between the earphone and ear canal does not change during the sensing process. One possible solution to save power is that we can start the acoustic sensing process only when a targeted event is detected and employ the low-power IMU sensor to serve as the trigger.

The negative impact generated by device mobility is a technical challenge. We demonstrated one possible solution to solve the issue for social distance measurement. For other applications such as earable sensing, new solutions need to be proposed to address the device motion issue. The IMU data is usually too coarse to be utilized to cancel the device motion for fine-grained ear canal monitoring. Feeding the IMU data and acoustic data into deep learning networks may be the direction to explore.

8 CONCLUSION

This paper shares our experience and findings on multiple practical challenges during the process of developing and deploying acoustic sensing systems in real-world settings. Four challenges are discussed, including annoying audible sound leakage, negative impact of sensing on music play/voice call, large power consumption, and degraded performance in the presence of device motions. We hope our insights can trigger future efforts that are devoted to solving real-world challenges on acoustic sensing and pushing acoustic sensing from the laboratory to the real world.

ACKNOWLEDGMENTS

This work is partially supported by National Institutes of Health (NIH) under Award Number 5R01MH122371-02.

REFERENCES

[1] Yang Bai, Li Lu, Jerry Cheng, Jian Liu, Yingying Chen, and Jiadi Yu. 2020. Acoustic-based sensing and applications: A survey. *Computer Networks* 181 (2020), 107447.

[2] Martin Bor and Utz Roedig. 2017. LoRa transmission parameter selection. In *2017 13th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 27–34.

[3] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*. IEEE, 1–5.

[4] Chao Cai, Zhe Chen, Henglin Pu, Liyuan Ye, Menglan Hu, and Jun Luo. 2020. Acute: Acoustic thermometer empowered by a single smartphone. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 28–41.

[5] Chao Cai, Rong Zheng, and Jun Luo. 2022. Ubiquitous Acoustic Sensing on Commodity IoT Devices: A Survey. *IEEE Communications Surveys & Tutorials* (2022).

[6] Yuchi Chen, Wei Gong, Jiangchuan Liu, and Yong Cui. 2018. I can hear more: Pushing the limit of ultrasound sensing on off-the-shelf mobile devices. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2015–2023.

[7] Sony Group Corporation. 2022. *Sony PCM-D100 High Resolution Audio Recorder*. https://pro.sony/ue_US/products/portable-digital-recorders/pcm-d100

[8] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2021. EarGate: gait-based user identification with in-ear microphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 337–349.

[9] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using ear canal echo for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–24.

[10] Zhihui Gao, Ang Li, Dong Li, Jialin Liu, Jie Xiong, Yu Wang, Bing Li, and Yiran Chen. 2022. MOM: Microphone based 3D Orientation Measurement. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 132–144.

[11] Yitao He, Junyu Bian, Xinyu Tong, Zihui Qian, Wei Zhu, Xiaohua Tian, and Xinbing Wang. 2019. Canceling inaudible voice commands against voice control systems. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–15.

[12] Analog Devices Inc. 2022. *Class D Audio Amplifier - MAX98306*. https://datasheets.maximintegrated.com/en/ds/MAX98306.pdf

[13] Diodes Incorporated. 2022. *Class D Audio Amplifier - PAM8406*. https://www.diodes.com/assets/Datasheets/PAM8406.pdf

[14] Texas Instruments Incorporated. 2022. *Class D Audio Amplifier - TAS2770*. https://www.ti.com/lit/ds/symlink/tas2770.pdf

[15] Texas Instruments Incorporated. 2022. *Class D Audio Amplifier - TPA2016D2*. https://www.ti.com/lit/ds/symlink/tpa2016d2.pdf

[16] Texas Instruments Incorporated. 2022. *PCM5242RHBEVM DAC Evaluation Module*. https://www.ti.com/tool/PCM5242RHBEVM

[17] Adafruit Industries. 2022. *Class D Audio Amplifier - PAM8302*. https://www.adafruit.com/product/2130

[18] Pulsar Instruments. 2022. *Decibel chart.* https://pulsarinstruments.com/news/decibel-chart-noise-level/

[19] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology* 7, 4 (2015), 396.

[20] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2020. FM-track: pushing the limits of contactless multi-target tracking using acoustic signals. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems.* 150–163.

[21] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2022. LASense: Pushing the Limits of Fine-grained Activity Sensing Using Acoustic Signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.

[22] BDNC (HOLDING) LIMITED. 2022. *General Purpose Speaker.* http://www.newbdnc.com/wp-content/uploads/datasheets/BFC-4448-24-4-006.pdf

[23] Jialin Liu, Dong Li, Lei Wang, and Jie Xiong. 2021. BlinkListener: " Listen" to Your Eye Blink Using Your Smartphone. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–27.

[24] Jialin Liu, Dong Li, Lei Wang, Fusang Zhang, and Jie Xiong. 2022. Enabling Contact-free Acoustic Sensing under Device Motion. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–27.

[25] Google LLC. 2022. *Android Audio Mixer.* https://android.googlesource.com/platform/frameworks/av/+/refs/heads/master/media/libaudioprocessing/AudioMixerBase.cpp

[26] Amjad Yousef Majid, Patrick Schilder, and Koen Langendoen. 2020. Continuous sensing on intermittent power. In *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN).* IEEE, 181–192.

[27] Wenguang Mao, Wei Sun, Mei Wang, and Lili Qiu. 2020. DeepRange: Acoustic Ranging via Deep Learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–23.

[28] Wenguang Mao, Mei Wang, Wei Sun, Lili Qiu, Swadhin Pradhan, and Yi-Chao Chen. 2019. RNN-Based Room Scale Hand Motion Tracking. In *The 25th Annual International Conference on Mobile Computing and Networking.* ACM, 38.

[29] Sky McKinley and Megan Levine. 1998. Cubic spline interpolation. *College of the Redwoods* 45, 1 (1998), 1049–1060.

[30] Rajalakshmi Nandakumar, Alex Takakuwa, Tadayoshi Kohno, and Shyamnath Gollakota. 2017. Covertband: Activity information leakage using music. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–24.

[31] Kun Qian, Chenshu Wu, Fu Xiao, Yue Zheng, Yi Zhang, Zheng Yang, and Yunhao Liu. 2018. Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices. In *IEEE INFOCOM 2018-IEEE conference on computer communications.* IEEE, 1574–1582.

[32] Yanzhi Ren, Chen Wang, Yingying Chen, Jie Yang, and Hongwei Li. 2019. Noninvasive fine-grained sleep monitoring leveraging smartphones. *IEEE Internet of Things Journal* 6, 5 (2019), 8248–8261.

[33] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. 2017. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services.* 2–14.

[34] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. 2018. Inaudible Voice Commands: The {Long-Range} Attack and Defense. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18).* 547–560.

[35] Xingzhe Song, Boyuan Yang, Ge Yang, Ruirong Chen, Erick Forno, Wei Chen, and Wei Gao. 2020. SpiroSonic: monitoring human lung function via acoustic sensing on commodity smartphones. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking.* 1–14.

[36] Ke Sun and Xinyu Zhang. 2021. UltraSE: single-channel speech enhancement using ultrasound. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking.* 160–173.

[37] VLIKE. 2022. *VLIKE LCD Digital Sound Level Meter.* https://www.amazon.com/VLIKE-Digital-Measurement-Measuring-Function/dp/B01N2RLJ32

[38] Anran Wang, Jacob E Sunshine, and Shyamnath Gollakota. 2019. Contactless infant monitoring using white noise. In *The 25th Annual International Conference on Mobile Computing and Networking.* 1–16.

[39] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking.* ACM, 82–94.

[40] David A Winter. 1995. Human balance and posture control during standing and walking. *Gait & posture* 3, 4 (1995), 193–214.

[41] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. Breathlistener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services.* 54–66.

[42] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services.* ACM, 15–28.

[43] Fusang Zhang, Zhi Wang, Beihong Jin, Jie Xiong, and Daqing Zhang. 2020. Your Smart Speaker Can" Hear" Your Heartbeat! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–24.

[44] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security.* 103–117.

[45] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2021. SoundLip: Enabling Word and Sentence-level Lip Interaction for Smart Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–28.

[46] Qian Zhang, Dong Wang, Run Zhao, Yinggang Yu, and Junjie Shen. 2021. Sensing to hear: Speech enhancement for mobile devices using acoustic signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–30.