Cocoon: On-body Microphone Collaboration for Spatial Awareness

Bhawana Chhaglani*
UMass Amherst, USA
bchhaglani@umass.edu

Utku Günay Acer
Nokia Bell Labs, Belgium
utku_gunay.acer@nokia-bell-labs.com

Si Young Jang
Nokia Bell Labs, UK
siyoung.jang@nokia-bell-labs.com

Fahim Kawsar
Nokia Bell Labs, UK
fahim.kawsar@nokia-bell-labs.com

Chulhong Min
Nokia Bell Labs, UK
chulhong.min@nokia-bell-labs.com

ABSTRACT

We are now surrounded by multiple microphones available on various devices around us including wearables. This opens unique opportunities for acoustic sensing applications to enhance the spatial resolution of audio signals on the body. However, state-of-the-art acoustic sensing applications still mostly utilize a single microphone or a microphone array on a single device. In this paper, we present *Cocoon*, a case study for on-body microphone collaboration for spatial awareness. Cocoon is a novel wearable system that automatically provides users with situational services based on their location. To this end, it combines spatial profiles from multiple on-body microphones on the fly and identifies a user's location by matching the profile against the pre-registered ones. Our experimental results show that Cocoon outperforms the existing single microphone-based methods, 10.0% points and 21.5% points accuracy increase in the controlled and real-world setup, respectively. *Cocoon* also improves the robustness to slight movements and orientation changes of the microphone, reducing the error rate by 17.5% points.

CCS CONCEPTS

• **Human-centered computing** → **Mobile devices**; *Collaborative and social computing devices.*

KEYWORDS

Microphone, Wearable Collaboration, Spatial Sensing

ACM Reference Format:
Bhawana Chhaglani, Utku Günay Acer, Si Young Jang, Fahim Kawsar, and Chulhong Min. 2023. Cocoon: On-body Microphone Collaboration for Spatial Awareness. In *The 24th International Workshop on Mobile Computing Systems and Applications (HotMobile '23), February 22–23, 2023, Newport Beach, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3572864.3580340

*This work was done when the author was on an internship at Nokia Bell Labs.

HotMobile '23, February 22–23, 2023, Newport Beach, CA, USA

ACM ISBN 979-8-4007-0017-0/23/02…$15.00
https://doi.org/10.1145/3572864.3580340

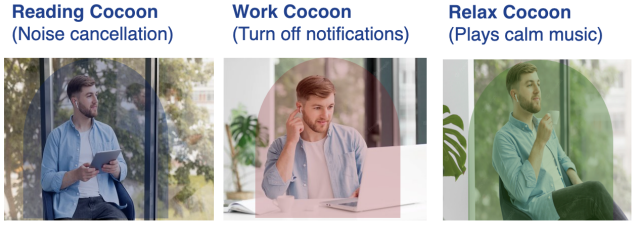

Figure 1: Operational scenarios of *Cocoon*

1 INTRODUCTION

Thanks to its versatile capability, acoustic sensing has received significant attention over the decades. Recently, a number of interesting acoustic sensing applications have been proposed, addressing a variety of problems including localization [9, 17, 20], user interface [1, 16, 22], ventilation sensing [5], temperature sensing [21], human activity recognition [14], safety [7], health sensing [6, 15, 25], authentication [4, 24] and surrounding event detection [10].

Due to their ultra-compact form factor, microphones are also embedded in many smart devices, e.g., smartphones, smart earbuds, smartwatches, and other Internet of Things (IoT) devices. It is now common to find ourselves surrounded by multiple microphones, which can make acoustic sensing even more ubiquitous. However, despite the ubiquitous nature of microphones, the capability of acoustic sensing on the body is still often limited to a single device.

The use of multiple microphones offers exciting opportunities for acoustic sensing [2, 3]. It can provide robust and reliable sensing by selectively choosing the best-quality audio streams [13]. It can also reduce interference and improve the quality of speech signals by focusing a receiving radiation pattern in the direction of the desired signal [26]. In addition, it enables applications to capture spatial features of audio signals by measuring sound waves in both time and space. While these ideas have been realized on microphone arrays on a single device, e.g., active noise cancellation on earbuds and speech enhancement on smart speakers such as Amazon Alexa, very few attempts have been made using multiple wearable devices.

This paper presents *Cocoon*, a case study of on-body microphone collaboration for spatial awareness. We define "Spatial awareness" as a key factor that enables user-defined region-specific services, i.e., whenever a user enters a dedicated space, the functions associated with that space automatically begin. *Cocoon* is a novel wearable system that automatically provides users with situational

services based on their location, as shown in Figure 1. For example, a user can register specific places as a brainstorming region (e.g., a working desk, a sofa), and it automatically activates active noise cancellation and turns off notifications when the user enters the region. To enable spatial awareness, *Cocoon* combines multiple on-body microphones on the body and creates spatial profiles using the sound they capture. Then, it identifies a user's location (whether a user stays in the registered region) by matching the profile with the pre-registered one.

To enable spatial awareness, we adopt an impulse response as a spatial profile, inspired by acoustic fingerprinting [18, 19]. Fingerprinting mechanisms are based on the principle that, while a speaker emits a chirp signal, the sound captured by the microphone is used to calculate the impulse response of the acoustic medium between the source of the signal and the microphone. In *Cocoon*, a pair of earbuds with microphones captures the chirp signal from a nearby smart speaker. Using the captured sound, each microphone computes the impulse response, associated with the sound. Then, *Cocoon* extracts key features such as peak amplitudes and peak time differences from both signals and runs classification on each of these features. The final classification is made using majority voting.

2 BACKGROUND AND RELATED WORK

Microphone collaboration: The idea of utilizing multiple microphones has been intensively studied using microphone arrays for a few decades [2, 3]. The most representative technique developed with microphone arrays is acoustic beamforming. It captures spatial samples of the propagating wave from multiple microphones placed at different spatial locations and manipulates them to enhance the target sound source or suppress unwanted interference. Microphone arrays are common now in mobile, wearable devices and smart-home appliances, and are actively used for audio source separation and enhancement [23]. Despite such benefits, few attempts have been made yet to extend this mechanism to multiple microphones on on-body wearables. Authors of [26] studied beamforming for meeting transcription on asynchronous audio-capturing devices such as mobile phones and laptops, and showed that the word error rate decreases as the number of deployed microphones increases. However, it does not apply to on-body wearable scenarios due to their unique challenges, such as constantly moving and rotating on-body microphones. In this paper, as an initial attempt to realize on-body microphone collaboration, we target spatial awareness, i.e., instantly identifying a user's surroundings, and show its feasibility through the end-to-end implementation.

Acoustic fingerprinting on mobile devices: Acoustic fingerprinting has been investigated for indoor localization using audio signals. EchoTag [20] enables phones to tag and remember indoor locations by generating acoustic signatures, which are obtained by transmitting a sound signal with a phone's speakers and then sensing its reflections with the phone's microphones. SweepSense [9] sweeps through a range of inaudible frequencies and measures the intensity of reflected sound to deduce information about the immediate environment, chiefly the materials and geometry of proximate surfaces. Authors of [17] proposed a single-step calibration-free fast indoor space mapping solution that quickly maps an indoor space

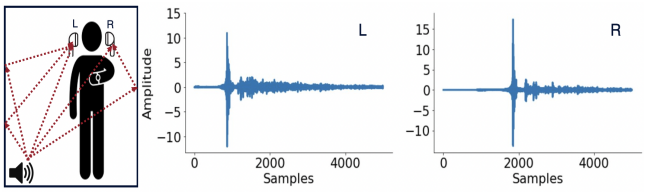

Figure 2: Example of spatial profiles captured from two microphones L (left) and R (right).

by simply walking around while holding a phone in a user's hand. Despite extensive and rich literature on audio localization, no prior work has used multiple on-body microphones for spatial awareness and fingerprinting. Here, we show that we can achieve enhanced spatial resolution with their collaboration. Note that Cocoon is not a competing technology with the existing acoustic fingerprinting techniques, but it complements them.

3 MICROPHONE COLLABORATION

3.1 Multi-microphone Fingerprinting

While acoustic fingerprinting-based localization can be implemented on a single microphone or multiple microphones on a single device (e.g., a smartphone and a noise-cancelling earbud), using multiple wearable microphones leads to the following benefits.

- **Enhanced spatial resolution**: Wearables are naturally worn apart from each other, which enhances the spatial resolution of sensing. Figure 2 illustrates different impulse responses captured by two microphones on earbuds, which shows the reflections of sound signals; we explain its details in §4. We further show that multiple microphones can capture richer information when they are far apart. Since wearables are significantly farther apart than multiple microphones equipped with a single device, they can provide additional information about surroundings.
- **Increased feature dimensionality**: As we can sense multiple spatial profiles from multiple vantage points, we have more feature dimensionality to better distinguish multiple locations. Figure 2 shows that the dominant reflections arriving at the two microphones will be significantly different.
- **Robustness to slight movements**: Multiple microphones will be more robust and reliable in detecting slight movements than a single microphone. This is because slight changes are very unlikely to cause a significant difference in all the captured profiles simultaneously.

3.2 Challenges

Despite the aforementioned benefits, it is not straightforward to realize the collaboration of multiple microphones on different wearables. We present several challenges that need to be addressed for on-body microphone collaboration.

- **Time synchronization:** It is important to maintain tight time-synchronization of multiple audio streams for microphone collaboration. However, many small-form factor wearable platforms such as commercial earbuds and smart wristbands do not support time-synchronization capability for audio streaming. Also, even

with powerful devices supporting time-synchronization, such as smartphones and smartwatches, it is not easy to synchronize the audio streaming from different devices as the audio streams do not contain the timestamp information.

- **User movement:** Since wearables are worn on the body, they move and rotate as the wearer moves around. This poses two problems. First, the topology of microphones dynamically changes causing relative differences in angles and distances between microphones. Second, the spatial profiles can be corrupted if microphones move around while capturing acoustic signals.
- **Efficient and robust collaboration:** The most representative method for processing multiple data streams is the sensor fusion technique, i.e., designing a fusion model with concatenated data streams. However, they often demand significant system resources as the data streams need to be continuously transmitted. Also, more importantly, the fusion model needs to be re-trained from scratch if a new device is added or an existing device is removed, which is impractical.

4 COCOON SYSTEM

Cocoon enables spatial awareness via wearable microphone collaboration. Specifically, we chose earables (e.g., [8]) as target devices because a pair of earbuds are constantly distanced apart, leading to a good spatial resolution, relatively stable and socially acceptable. As a sound source, we use a nearby smart speaker under the assumption that the speaker is located in a fixed position.

The overall operational flow consists of two stages. First, a user registers regions of interest where she wants to use situational services. Once the registration is requested, *Cocoon* collects the spatial profile from two earbuds and uses it as training data during the training phase. Note that our objective is not localization, thus *Cocoon* does not need to collect the data for all positions in a place. Second, during the online phase, *Cocoon* detects if a user enters the pre-registered regions by monitoring the spatial profiles. We assume that a smart speaker in a place periodically emits a chirp sound or is triggered by a switch or gesture. Then, situational services (e.g., muting notification) are activated based on the detection results.

Figure 3 shows the *Cocoon* pipeline with 5 main components: (1) Spatial sensing, (2) Spatial profile alignment, (3) Feature extraction and ranking, (4) local classification and (5) collaboration.

Spatial sensing: The first step is to sense spatial profiles using two microphones present on each ear. To enable spatial awareness, we use *impulse response* (IR) as acoustic fingerprint features for the following reasons. First, it captures the reflections of sound signals in terms of amplitude and time shift, which indicates the distance and amplitude of acoustic reflections coming from walls and objects in the vicinity. Second, it can be generated from the acoustic signals captured with a very short duration (in our implementation, 0.02s), which is robust to the movement of microphones. We compute IR using the Exponential Sine Sweep (ESS) technique. We first emit a chirp signal from the speaker with a frequency range of 6-20 kHz, a duration of 0.02s, and a sampling rate of 48 kHz and simultaneously capture the microphone signal. The speaker keeps sending this ESS signal every 2 seconds. Then, we convolve the captured microphone signal with the inverse of the transmitted chirp signal. This resulting IR contains peaks corresponding to a direct path and reflections from nearby walls and objects, as shown in Figure 2. Thus, we can expect that the impulse response would be similar if a speaker and microphones stay at the same location. The transmitted sine sweep signal is given by:

$$x(t) = sin\left(\frac{2\pi f_1 T}{R}\left(e^{\frac{tR}{T}} - 1\right)\right)$$

where f_1, f_2 are the initial and final frequency of the sine sweep signal, T is the duration of the sweep, and $R = \ln(f_2/f_1)$ is the sweep rate. Now, the inverse filter is calculated by scaling the amplitude of time-reversed $x(t)$ by $k = \exp(tR/T)$ which will give $f(t) = x_{inv}/k$. Finally, we get IR $h(t)$ by: $h(t) = s(t) * f(t)$ where $s(t)$ is the signal captured by the microphone.

Spatial profile alignment: After capturing the IR from both microphones, we perform a number of signal processing steps before extracting features. First, we synchronize the two IRs captured from the two microphones. To do this, we identify the direct path or the strongest reflection observed at both microphones and remove the signal before the highest peak from both IRs. Next, we clip the signal after 6000 samples as this is equivalent to a reflection coming from 42 meters away, which is not practical. Figures 4 (a) and (b) show the signal representation before and after this pre-processing step, respectively.

Feature extraction and ranking: After pre-processing, we extract features from the signal for the classification model. As the peaks in the IR correspond to the reflections coming from the nearby walls and objects, we select the prominent peaks in the signal because we can get more robust and distinguishable features with stronger signals. Figure 4(c) shows the prominent peaks (in red) extracted from the signal. Features such as peak amplitudes, peak times or peak time differences can be chosen to represent the selected peaks. Peak amplitude is the amplitude corresponding to the selected prominent peaks, peak times are the time values corresponding to the peaks, and peak time difference is the time difference between consecutive prominent peaks. We experimentally find that the peak amplitude provides the best accuracy, whereas peak time offers the worst. Hence, we use peak amplitudes and peak time differences as our features.

Local classification: We use the K-nearest neighbours based Time Series Classification (TSC). For the distance metric between the captured sound and the training data, we rely on Dynamic Time Warping (DTW). This metric computes the similarity between two temporal sequences. We train ML models for both peak amplitudes and peak time differences, for each microphone. In other words, we train 4 classifiers using four different sets of features, peak amplitudes (L), peak time differences (L), peak amplitudes (R), and peak time differences (R). These models take the respective features as input and predict a user's location or place.

***Cocoon* collaboration**: For combining spatial profiles, we consider two techniques. The first is feature concatenation, where we train a model comprehensively on all the features of both left and right microphones together. The second is ensemble learning, where we train individual models for each microphone and then combine their predictions. We compare these two collaboration techniques from the system's point-of-view and choose the ensemble learning technique as it has various advantages, such as less communication requirement and flexibility to adding/missing wearables cases. The

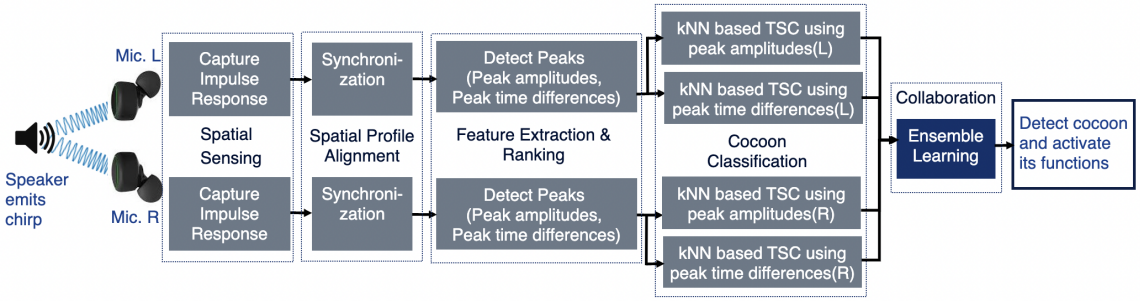

Figure 3: *Cocoon* overview

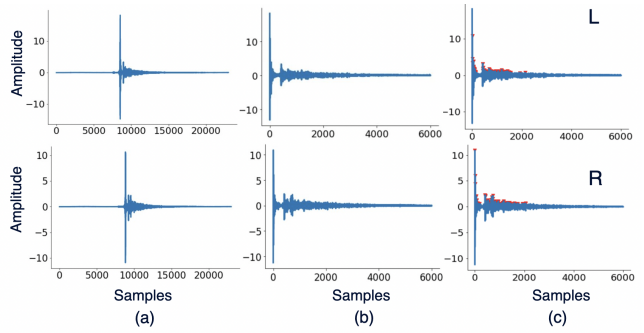

Figure 4: (a) Captured IR (b) after alignment (c) feature extraction

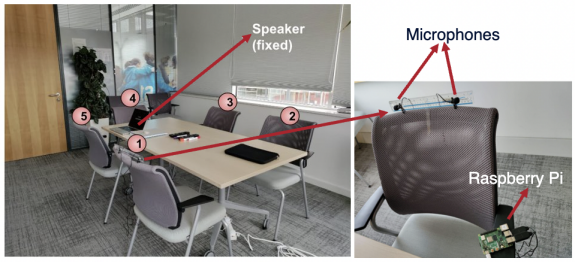

Figure 5: Data collection setup

Methods	Controlled setup	Real-world setup
Cocoon	100%	98%
L (Left)	90%	80%
R (Right)	90%	73%

Table 1: Accuracy with single and dual microphones in controlled and real-world setup.

proposed system fits one classifier for each time series data and then aggregates their predictions. *Cocoon* brings together classifications from 4 different models. We then use majority voting to come up with the final outcome on the spatial profile; in a tie situation, we follow the classifier with the highest confidence value.

5 EVALUATION

We performed experiments using the *Cocoon* prototype to validate its performance. In this section, we describe our prototype and data collection setup. Next, we show the accuracy and robustness of *Cocoon* in comparison with the single microphone-based method. For in-depth analysis, we further compare the detection accuracy between audible and inaudible emitting sounds.We develop the *Cocoon* prototype using two Raspberry Pi 4 boards, each of which is connected to a USB microphone, and a MacBook laptop. The USB microphones allow 16-bit recording and support a sampling rate of up to 48 kHz. We use the laptop as a speaker that emits the ESS signal with a delay of two seconds.

We use 2 different setups for data collection. Since real-life sound sensing is affected by a number of factors simultaneously, we first conduct a comprehensive study using the dataset collected in a *controlled* setup in order to isolate the impact of other factors and also better control the level of the movement. Then, we evaluate the performance and robustness of *Cocoon* using a *real-world* setup.

Controlled setup: In the controlled setup, the two USB microphones are placed on a ruler 20 cm apart, as shown in Figure 5. For data collection, we collected data in 3 rooms: 2 conference rooms and 1 bedroom. For each of these three rooms, the laptop (speaker) was kept at a fixed location inside the room, and we collected data for 20 positions in total (which can be assumed as a region of interest). For each position, we collected the spatial profiles 10 times. We collect data in two ways: one subset of the dataset consists of IRs captured without any movement, and the second subset of the dataset consists of IRs captured when the microphones are slightly moved or changed in orientations.

Real-world setup: In the real-world setup, microphones are clipped to human ears. We collect the data in the conference room at five different locations with two participants. Note that we used USB microphones and clipped them on ears because no commercial earables provide yet access to simultaneous audio streams of their microphones from two earbuds via Bluetooth classic. For the data collection, the participant goes to the location, sits stationary on the chair, and we record the spatial profile data.

5.1 Collaboration Accuracy

Table 1 shows the accuracy of the *Cocoon* detection in the controlled and real-world setup. For the controlled setup, *Cocoon* achieves the same accuracy, precision, recall, and F1 of 100% while both single microphone scenarios achieve accuracy, precision, recall, and F1

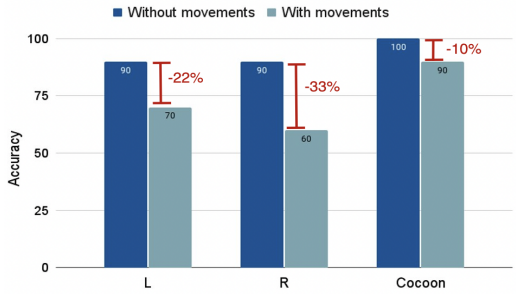

Figure 6: Decrease in accuracy with slight movements and orientation changes; L and R represent left and right, respectively.

Methods	Audible Signal	Inaudible signal
Cocoon	100%	80%
L (Left)	90%	65%
R (Right)	90%	70%

Table 2: Accuracy with audible and inaudible signals

of 90%, 93%, 93%, and 92% respectively. It can be clearly seen that *Cocoon* achieves the higher accuracy of region detection by collaborating two microphones, compared to the single microphone-based approach. This validates our hypothesis that wearable collaboration leads to better spatial profiling by having increased dimensionality and enriched spatial information. More specifically, *Cocoon* achieves a 10% and 21.5% points increase in detection accuracy on average in both controlled and real-world setups, respectively. Interestingly, we observe that performance improvement is more significant in the real-world setup. The single microphone-based approach shows satisfactory performance (90%) in the controlled setup, but performance degrades much in the real-world setup, i.e., 10% and 17% point decrease in the left and right microphones, respectively. However, the performance degradation of *Cocoon* is just 2%, which implies more robust spatial sensing from the collaboration of multiple microphones. We omit per-room and per-participant results as we did not find a statistically significant difference.

5.2 Collaboration Robustness

As discussed earlier, we record the data with and without slight movements and orientation changes in the controlled setup. To incur movement and orientation changes, we change the chair's position and orientation slightly. We train a model using a subset of the dataset that does not involve any changes to the microphones and we train another model using the entire dataset. Figure 6 shows the accuracy of both of these models. We observe that the accuracy decrease in *Cocoon* is less than that in single microphone scenarios. This corroborates our intuition that multiple microphone collaboration leads to robustness as slight changes are very unlikely to cause a significant difference in all the captured profiles simultaneously.

5.3 Audible and Inaudible Chirp Signal

Chirp signals, an essential requirement in our system, can be emitted from the speaker with a wide range of frequencies. We study

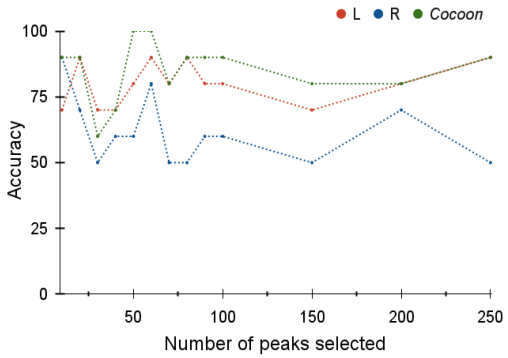

Figure 7: Accuracy variation with # of peaks selected.

the performance of *Cocoon* under the influence of audible and inaudible chirp signals using the controlled dataset. We use 6-20 kHz signal frequency and 15-20 kHz signal frequency for audible and inaudible signals, respectively. Table 2 shows the accuracy of *Cocoon* with inaudible and audible chirp signals. While *Cocoon* shows a reasonable performance with the inaudible sound (i.e., 80%), its performance is lower than that with the audible chirp signal. This is because higher frequency signals are less likely to penetrate objects. Thus, the reflections are not strong enough, and IR will be very sensitive to objects in the vicinity. On the other hand, audible chirp signals are more robust while being audible to naked ears. However, to make this system more practical, it is essential that we use an inaudible signal so that it is not hearable. This will also help in eliminating the effect of ambient noise as we can filter out the chirp signal from background noise. We leave performance improvement of the inaudible sound as future work.

5.4 Feature Analysis

As mentioned in Section 4, one important hyper-parameter that affects the detection performance is the number of prominent peaks from the IRs. The more peak we take into consideration while training, the more descriptive the acoustic footprint of a location is. We call this the peaks parameter p. We change this parameter and observe the accuracy of the models. We use the dataset recorded in the controlled setup for this experiment. Figure 7 shows the accuracy variation (y-axis) with a change in p value (x-axis).

The results show two notable implications. First, the accuracy of *Cocoon* is always higher than the single microphone approach regardless of the p value. This is because two acoustic signals can describe the location with more spatial information. Second, the accuracy peaks when $p = 50$ and then fluctuates as p increases. Initially, when p is low, accuracy is lower because less information regarding the location is available. As more number of peaks are trained, more information about the location is trained in the model. However, after the accuracy has peaked, the peaks might not be very prominent to reflect unique features, and it might get affected by interference from the ambient noise of the region. We configure the value of p to be 50 because the accuracy saturates and then decreases eventually after this value.

5.5 System Cost

In this paper, we prototyped our system using Raspberry Pi 4 boards because no commercial earable platforms provide yet access to simultaneous audio streams from two microphones and ML programmability on the device. However, to get some insights about the system cost of *Cocoon* in the deployment setup, we measure the execution latency on Raspberry Pi 4 and discuss its implication on commercial earable platforms.

The major operations for runtime inference are spatial sensing, spatial profile alignment, feature extraction & ranking, classification, and collaboration as described in Figure 3. The main bottleneck is the classification; the rest of the operations take less than 10ms in total since the input data size is quite short, i.e., 6000 samples. For the classification, the individual model size varies from 78-90 kilobytes and it takes around 80ms for each inference, i.e., around 160ms on one device. However, when we select the peaks parameter, $p = 50$, the maximum time taken for model inference is reduced to 3 ms and the model size is reduced to around 10 kilobytes, which is a drastic reduction from the previous case. Considering that today's smart earbuds are increasingly equipped with powerful processors, we envision that *Cocoon* can easily be adopted into earbuds. For example, the 2nd generation of AirPods Pro is equipped with the Apple H2 chip and Galaxy Buds Pro is equipped with Broadcom BCM43015; the detailed processing capability of these chips is not open yet, but we can expect that they are powerful enough to support many features, e.g., active noise cancellation and on-device AI using various sensors on earbuds. It is also important to note that the main operations of *Cocoon* can be easily offloaded to the smartphone if the resource of earbuds is not sufficient.

6 DISCUSSIONS

In this paper, we have explored the feasibility of our Cocoon system from multiple perspectives. However, there are a few limitations of Cocoon which we plan to investigate in our future work.

Chirp sound: *Cocoon* relies on chirp sounds for detecting a pre-registered region of interest. It requires a speaker in the vicinity to emit either an audible or inaudible sound, which is not a practical assumption. In addition, while audible chirp signal supports the system with better accuracy performance, it could be annoying to the user as discussed in Section 5.3. On the other hand, using inaudible chirp sound lowers the accuracy of detecting pre-registered regions. We plan to eliminate this requirement by performing additional experiments using ambient sound sources present in a user's environment like vent noise or refrigerator noise to capture spatial profiles by multiple microphones.

Environmental factors: As our *Cocoon* system relies on audio signals for spatial awareness, performance is impacted by environmental factors such as surrounding obstacles and ambient noise. For the proposed application of *Cocoon*, the system should be robust to these factors to eliminate the need for training for every single situation and setup. To ensure that the proposed system is robust to environmental factors, we plan to opportunistically sense spatial profiles whenever the user is inside a pre-registered region. To detect the presence of the user in a *Cocoon*, we can utilize other sensing modalities such as IMU for human activity detection and understand if the user walked away. Then, the system can group these opportunistically captured spatial profiles with the previously trained profiles. This can be achieved with incremental learning on the deployed model during run-time. Additionally, it might be useful to integrate the user's expected position based on past localization attempts into the collaboration to enhance the robustness to movements. We also plan to investigate the frequency spectrum (transfer function) of the captured spatial profiles to gather more location-specific information. This frequency spectrum will tell us about which frequencies are more absorbed than the others. Depending on the environment, *Cocoon* can decide which frequency to be selected to enhance accuracy performance.

Number and orientation of wearable devices: In this work, we use two earables' microphones for *Cocoon* profiling. Our collaborative microphone system can be extended to microphones present on different on-body devices such as smartwatch and smart glasses. Based on a few studies [26], adding more microphones to the system can increase accuracy and efficiency. However, at the same time, including additional microphones from wearable devices poses additional challenges such as inconsistent orientation and distance between microphones and hardware heterogeneity of the microphones on each wearable which all contribute to performance degradation. In our setup, we experiment with two fixed distanced microphones which provides us with significantly different acoustic information. When adding a new device such as smartwatch to the existing orientation, the mobility of the smartwatch causes inconsistent orientation and distance across the microphones on wearables in our *Cocoon* system. This inconsistency may create issues such as variable time differences between IRs leading to false inference results of pre-registered regions. To alleviate this issue, the continuous distant measurement between wearables using acoustic/BLE ranging techniques and training on different distances between wearables may be required. For the proposed system, the microphones should not be too close or too far. This is because if they are too close, they will not capture significantly different information about the environment. When they are too far, they cannot be used to localise a small space. Wearables' microphones are at a significant distance apart from each other and yet constrained on the human body, making on-body microphones a good choice for localizing a user in a region. With multiple present devices, we also have to address the heterogeneity in the microphone hardware. In such cases, the spectrum frequency which the sensor can capture may be different [11, 12]. For instance, microphone on constrained smartwatch may capture less information compared to the one on the smartphone. We aim to explore this situation with microphones of different quality as commodity microphones are prone to non-linearity that can significantly disrupt the recorded audio.

7 CONCLUSION

We present *Cocoon*, a collaborative acoustic sensing mechanism using on-body wearables equipped with microphones. Particularly, it is used to localise users by detecting their spatial profile. *Cocoon* captures the impulse response associated with audio channels between a speaker that emits chirp signals and each microphone that capture these signals. It leverages intrinsic separation between the microphones to enrich spatial information retrieved from the audio signals.

REFERENCES

[1] Takashi Amesaka, Hiroki Watanabe, Masanori Sugimoto, and Buntarou Shizuki. 2022. Gesture Recognition Method Using Acoustic Sensing on Usual Garment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–27.

[2] Jacob Benesty, Jingdong Chen, and Yiteng Huang. 2008. *Microphone array signal processing*. Vol. 1. Springer Science & Business Media.

[3] Michael Brandstein and Darren Ward. 2001. *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media.

[4] Jagmohan Chauhan, Yining Hu, Suranga Seneviratne, Archan Misra, Aruna Seneviratne, and Youngki Lee. 2017. BreathPrint: Breathing acoustics-based user authentication. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 278–291.

[5] Bhawana Chhaglani, Camellia Zakaria, Adam Lechowicz, Jeremy Gummeson, and Prashant Shenoy. 2022. FlowSense: Monitoring Airflow in Building Ventilation Systems Using Audio Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–26.

[6] Yanbin Gong, Qian Zhang, Bobby HP NG, and Wei Li. 2022. BreathMentor: Acoustic-based Diaphragmatic Breathing Monitor System. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–28.

[7] Wenqiang Jin, Srinivasan Murali, Youngtak Cho, Huadi Zhu, Tianhao Li, Rachael Thompson Panik, Anika Rimu, Shuchisnigdha Deb, Kari Watkins, Xu Yuan, et al. 2021. CycleGuard: A Smartphone-based Assistive Tool for Cyclist Safety Using Acoustic Ranging. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–30.

[8] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Alessandro Montanari. 2018. Earables for Personal-Scale Behavior Analytics. *IEEE Pervasive Computing* 17, 3 (2018), 83–89. https://doi.org/10.1109/MPRV.2018.03367740

[9] Gierad Laput, Xiang'Anthony' Chen, and Chris Harrison. 2016. Sweepsense: Ad hoc configuration sensing using reflected swept-frequency ultrasonics. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 332–335.

[10] Hong Lu, Wei Pan, Nicholas D Lane, Tanzeem Choudhury, and Andrew T Campbell. 2009. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*. 165–178.

[11] Akhil Mathur, Anton Isopoussu, Fahim Kawsar, Nadia Berthouze, and Nicholas D. Lane. 2019. Mic2Mic: Using Cycle-Consistent Generative Adversarial Networks to Overcome Microphone Variability in Speech Systems. In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks* (Montreal, Quebec, Canada) *(IPSN '19)*. Association for Computing Machinery, New York, NY, USA, 169–180. https://doi.org/10.1145/3302506.3310398

[12] Chulhong Min, Akhil Mathur, Alessandro Montanari, and Fahim Kawsar. 2022. SensiX: A System for Best-effort Inference of Machine Learning Models in Multi-device Environments. *IEEE Transactions on Mobile Computing* (2022).

[13] Chulhong Min, Alessandro Montanari, Akhil Mathur, and Fahim Kawsar. 2019. A Closer Look at Quality-Aware Runtime Assessment of Sensing Models in Multi-Device Environments. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems* (New York, New York) *(SenSys '19)*. Association for Computing Machinery, New York, NY, USA, 271–284. https://doi.org/10.1145/3356250.3360043

[14] Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. 2022. SAMoSA: Sensing Activities with Motion and Subsampled Audio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–19.

[15] 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.

[16] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. Fingerio: Using active sonar for fine-grained finger tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1515–1525.

[17] Swadhin Pradhan, Ghufran Baig, Wenguang Mao, Lili Qiu, Guohai Chen, and Bo Yang. 2018. Smartphone-based acoustic indoor space mapping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–26.

[18] Mirco Rossi, Julia Seiter, Oliver Amft, Seraina Buchmeier, and Gerhard Tröster. 2013. RoomSense: an indoor positioning system for smartphones using active sound probing. In *Proceedings of the 4th Augmented Human International Conference*. 89–95.

[19] Stephen P Tarzia, Peter A Dinda, Robert P Dick, and Gokhan Memik. 2011. Indoor localization without infrastructure using the acoustic background spectrum. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*. 155–168.

[20] Yu-Chih Tung and Kang G Shin. 2015. EchoTag: Accurate infrastructure-free indoor location tagging with smartphones. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. 525–536.

[21] Haoran Wan, Lei Wang, Ting Zhao, Ke Sun, Shuyu Shi, Haipeng Dai, Guihai Chen, Haodong Liu, and Wei Wang. 2022. VECTOR: Velocity Based Temperature-field Monitoring with Distributed Acoustic Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–28.

[22] Junjue Wang, Kaichen Zhao, Xinyu Zhang, and Chunyi Peng. 2014. Ubiquitous keyboard for small mobile devices: harnessing multipath fading for fine-grained keystroke localization. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. 14–27.

[23] Weiguo Wang, Jinming Li, Yuan He, and Yunhao Liu. 2020. Symphony: localizing multiple acoustic sources with a single microphone array. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 82–94.

[24] Zi Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. Toothsonic: Earable authentication via acoustic toothprint. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–24.

[25] Xuhai Xu, Ebrahim Nemati, Korosh Vatanparvar, Viswam Nathan, Tousif Ahmed, Md Mahbubur Rahman, Daniel McCaffrey, Jilong Kuang, and Jun Alex Gao. 2021. Listen2cough: Leveraging end-to-end deep learning cough detection model to enhance lung health assessment using passively sensed audio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–22.

[26] Takuya Yoshioka, Dimitrios Dimitriadis, Andreas Stolcke, William Hinthorn, Zhuo Chen, Michael Zeng, and Xuedong Huang. 2019. Meeting Transcription Using Asynchronous Distant Microphones. In *Proc. Interspeech 2019*. 2968–2972. https://doi.org/10.21437/Interspeech.2019-3088