Title

Cyber-Physical Augmentation for Robust and Scalable Occupant Monitoring

Author

Zhang, Yue

Publication Date

2024

UNIVERSITY OF CALIFORNIA, MERCED

Cyber-Physical Augmentation for Robust and Scalable Occupant Monitoring

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering and Computer Science

by

Yue Zhang

Committee in charge:

Professor Shijia Pan, Chair

Professor Shawn Newsam

Professor Wan Du

Summer 2024

The dissertation of Yue Zhang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Professor Shawn Newsam

Professor Wan Du

Professor Shijia Pan, Chair

University of California, Merced

Summer 2024

DEDICATION

I dedicate this dissertation to myself, my friends and families, without whose support and encouragement I could not have completed this work. I am deeply appreciative of your kindness and generosity in helping me through difficult times.

VITA

2012 - 2016	B. Eng. in Electronic Engineering, Tsinghua University
2016 - 2019	M. S. in Electronic Engineering, Tsinghua University
2019 - Now	Ph. D. in Electrical Engineering and Computer Science, University of California, Merced

SELECTED PUBLICATIONS

Dong Yoon Lee*, Yihong Li*, **Yue Zhang***, Hao-Chuan Wang, Alyssa Mae Weakley, and Shijia Pan. Kap: Kinetic augmented pill bottle for vibration-based medication interaction recognition. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*, pages 27–32, 2024.

Yue Zhang, Zhizhang Hu, Uri Berger, and Shijia Pan. Cma: Cross-modal association between wearable and structural vibration signal segments for indoor occupant sensing. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*, pages 96–109, 2023.

Yue Zhang, Carlos Ruiz, Shubham Rohal, and Shijia Pan. Cpa: Cyber-physical augmentation for vibration sensing in autonomous retails. In *Proceedings of the 24th International Workshop on Mobile Computing Systems and Applications*, pages 8–14, 2023.

Yue Zhang, Zhizhang Hu, Susu Xu, and Shijia Pan. Autoqual: task-oriented structural vibration sensing quality assessment leveraging co-located mobile sensing context. *CCF Transactions on Pervasive Computing and Interaction*, 3:378–396, 2021.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

As I reflect on this journey, I am filled with profound gratitude for the numerous individuals who have contributed to the completion of this thesis.

First and foremost, I would like to acknowledge my own persistence and courage. This 5-year PhD journey has been far more challenging and demanding than I ever anticipated.

I am deeply indebted to my esteemed advisor, Professor Shijia Pan, whose invaluable guidance, encouragement, and insights have been instrumental throughout this process. I am also profoundly grateful to my dissertation committee members, Professor Shawn Newsam and Professor Wan Du. Their insightful comments, thought-provoking questions, and valuable suggestions have significantly enriched the quality and depth of this research.

My heartfelt appreciation extends to my mentors and advisors at Futurewei Technologies Inc. The internship experiences in the summers of 2023 and 2024 were transformative, allowing me to regain confidence and rediscover my passion. It was a pleasure to work with Dr. Xiyun Song, Dr. Heather Yu, Dr. Zongfang Lin, and Dr. Liang Peng. I also want to express my gratitude to my fellow interns at Futurewei: Seokmin Choi (University at Buffalo, SUNY), Achleshwar Luthra, and Jingyu Shi (Purdue University). Together, we shared a fulfilling and enriching summer.

I am grateful to my friends at the University of California Merced and collaborators who have provided a supportive and collaborative environment. The moral support from Zhizhang Hu, Shangjie Du, Yuning Chen, Yuan Feng, Dr. Songtao Ye, Dr. Miaomiao Liu, Dr. Yuxin Tian, and Jiqing Wen (Arizona State University) was crucial in helping me persevere through challenging times.

A special thanks goes to Yosemite National Park. Its breathtaking views and challenging trails offered much-needed respite and rejuvenation during the rigorous days of my studies. I cannot imagine how much more difficult this journey would have been without the opportunity to hike and camp in Yosemite.

Finally, I owe an immeasurable debt of gratitude to my family for their unconditional love, encouragement, and unwavering belief in me. Their steadfast support and understanding have been the foundation of my success. I am eternally grateful to my parents and siblings for their constant inspiration and motivation throughout this arduous journey.

ABSTRACT OF THE DISSERTATION

Cyber-Physical Augmentation for Robust and Scalable Occupant Monitoring

by

Yue Zhang
Doctor of Philosophy in Electrical Engineering and Computer Science

University of California Merced, Summer 2024

Professor Shijia Pan, Chair

With the development of information technology and Internet-of-Things (IoT), various smart home applications have been proposed to improve the safety and comfortableness of people's lives. Occupant monitoring is essential to enable these smart home applications, and multiple systems have been explored, including vision-, audio-, radio- and wearable-based methods. However, these methods face certain limitations and requirements in real-world environment such as line-of-sight constraints for vision-based methods, the need for silent environment for audio-based methods, sensitive to multi-user movement for radio-based methods, and considerations of comfort and battery life for wearable-based methods. There is a pressing need for a robust and scalable solution for occupant monitoring in complex and diverse real-world deployments.

In this dissertation, we introduce a cyber-physical augmented structural vibration sensing system for occupant monitoring. Compared to other methods, the advantages of structural vibration sensing are non-intrusive (device free), enabling sparse deployment and decreasing privacy concerns. However, these advantages are accompanied by limitations that restrict the system's robustness and scalability in real-world deployment.

To improve the robustness and scalability, we approach this problem from two aspects: **cyber** and **physical augmentation** to enhance robust data acquisition and scalable information inference. First, we leverage the individual information from wearable sensing to **cyber**-augment the scalability of information inference. We introduce *CMA*, a cross-modal signal segment association scheme that associates the identity information from wearable sensing to enable individual occupant monitoring for structural vibration sensing without label data. Second, we utilize activity information from wearable sensing to **cyber**-augment the robust (high-quality) data acquisition. We introduce *AutoQual*, an autonomous sensing quality assessment framework to quantify the impacts of deployment environment on the sensing task performance. The wearable sensing enables the system to automatically select the occupant induced signal for quality assessment without requiring additional human effort. Third, we combine **cyber** and **physical** augmentation to enhance scalable information inference and robust data acquisition. We present *CPA*, a cyber-physical augmentation scheme to enhance the vibration sensing signal via a physical arc structure and achieve high-accuracy, labeling-free event detection. Our experiments verified the efficiency of *CMA*, *AutoQual*, and *CPA* in real-world experiments. Fourth, we explore **cyber** and **physical** augmentation to enhance the capability of single-point sensing for occupant tracking. We introduce *LEVO*, which utilizes the metamaterial filter made of LEGO® bricks to embed the direction information into the waveform. The preliminary experiment shows the feasibility of *LEVO* for occupant tracking by signal sensor.

Chapter 1

Introduction

Occupant monitoring is essential to enable various smart home applications such as energy management [83, 91, 137], smart healthcare [138, 138], and elderly care [6, 15, 70, 87]. Researchers have explored various methods to sense the occupant information, including vision-based [28, 29], audio-based [62, 71], radio-based [8, 141], and wearable-based [92, 106, 106] methods. Generally, these systems fall into two categories: device-based and device-free. Device-based methods require occupants to carry special devices to monitor their information, which are usually powered by batteries. However, considering comfortableness and battery life, long-term and continuous monitoring is critical for device-based methods. On the other hand, device-free approaches, such as visual- and acoustic-based methods, often have certain requirements (e.g. line-of-sight, and silent environment). The radio-based methods are sensitive to multi-user movement and high equipment costs. The robustness and scalability of the device-free methods in complex deployment environments are critical. We summarize the **research questions** as: how to achieve a robust and scalable occupant monitoring for end users without additional resource requirements?

In this dissertation, we introduce a set of cyber-physical augmentations to enhance the structural vibration sensing system for robust and scalable occupant monitoring. Compared to other sensing systems, the structural vibration sensing system offers several advantages: 1) non-intrusive, 2) sparse deployment, and 3) less privacy concern. The structural vibration-based sensing system detects the occupant activity (e.g., footstep, pick-up and put-down things) induced vibrations to infer their information, which belongs to a device-free sensing method. Due to the long propagation distance of vibration signals in solids, each sensor can cover a large area, enabling sparse deployment. The sensor captured vibration data is an indirect measurement (structure vibration) of the occupant activity, which makes it challenging to directly infer occupant information for humans. Compared to the direct sensing system (e.g., vision, audio), vibration sensing raises less privacy concerns. However, these advantages come with challenges that limit the robustness and scalability in real-world deployment. First, multiple factors in the physical environment impact the quality of sensor acquired data. These impacts are more severe for indirect and sparse sensing system, which cause the system performance (robustness) varies. Second, collecting label data is time consuming and labor intensive, especially for an indirect sensing signal which is difficult to recognize labels from raw data. The requirement of label data limits the scalability of information inference in various environments. To address these challenges, we approach this problem from two perspectives: cyber and physical augmentation to enhance the robustness of data acquisition and scalability of information inference.

We leverage the complementary information from other modalities to decrease the requirement of human effort to enhance the scalability (**cyber** augmentation), and propose a set of physical structure to enhance the high-quality data acquisition (**physical** augmentation). In summary, to improve the robustness and scalability of the vibration-based occupant monitoring system, 1) we propose a cyber augmentation solution to enhance the scalability of information inference without labeled data, 2) we propose a cyber augmentation solution to enhance the robustness of data acquisition in complicated environments, 3) we propose a cyber-physical augmentation solution to enhance the robustness of data acquisition and scalability of information inference under sparse deployment.

In Chapter 2, we focus on leveraging the identity information from wearable sensing to enhance the identification capabilities of structural vibration sensing system. In the structural vibration-based sensing system, the occupant identity information is missing when multiple users are co-active. The traditional solution is collecting labeled data for each occupant and build an occupant identification model to identify each occupant, which is labor intensive and time consuming. We notice that the wearable sensing data contains both identity and activity information. We can leveraging the co-located wearable sensing data of each occupant to 'label' the identity of the vibration signals without requiring any additional human effort. We propose *CMA*, a cross-modal signal segment association scheme between wearable and structural vibration sensors. To determine whether two signal segments from different modalities over the same period are associated, we calculate an *association probability* (AP). The intuitions to calculate this association probability are twofold: 1) as long as the sensors are capturing the same physical activity, there will be an implicit shared context between two signal segments, and 2) we assume that for the structural vibration signals that are segmented as one activity (e.g., five seconds), there will be only one wearable sensor associated to it. We evaluate *CMA* via a public multimodal dataset for systematic evaluation, and we collect a continuous uncontrolled dataset for robustness evaluation. *CMA* achieves up to 37% AUC value, 53% F1 score, and 43% accuracy improvement compared to baselines. The material in this Chapter appears in the 22nd International Conference on Information Processing in Sensor Networks. The authors are Yue Zhang, Zhizhang Hu, Uri Berger, and Shijia Pan. © 2023 Copyright held by the authors.

In Chapter 3, we look into utilizing the activity information from wearable sensing to enhance the high-quality data acquisition. The acquired data quality varies in different environments and locations, which leads to the system performance variation. The system's information inference performance (e.g., detection rate, learning accuracy) is impacted by the deployment environment. Quantifying sensing quality allows further enhancement of deployment efficiency to improve IoT sensing systems' performance. Compared to prior work on signal quality assessment, which is mainly used in the domains of communication [9, 16, 64, 119] and computer vision [75, 125, 132]. These methods do not reflect effects of the deployment environment on the sensing task performance, and do not quantify the environmental impacts to acquired data characteristics. We present *AutoQual*, an autonomous sensing quality assessment framework to quantify the impacts of the deployment environment on IoT sensing system performance. To systematically understand these deployment environment impacts, we define **sensing quality** as a series of measurable factors/models reflecting how they impact a given information inference task. To autonomous quantify sensing quality, we 1) utilize domain knowledge on wave propagation and structural properties to identify a set of assessment factors (AF) and design the measuring method accordingly, 2) adopt a data-driven approach to estimate the relationship between measured AFs and sensing tasks' performance, 3) automate AF measurements using human-induced vibration signals extracted by co-located mobile devices. We apply *AutoQual* for an autonomous sensing quality assessment on multiple sensing tasks, including event detection and occupant identification. The real-world experiment shows *AutoQual* achieves less than 0.10 absolute error, which is an 2× improvement to baselines. The material in this Chapter appears in the CCF Transactions on Pervasive Computing and Interaction. The authors are Yue Zhang, Zhizhang Hu, Susu Xu, and Shijia Pan. © 2021 Copyright held by the authors.

In Chapter 4, we focus on fusing cyber and physical augmentation to enhance high-quality data acquisition and information inference. In the structural vibration-based sensing system, the system relies on the detected event signal to infer occupant information, such as identifying the occupant through footstep-induced vibration events. The performance of event detection is important for further information inference. However, the vibration wave attenuates as it propagates in structures, and causes the sensor acquired data to have a low Signal-to-Noise Ratio (SNR) after a certain distance. The traditional threshold-based event detection algorithm has a bed performance for the low Signal-to-Noise Ratio (SNR) events detection. We introduce *CPA*, a cyber-physical augmentation scheme to enhance the vibration sensing signal via a physical arc structure and achieve high accuracy labeling-free event detection. We evaluate *CPA* in a real-world scenario by deploying a vibration sensor on a retail gondola to monitor interactions across multiple shelves. Our approach demonstrates up to 2.9× improvement for light-weight product pickup event detection. The material in this Chapter appears in the 24th International Workshop on Mobile Computing Systems and Applications. The authors are Yue Zhang, Carlos Ruiz, Shubham Rohal, and Shijia Pan. © 2023 Copyright held by the authors.

In Chapter 5, we briefly introduce a cyber and physical augmentation to enhance occupant tracking from single-point sensing (*LEVO*), which is a potential research direction we plan to explore in the future. *LEVO* utilizes the metamaterial filter made of LEGO® bricks to manipulate the frequency components of the mechanical wave, which embeds the direction information into the waveform. The material in this Chapter appears in the 21th ACM Conference on Embedded Networked Sensor Systems. The authors are Yue Zhang, Shikha Patel, Dong Yoon Lee, Paolo Celli, Amelie Bonde, and Shijia Pan. © 2023 Copyright held by the authors.

Chapter 2

Cyber Augmentation for Scalable Information Inference

2.1 Introduction

Indoor occupant sensing enables many smart home applications, such as elderly care, building management, and personalized service. Various sensing modalities have been explored, and these systems fall into two categories based on whether they require the occupant to carry extra devices: on-body and off-body sensing. Fusing on-body and off-body sensing is prevalent in indoor occupant sensing, given multimodal signals can provide complementary information for the same target, and therefore achieve robust information inference [23, 31, 37, 60, 113]. Among these combinations, wearable and structural vibration sensing have demonstrated efficient complementarity for various inference tasks [51, 60]. However, when the size of these IoT systems increases, they may sense multiple physical activities occurring at the same time. For example, for an IoT system deployed over different areas in a house, they may sense people doing different activities in different areas. It also means that for any pair of cross-modal sensors, the physical activity they are sensing may or may not be the same. If signal segments of two sensing modalities that capture different activities are used for inference, a spurious complementary relationship will be used. Therefore, it is of great importance to establish correct association relationships for signal segments from co-located sensors of different modalities.

This cross-modal association relationship is beneficial for multiple use cases: *1) User signal segment annotation.* When wearable and structural vibration sensors are used together, with this signal segment level association, the wearable sensors can be used as the identity annotation tool for the structural vibration sensors' signal segments, since the wearable is by nature associated with their user already. This could further advance the structural vibration sensing-based IoT system's usability and scalability as a zero-effort bootstrapping user annotation scheme. *2) Enhancing multimodal learning efficiency.* With a high-accuracy signal segment association, multimodal learning would be able to leverage this prior knowledge to achieve more accurate modeling, since falsely associated signal pairs may result in the spurious complementary relationship being modeled.

Therefore, we formulate this *cross-modal signal segment association* problem between wearable and structural vibration sensors [148, 149, 151] as illustrated in Figure 2.1. Given a set of segments from two co-located sensing modalities collected during the same period (e.g., Seg1-1,...Seg3-1), our goal is to learn a segment-level association cross modalities (e.g., Seg1-1: P1-I1, Seg2-1: P2-I1). However, this cross-modal signal segments association has the following **challenges**: *1) Indirect sensing leads to the lack of direct comparable information.* For indirect sensing systems of structural vibration and IMU, their raw measurements are often not directly interpretable, and therefore, can not be easily compared for shared context (signal examples in Figure 2.5 later). *2) Complementary leads to disassociation.* IoT systems that adopt multiple modalities often leverage their complementarity to achieve more efficient modeling. On the other hand, the more complementary the two modalities are, the less shared information they capture, and hence their signal segments are

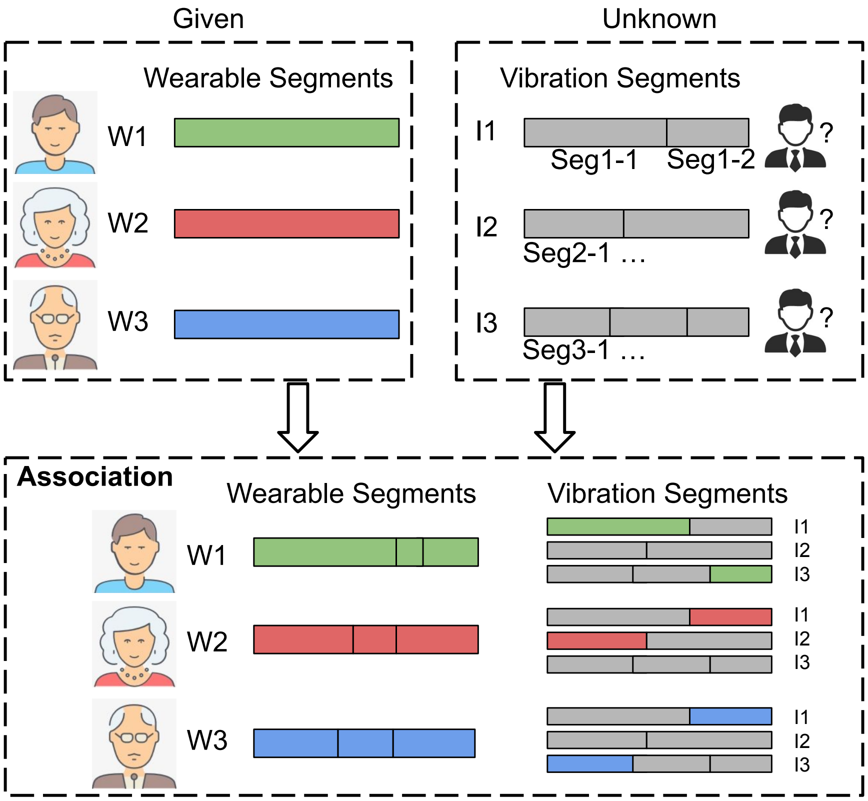


Figure 2.1: Cross-modal signal segment association problem: given a set of segments from two sensing modalities collected during the same period, we aim to identify segment pairs that are associated. We refer to the segment pair that contains signals induced by the same physical activity as 'associated'.

more difficult to be associated with. For example, prior work that conducts location association between the electric load sensor and microphone [51] requires longer measurements than that of the camera and IMU [114], because the latter leverages a clear shared context of acceleration. *3) Mobility variance leads to spatiotemporal variation.* For modalities with different levels of mobility, this association may vary over time. For example, occupants who carry an on-body sensor may move in the house and are captured by different off-body sensors. Therefore, this association relationship varies over time due to occupants' mobility. We form our **research question** as How do we learn the segmentation-level association relationship between wearable and structural vibration sensors with constrained shared context and without labeled data?

We present *CMA*, a cross-modal signal segment association scheme between wearable and structural vibration sensors. To determine whether two signal segments from different modalities over the same period are associated, we calculate an *association probability* (AP). The intuitions to calculate this association probability are twofold: 1) as long as the sensors are capturing the same physical activity, there will be an implicit shared context between two signal segments, and 2) we assume that for the structural vibration signals that are segmented as one activity (e.g., five seconds), there will be only one wearable sensor associated to it. The temporal convolutional network (TCN) has shown efficient learning ability for the temporal representation features from time series signals [74]. We propose *AD-TCN*, a framework built upon TCN to calculate the amount of shared context between signal segments from different modalities. First, *AD-TCN* takes all candidate wearable segments and the vibration segment history values to predict the vibration segment's current time step value. Then we train the model and calculate the association probability between signal segments from two modalities based on the weights of the trained *AD-TCN*. The association probability reflects the contribution of one signal segment for predicting the other. If the contribution of a signal segment is higher than a threshold, we consider this wearable signal segment is associated with the vibration signal segment, i.e., they detect the same physical activity. In summary, the contributions of this work are as follows:

- We introduce *CMA*, a cross-modal sensing signals' segment-level association scheme for multimodal IoT systems.
- We present *AD-TCN* that learns the segment-level cross-model representation

and uses the learned model parameters to calculate the amount of shared context between modalities.

- We evaluate *CMA* through both a public dataset and an uncontrolled real-world dataset for robustness analysis.

The rest of the Chapter is organized as follows. First, we investigate and compare *CMA* to related work in Section 2.2. Then, we illustrate the details of *CMA* in Section 2.3. Next, we demonstrate the experiments and analysis in Section 2.4 and Section 2.5. Finally, we discuss future directions in Section 2.6 and summarize the Chapter in Section 2.7.

2.2 Related Work

We investigate prior research on device pairing/identification, and occupant identification, and discuss the research gap that we focused on in this work.

2.2.1 Cross-modal IoT Device Identification

Cross-modal IoT device pairing/identification is a relevant topic to cross-modal signal segment association. Prior work on cross-modal pairing relies on the shared context that can be sensed by both sensing modalities and compare the similarity of the acquired shared context to achieve the paring or identification. Ruiz et. al. leverages the shared 3D motion (spatial context) of human body parts captured by both camera and IMU sensor to achieve IoT device identification [98,114]. Han et. al. utilize the shared context of activity start/end time (temporal context) to generate fingerprints for co-located device pairing [51]. However, these prior works do not directly apply to our target scenario due to the challenges from *constrained shared context*. *CMA* solves these challenges by using the temporal convolutional network to

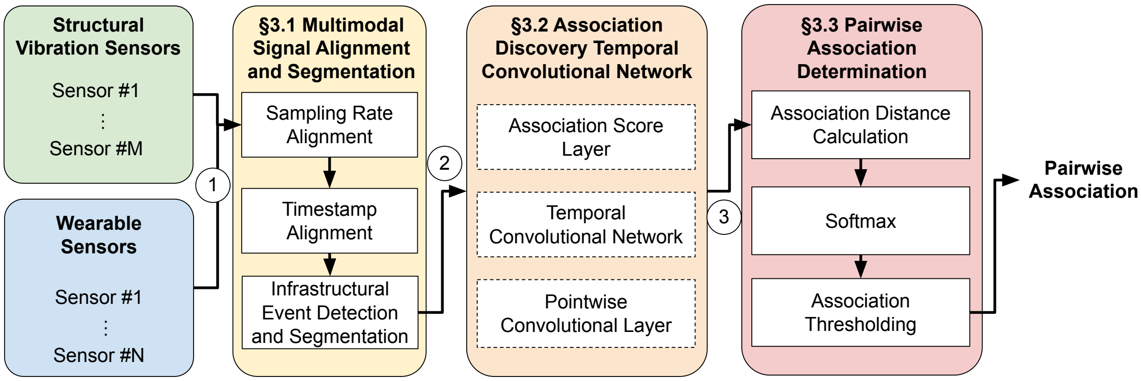

Structural Vibration Sensors
Sensor #1
⋮
Sensor #M
Wearable Sensors
Sensor #1
⋮
Sensor #N
1
§3.1 Multimodal Signal Alignment and Segmentation
Sampling Rate Alignment
Timestamp Alignment
Infrastructural Event Detection and Segmentation
2
§3.2 Association Discovery Temporal Convolutional Network
Association Score Layer
Temporal Convolutional Network
Pointwise Convolutional Layer
3
§3.3 Pairwise Association Determination
Association Distance Calculation
Softmax
Association Thresholding
Pairwise Association

Figure 2.2: System overview. *CMA* consists of three modules to estimate the association relationship between the structural vibration and wearable sensors: 1) multimodal signal alignment, 2) Association Discovery Temporal Convolutional Network (*AD-TCN*), and 3) association probability estimation.

efficiently discover the limited association information without an explicitly shared context.

2.2.2 IoT for Occupant Identification

The fundamental problem solved by this work is to associate the infrastructure sensor signals with the person who induces it, which is also relevant to the sensor signal-based identification problem. Prior work on occupant identification has explored the possibility to identify the person based on how their behavior or interaction with the environment varies [52]. A more specific description of human behavior is the walking pattern or gait, which can be observed by a wide range of sensors [100, 130, 146]. Other biometrics are also explored to enable ubiquitous occupant identification in the smart home setting such as voice [80], human body's reflection, refraction, diffraction, and even absorption of radio signals [143]. However, all the identification systems require the occupant identity label to create the corresponding classifier model to achieve the identification. In our scenario, it is difficult and impractical to assume the availability of labeled data for each deployment. Instead, we leverage the wearable sensor and their nature association with individuals who wear them to 'label' the identity of the infrastructure sensing segment as a signal association problem.

2.3 *CMA* Design

We present *CMA*, a cross-modal signal segment association scheme. Figure 2.2 describes *CMA*'s architecture, which consists of three modules. First, *CMA* aligns signals ① from all sensors by aligning their timestamp and sampling rate, so that these signals are comparable temporally (Section 2.3.1). Then a threshold-based event detection algorithm is applied to detect the valid events from the structural vibration data, and the timestamp of the structural vibration events are utilized to segment the wearable IMU data. The segmented multimodal events ② are then sent to the Association Discovery Temporal Convolutional Network (*AD-TCN*), where for each structural vibration sensor, an *AD-TCN* is trained and the weight values of the association score layer ③ are output (Section 2.3.2). Finally, *CMA* calculates the pairwise **association probability (AP)** between each structural vibration sensor and each wearable (Section 2.3.3). We consider the pair of the wearable and the structural vibration sensor with the association probability higher than a threshold is associated (i.e., they detect the same occupant).

2.3.1 Multimodal Signal Alignment and Segmentation

Due to the heterogeneity of the two sensing modalities, *CMA* first preprocesses the incoming signals by aligning and segmenting the signal of interest. Since different types of sensors are sampled at different rates, the number of samples in the same event duration may vary. Furthermore, since we utilize TCN architecture for association discovery (Section 2.3.2), the architecture takes the same length of time series data points as input and outputs. Therefore, it is important to ensure that all the sensor inputs have the same number of samples in each second, and samples over all the sensor inputs are temporally aligned (Section 2.3.1). In addition, since in our application scenarios, the wearable sensors are directly associated with the user identities and the structural vibration sensor signals need to be associated with the user identities, *CMA* only conducts association when there is vibration signal detected (Section 2.3.1).

Sampling Rate and Timestamp Alignment.

To ensure accurate multimodal temporal information modeling, we first align the sampling rate over all the sensor inputs. We select the lowest sampling rate Q (reference) of all available sensors as the reference. Then we conduct resampling [123] on each of the other sensor inputs. Using a signal with an original sampling rate of P Hz as an example ($P \geq Q$, and P, $Q \in \mathbb{N}^{+}$). To resample the signal, first, the least common multiple (LCM) of P and Q is calculated. Then the linear interpolation is conducted to up-sampling the P Hz sampling rate data to LCM Hz. Next, a low-pass filter is applied to remove the higher frequency ($> P$) components in the up-sampling series. Finally, the up-sampling series is down-sampled to Q Hz [81].

Since the TCN leverages the temporal relationship between historical samples and current samples to establish models, it is important to have samples from all sensors time-aligned. Therefore, based on the periodically provided timestamp, *CMA* interpolates the timestamp for each sample for high-resolution alignment.

Structural Vibration Event Detection and Activity Segmentation

To detect the event of interest to conduct the temporal association on, we further conduct a threshold-based event detection algorithm on the vibration data. We first apply a sliding window on the time sequence data of the vibration sensor and calculate the energy of the windowed signal (Figure 2.3(b)). We characterize the ambient noise's windowed signal energy as Gaussian noise (μ_n, σ_n) [100]. Then, we select a lower bound θ_e as the energy threshold of the windowed signal. If the energy

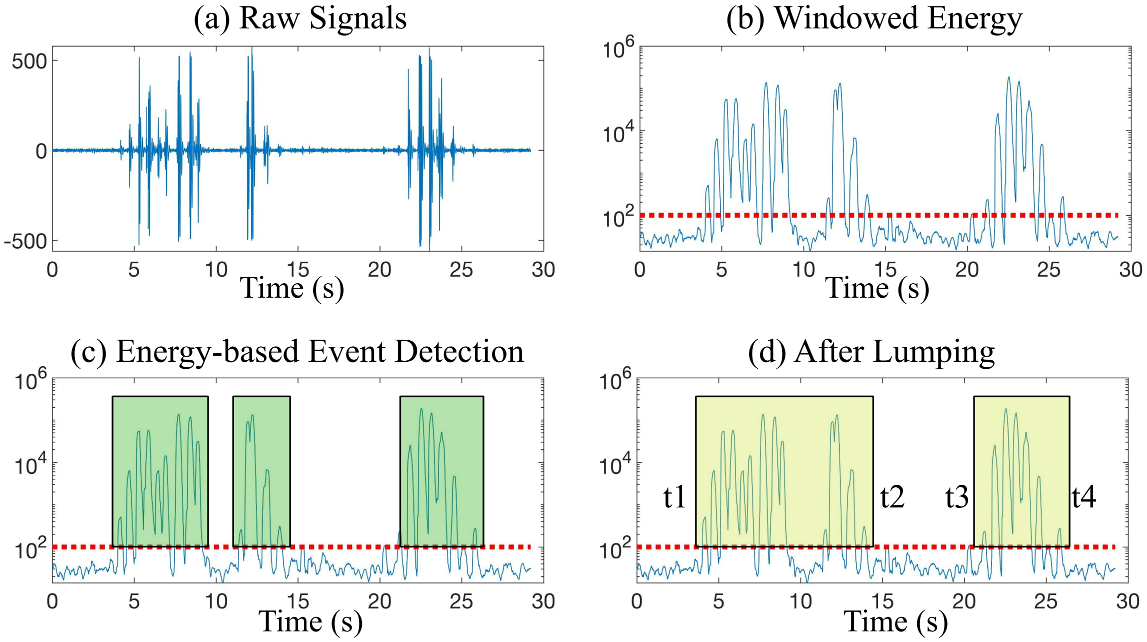


Figure 2.3: Structural vibration event detection and activity segmentation signal examples. (a) depicts examples of raw signals of human-induced structural vibration. (b) shows the signal energy of the sliding window applied to the signal in (a). (c) we conduct an energy-based event detection on the windowed signal energy, where the detected events are marked by green boxes. (d) finally, events with intervals lower than a pre-selected threshold are lumped [51] as one activity, which is marked by the yellow box. For example, the segment from t1 to t2 contains signals of one activity.

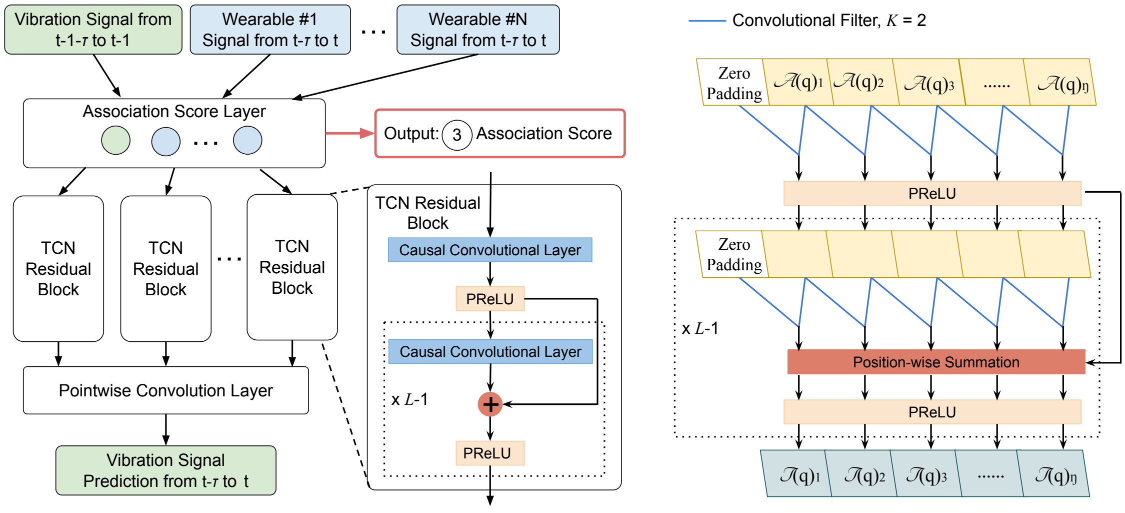
of this windowed signal is larger than $(\mu_n + \theta_e * \sigma_n)$, we consider this window is an **event** (Figure 2.3(c)).

Next, we conduct activity segmentation with an interval-based lumping method [51], where we segment the consecutive events that are less than the event interval threshold $\Delta\tau$ as one **activity segment (AS)** (Figure 2.3(d)). We segment the aligned IMU data consistently with the structural vibration sensor segments. The activity segment's start and end time is data-driven, therefore it does not have semantic meanings, and one segment maybe two people's activity occurring consecutively within $\Delta\tau$. To ensure efficient association, we further segment the activity segments into **association units (AU)** to unify the association signal with lower and upper bounds, τ_l and τ_u. We assume the association of the signals does not change within an AU. For a segmented AS, if the duration is shorter than τ_l, *CMA* discards it because there is not enough information to perform the association. On the other hand, if the duration is longer than τ_u, we will divide the AS into multiple AUs by the duration $\leq \tau_u$, and discard ones with a duration $< \tau_l$. The aligned AUs from two modalities are inputs for the next module.

2.3.2 Association Discovery Temporal Convolutional Network (*AD-TCN*)

In our cross-modal association problem, the wearable IMU measures the occupant's motion, which causes the structure to vibrate. Inspired by the prior work that utilizes the TCN architecture to infer Granger causality [90], we model the cross-modal signal association problem as a time series prediction problem and quantify the contribution of one segment (X) on the prediction of another segment (Y) as an indicator of such association relationship. In our model, for an AU of duration τ at time step t, we consider X is the raw signal of the wearable sensor between $t-\tau$ and t, and Y is the raw signal of the structural vibration sensor between $t-\tau$ and $t-1$. If X's past value at $t-\tau$ to t contributes to predict Y at t, then X and Y are associated with an association probability proportional to this contribution.

We present *AD-TCN*, an association discovery network built upon the TCN architecture to infer causal relationship [90] between pairs of multimodal sensing signals. Figure 2.4(a) shows the overview of the *AD-TCN*. The network has three parts, namely the association score layer, TCN residual block, and pointwise convolution layer. The network takes aligned AU of duration τ with $\eta = \tau \times Q$


Vibration Signal from t-1-τ to t-1
Wearable #1 Signal from t-τ to t
Wearable #N Signal from t-τ to t
Association Score Layer
Output: 3 Association Score
TCN Residual Block
TCN Residual Block
TCN Residual Block
TCN Residual Block
Causal Convolutional Layer
PReLU
Causal Convolutional Layer
x L-1
PReLU
Pointwise Convolution Layer
Vibration Signal Prediction from t-τ to t
Convolutional Filter, K = 2
Zero Padding
𝒜(q)₁
𝒜(q)₂
𝒜(q)₃
......
𝒜(q)η
PReLU
Zero Padding
x L-1
Position-wise Summation
PReLU
𝒯(q)₁
𝒯(q)₂
𝒯(q)₃
......
𝒯(q)η

(a) Network Architecture

(b) Causal Convolution

Figure 2.4: The architecture of the association discovery temporal convolutional network (*AD-TCN*). The network consists of the association score layer, the TCN residual block, and the pointwise convolution layer. The model is trained over multiple epochs, and the association score layers' node weights are the output of the *AD-TCN* model, marked as ③ in Figure 2.2.

samples from index η_0 as inputs. For wearable sensor signals, the input is signal indexing between η_0 and $\eta_0 + \eta$. For structural vibration sensor signals, the input ranges from $\eta_0 - 1$ to $\eta_0 - 1 + \eta$. The prediction output is the structural vibration sensor signal indexing from η_0 to $\eta_0 + \eta$. For each structural vibration sensor's AUs and n available wearable sensors' signal, an *AD-TCN* network is trained independently to estimate the association relationship.

Association Score Layer

We introduce a trainable association score layer to measure the weight put on each channel of sensor signals by the network. Figure 2.4(a) shows the architecture of the association score layer and its inputs and outputs. For a multimodal sensing system with M wearable sensors (the i^{th} sensor has C_i channels), the association score layer contains $h = 1 + \sum_{i=1}^{M} C_i$ nodes (shown as circles in Figure 2.4(a)), each contains a weight value. In the beginning, all nodes are initialized with the same weight value, i.e., each input equally contributes to structural vibration signal prediction. These weights are updated during model training by the gradient descent algorithm [111]. The association score is calculated from the weight via the softmax function as the layer's activation function. When the model training is finished, the final association score outputs to the Association Probability Estimation module (③ in Figure 2.2). A high association score indicates that this node's input has more contribution to predicting the structural vibration signal, and the input signal of this node is more likely associated with the structural vibration signal. On the other hand, during model training, the association scores are multiplied with their corresponding input signal as the output of the layer. For input multimodal signal segments with length $\eta = \tau \times Q$, the output of the association score layer is shown as follows:

$$\begin{aligned} &\mathcal{A}(q) = \alpha_q \cdot SE_q = \frac{\exp^{W_q}}{\sum_{j=1}^{h} \exp^{W_j}} \cdot SE_q \\ &\mathcal{A}(q) \in \mathbb{R}^{\eta \times 1}, q \in [1, h] \end{aligned} \tag{2.1}$$

Where $SE_q \in \mathbb{R}^{\eta \times 1}$ is the q th input, α_q and W_q are the association score and the weight of q th node, respectively.

Temporal Convolutional Network Residual Block

We adopt the temporal convolutional network (TCN) residual block [12] for its strong performance in time-series prediction. Transitional TCN is designed for univariate time-series prediction, i.e., predicting with one time-series data. However, *CMA* models the association problem as a time-series prediction problem with multiple time-series data inputs, i.e., multivariate time-series prediction. To adapt to the multivariate time-series prediction, we utilize a depthwise separable architecture to extend the univariate TCN architecture for multivariate prediction [30]. That is, for output from the association score layer's each node, they are separately sent to different TCN residual blocks, as shown in Figure 2.4(a). In total, there are h independent TCN residual blocks. Each block has the same architecture: L layers of 1-D causal convolutional network layers [126]. These layers have the same kernel size K. Figure 2.4(b) illustrates the mechanism of 1-D causal convolution in the TCN residual block. The 'causal' in this layer architecture name means that the prediction of time t data is generated only with data from time t and earlier. For instance, to predict $\mathcal{T}(q)_3$, only the data no later than $\mathcal{A}(q)_3$ is used. In this way, no future information is used in prediction i.e., no information leakage. Each causal convolutional layer has the same length (η) as the input time-series signal. Since only the history data can be used for prediction, in order to keep subsequent layers the same length as the first layer, a left zero-padding of size $K-1$ is added. After each causal layer, we adopt a Parametric Rectified Linear Unit (PReLU) [54] as the non-linear activation function, for its empirical strong performance on improving model fitting capability. A residual connection [55] is added before each PReLU activation result in the block, except the first one. The residual connection conducts a position-wise summation of the previous and current layers' results. This allows the block to learn modifications on the block input rather than the entire transformation, which has been shown to benefit scaling the network to very deep [12]. The set of calculations of the q th block can be described as follows:

$$\begin{aligned}
\mathcal{T}_1(q) &= PReLU(G_q^1 * \mathcal{A}(q) + b_q^1) \\
\mathcal{T}_l(q) &= PReLU(G_q^l * \mathcal{T}_{l-1}(q) + b_q^l) + \mathcal{T}_{l-1}(q) \\
\mathcal{T}_1(q), \mathcal{T}_l(q) &\in \mathbb{R}^{\eta\times 1},\ l \in [2, L]
\end{aligned} \tag{2.2}$$

Where $\mathcal{T}_1(q)$ and $\mathcal{T}_l(q)$ are the output of the first layer and l th layer, $G_q^1, G_q^l \in \mathbb{R}^{K\text{x}1}$ are weights of the convolution filters in the first layer and l th layer, and $b_q^1, b_q^l \in \mathbb{R}$

are bias terms of each layer. K is the kernel size of the convolution filter. $*$ denotes the convolution operator.

Receptive field is a term that describes how much history data is utilized in the prediction, and it has been proved the size of the receptive field has an impact on the prediction accuracy [126]. There are two hyper-parameters in the TCN residual block that jointly determine the receptive field size: L, number of causal convolutional layers; and K, kernel size of the 1-D convolution filter [12]. Additionally, we can achieve the same receptive field using a different composition of K and L, but the properties of the network may impact performance. For instance, a large L may make model training more difficult and cause overfitting [12]. The evaluation of receptive field size F and hyper-parameters setting (K and L) on the system performance is shown in Section 2.5.3.

Pointwise Convolution Layer

We apply a pointwise convolution layer to integrate the output of all h TCN residual blocks as the prediction of the structural vibration segment. The output of the pointwise convolution layer has the same length η of the input time-series signal segments. The calculation of the pointwise layer is as follows:

$$\begin{aligned} \hat{\mathcal{I}} &= \sum_{q=1}^{h} p_q \cdot \mathcal{T}_l(q) \\ \hat{\mathcal{I}} &\in \mathbb{R}^{\eta \times 1} \end{aligned} \tag{2.3}$$

Where $p_q \in \mathbb{R}$ is the weight of the pointwise convolution filter for the qth TCN block output.

Loss Function

We use the mean square error (MSE) as the loss function to measure the difference between the raw vibration sequence (I) and the predicted sequence ($\hat{\mathcal{I}}$). The calculation of MSE is as follows:

$$\mathcal{L} = \frac{\sum_{r=1}^{\eta}(I(r) - \hat{\mathcal{I}}(r))^2}{\eta} \tag{2.4}$$

Where η is the length of the AU. MSE reflects how similar the predicted sequence $\hat{\mathcal{I}}$ and the ground truth I are. The optimization goal is to minimize $\mathcal{L}$ during the model training.

2.3.3 Pairwise Association Determination

To enable explainable association, *CMA* estimates an AP for each τ seconds multimodal data based on the *AD-TCN* output. The output – association score– is the attention value [11] of the neural network for each input, and cannot represent the association relationship directly. Furthermore, since *AD-TCN* is applied to each structural vibration sensor, the weight values of different *AD-TCN* are not comparable. Therefore, a common representation of the association relationship between the structural vibration and wearable sensors is needed. To do so, we first calculate a 'divergence' between the structural vibration sensor and all the available wearable sensors using the association score. Next, we apply a softmax function to convert the association divergence to the AP between the structural vibration sensors and wearable sensors. In this way, we find a common measurement of the association relationship between multiple structural vibration sensors and wearable sensors.

The association divergence measures the association relationship between the structural vibration sensor and wearable sensor. A low association divergence value means the IMU has less contribution on the prediction of the target vibration sensor, i.e., they have a lower probability to be associated. For the wearable sensor q with C channels, *CMA* outputs C values of association score, as a vector $\mathbf{W_q}$. *CMA* integrates the C channels of the association score into a divergence D_q as the square root of Euclidean norm [121] of the vector $\mathbf{W_q}$.

$$D_q = \sqrt{\sum_{i=1}^{C} W_q(i)^2} \tag{2.5}$$

Note that this D_q alone, or the vector $\mathbf{W_q}$ alone is not comparable to each other, because the association score for each structural vibration sensor is calculated individually by a neural network. Therefore, they cannot be directly compared to a global threshold. To allow explainable and comparable outputs, we further normalize this divergence by softmax [3], and output the AP as

$$AP = \frac{\exp^{(D_q)}}{\sum_{i=1}^{N} \exp^{(D_i)}} \tag{2.6}$$

CMA reports an association if the AP value is larger than a threshold θ_{AP}.

Figure 2.5 shows an example AU of duration $\tau = 14s$ with the structural vibration segment in (a) and wearable segments in (b,c). By directly comparing Figure 2.5(a) to (b,c), we do not observe a clear association between their waveforms. However, by using our *CMA* with this AU as inputs, the predicted structural vibration segment is shown in Figure 2.5 (d), which shows a high similarity to (a). On the other hand, if we replace the input of the wearable segment with a signal segment of the same dimension with value 0, i.e., a segment has no information, the predicted segment is as shown in Figure 2.5(e), which demonstrates a lower similarity to (a). The AP of the associated IMU sensor (#1 47%) is higher than that of the other two IMU sensors (#2 26% and #3 27%). The AP between all zeros sequence (33%) and unassociated IMU sensors (#2 35% and #3 32%) are similar to an even distribution (random guess 33%). Therefore, the association probability can reflect the association for cross-modal signals.

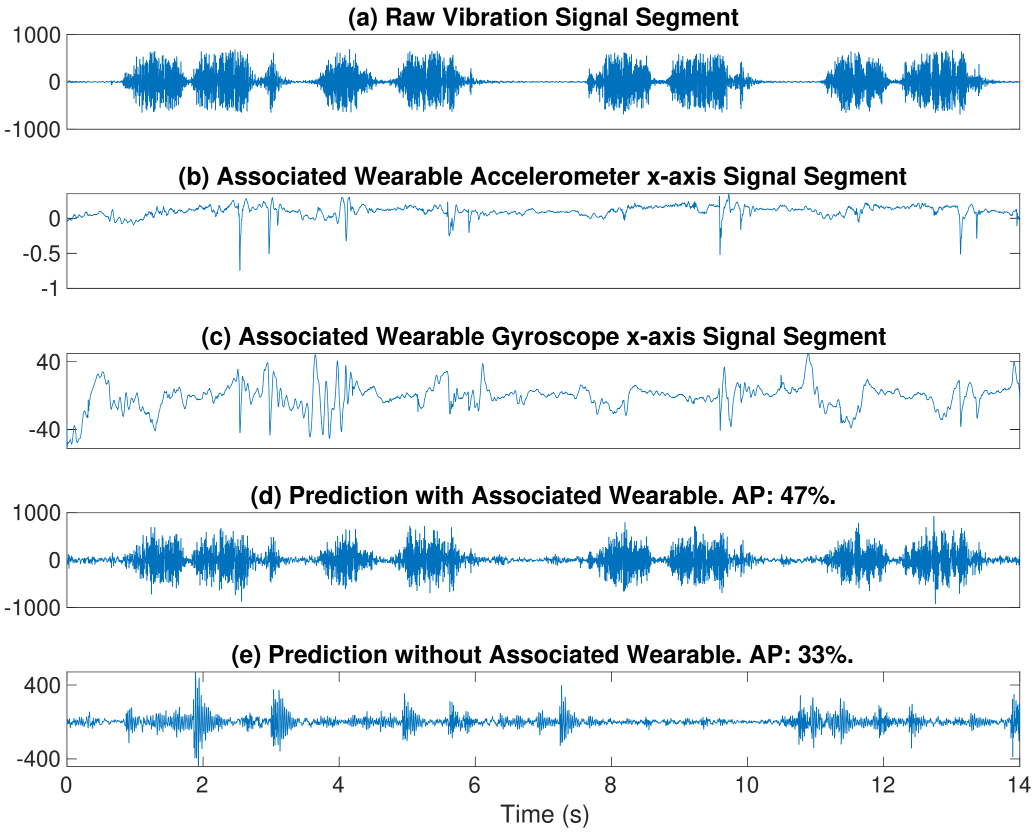


Figure 2.5: One example of associated structural vibration (a) and wearable (b, c) signal segments, and the predicted structural vibration segment with (d) and without (e) associated wearable segment. We observe that the structural vibration segment predicted with the associated wearable's signal shows higher similarity to the raw structural vibration signal segments. *CMA* outputs AP of (d) and (e) as 47% and 33%, respectively. This AP difference indicates that the *AD-TCN* learns the implicit shared context between the structural vibration and wearable segments.

2.4 Experiment Setup

We evaluate *CMA* from two aspects: 1) the association performance and system characterization on the public dataset and our collected uncontrolled dataset. 2) use case study for real application demonstration. We first conduct a set of controlled experiments for system characterization on the public dataset, including hyperparameter configuration, the impact of human activity category, and AP distribution. Then, we evaluate the performance of uncontrolled experiments for robustness verification. Finally, we implement two use cases on the public dataset to demonstrate how to adapt *CMA* in real applications, including occupant identification and multimodal human activity recognition.

In this section, we introduce the two datasets (one open-sourced and one real-world collected), ground truth, evaluation metrics, as well as the implementation of baselines, *CMA*, and two use cases. The experiments are conducted based on the guideline approved by the University Institutional Review Board (IRB) review.

2.4.1 Datasets Description

Public Dataset The dataset [59] includes both structural vibration and wearable sensors – floor vibration sensors and on-wrist IMU (6-axis) sensors. The dataset is collected over two buildings with six human subjects with nine types of in-home activities of daily living. The nine types of in-home activities of daily living are keyboard typing, using mouse, handwriting, cutting food, stir-fry, wiping countertop, sweeping floor, vacuuming floor, open/close drawer. For each **scenario**, i.e., one building one human subject conducting nine types of activities, signals from four vibration sensors deployed in the house, and one IMU sensor deployed on the human subject's wrist are collected. Each human subject conducts the same set of activities in each **scenario** for 10 times and each time for approximately 15 seconds. The sampling rates of the vibration sensor and the IMU sensor are 6500 Hz and 235 Hz, respectively. The dataset also contains the ground truth of activity types, and start and end timestamps.

Continuous Uncontrolled Dataset We adopt the same types of sensors, and sampling rate as the public dataset [59] and collect the continuous uncontrolled datasets over five houses. We recruit 11 human subjects in total, and maintain three subjects per house for the data collection. In each house, we deploy three vibration sensors on the surface of the furniture (desk, kitchen bar, etc.) to capture the subject induced vibration signals, including the kitchen area, living area, and dining area. Considering there are ~ 2.5 people per household on average in the United States in 2021 [20], we invite three participants to cohabit in each house, and each participant wears an IMU sensor on their wrist. We collect the six-axes IMU data (three-axes accelerometer and three-axes gyroscope) from three participants simultaneously. The duration of data collection in each house is around one hour. The participant conducts their daily activities in each area: cooking in the kitchen area, eating in the dining area, and watching TV or surfing on the Internet with a laptop in the living area. To reflect the diversity of participants' activities, the participant can do any activity in each area as natural as possible. For example, the subject can cook any food they like; some subjects cook potatoes, some cook sandwiches. In practice, the sampling rate of the vibration sensor and the IMU sensor are around 4000 Hz and 250 Hz, respectively. We also deploy a camera in each area to record which participant is active in this area.

2.4.2 Ground Truth of Pairwise Association and Dataset Preparation

The cross-modal association problem is described as determining if the signals from two sensing modalities for a given period are induced by the same physical event, which is the individual activity in our case. For an AU, the ground truth of the association between the vibration signal and the IMU signal is true if and only if the vibration signal is induced by the individual wearing the IMU.

Public Dataset. To utilize this dataset for evaluating *CMA* on the task of cross-modal association, we generate association ground truth based on the provided original activity ground truth. We first detect and segment each activity event based on the provided start and end timestamp of each activity event. For each activity segment with signals from four vibration sensors and one IMU sensor, we select the vibration sensor with the highest signal-to-noise ratio (SNR) as the signal associated with the corresponding IMU sensor. We go through the entire dataset and generate 1048 pairs of the cross-modal association data segments (each ∼10s). For any two cross-modal segments $VibSig_i$ and $IMUSig_j$, the association labe is true if $i = j$, otherwire is false.

For each **trial**, we randomly select N segment pairs from the candidate set (it can be the full set with 1048 pairs or a subset). We apply *CMA* on each $VibSig$ with all the $IMUSig_{1,\ldots,N}$ and output N APs between the $VibSig$ and N $IMUSig$. To reflect the practical scenario of a home with parents and children, we set the default value for N as 3. For each **experiment**, we repeat this **trial** at least 100 times to reduce the random selection bias.

Continuous Uncontrolled Dataset. For the continuous uncontrolled dataset, we first apply the event detection and activity segmentation (introduced in Section 2.3.1) on each vibration sensor. The vibration segment and other segmented IMU segments combine a AU. We determine the association ground truth of this AU by watching the recorded video in the vibration sensor deployed area, and we consider the human subject who appears in this area during this event period as the inducer of this event. For each experiment, we use all detected AUs in one house to evaluate the performance of *CMA* in real-world experiments and evaluate the robustness of *CMA* by comparing the performance variation in different houses.

2.4.3 Evaluation Metric

We consider two metrics in the evaluation: 1) the ROC curve and its AUC value to evaluate the performance in all thresholds, 2) F1 score and accuracy to evaluate the performance in a selected threshold. In this work, we usually use the former metric to evaluate *CMA* and the baseline methods, and use the latter metric to provide an intuitive evaluation of the overall performance in the public dataset and continuous uncontrolled dataset.

ROC Curve and AUC value

In our sensor signal association problem, both the true positive (i.e., the structural vibration sensor's signal is associated to the wearable sensor that causes vibration) and false positive (i.e., the structural vibration sensor's signal is not associated to the non-causal wearable sensor) are important performance indicators. Therefore, we adopt ROC (Receiver Operating Characteristic) curve and AUC (Area under the ROC Curve) [61] to evaluate each **experiment**. ROC curve is a probability curve that systematically depicts how the performance (true and false positive rates) change across the entire range of thresholds [40]. To generate the ROC curve, we apply different AP thresholds θ_{AP} and calculate the true positive and false positive rate. AUC measures the quality of the association irrespective of threshold values [56]. The higher AUC value indicates a better performance.

F1 score and Accuracy

Since the final output of *CMA* is a pairwise association between two modalities, we further threshold the AP and calculate the F1 score [124] and accuracy. For each AU, if the IMU segment association matches with the ground truth, we consider it as a true positive (TP). If the associated IMU ID does not match with the association ground truth, we consider it is a false positive (FP); and vice versa, for a false negative (FN).

Table 2.1: Signal similarity metrics for signals X, Y of length l.

Metric	Equation
MCC	MCC(X, Y) $= \max_{k=0,\ldots,l} \frac{\sum_{i=1}^{l} x_i \cdot y_{i+k}}{\sqrt{\sum_{i=1}^{l} x_i^2} \cdot \sqrt{\sum_{i=1}^{l} y_i^2}}$
Cosine Similarity	CS(X, Y) $= \frac{\sum_{i=1}^{l} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{l} x_i^2} \cdot \sqrt{\sum_{i=1}^{l} y_i^2}}$
Surface Similarity	SS(X, Y) $= \frac{\sqrt{\sum_{i=1}^{l} (x_i - y_i)^2}}{\sqrt{\sum_{i=1}^{l} x_i^2} + \sqrt{\sum_{i=1}^{l} y_i^2}}$

The precision and recall are calculated as $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$. The F1 score is a function of precision and recall, $F1\ score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$. The accuracy is the percentage of correctly determined association cases and unassociation cases over all cases.

2.4.4 Baseline Methods

We consider measuring the shared context or similarity between cross-modal signals as baselines, so we evaluate *CMA* against three commonly used signal similarity metrics [114]. For vibration data segments $VibSig_i$ and IMU data segments $IMUSig_j$, we calculate 1) Cosine similarity (CS), 2) max cross-correlation (MCC), 3) Surface similarity (SS) between them as shown in Table 2.1. For IMU signals with six axes, we calculate the signal similarity between each axis and the vibration signal and report the highest similarity over all six axes. For all the baseline methods, the higher value between $VibSig_i$ and $IMUSig_j$ means that the vibration segment i is more likely to be associated with IMU segment j.

2.4.5 *CMA* Implementation

Multimodal Signal Alignment Since the sampling rate for the vibration sensor and the IMU sensor are different in the two datasets, we resample the vibration sensing data from 6500 Hz to 235 Hz for the public dataset and resample the vibration sensing data from 4000 Hz to 250 Hz for the continuous uncontrolled dataset to align the multimodal signal inputs. We utilize the resample function [81] in Matlab to re-sample the data. We use the recorded timestamp to align the vibration sensing data with the IMU sensing data for the uncontrolled dataset. We empirically set the energy threshold θ_e as eight, and the threshold of event interval $\Delta\tau$ as four seconds. We set the upper bound of activity segments τ_u as 20 seconds and the lower bound of activity segments τ_l as eight seconds.

Association Discovery Then for the *AD-TCN* model training, we use the Stochastic Gradient Descent algorithm, and ADAM [72] as optimizer. We set the maximum training epochs as 6000. To avoid the impact of over-fitting or under-fitting of *AD-TCN*, we apply the early stopping method to automatically stop the training based on the loss decrease [140]. We use *ReduceLROnPlateau* function [35] which is integrated into PyTorch [102] to implement early stopping and set the factor and patience parameter as 0.5 and 4, respectively. We terminate the training when the learning rate drops to less than 0.001 (initially 0.01). Parameters of dilation and stride in *Conv1d* [34] are both set as 1.

Association Threshold We consider the output of the softmax function (section 2.3.3) as the estimated AP between N IMU segments. If all IMU segments are not associated with the vibration segment, the ideal distribution of AP should be a uniform distribution. So we select $1/N$ as the association threshold for *CMA*. For the baseline methods, we select the mean value over all detected events in each experiment set (100 trails in the public dataset) as the threshold to determine the association.

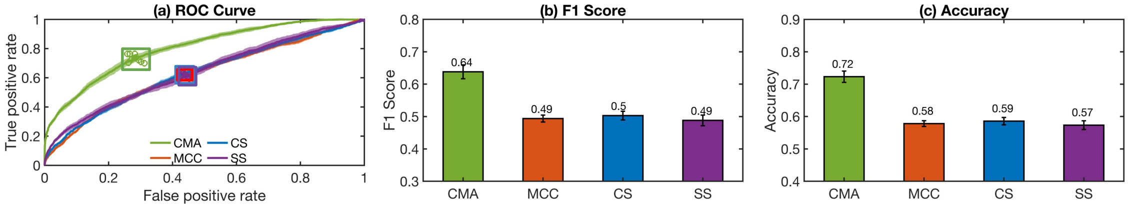


Figure 2.6: Association performance with the public dataset. (a) shows the average ROC curve and the standard deviation (width of the curve) of false positive rate and true positive rate in 10 experiments. (b) and (c) shows the F1 score and accuracy calculated from the circled data points in (a), respectively.

Once the baseline values (CS, MCC, SS) between the vibration segment and the IMU segment is larger than this threshold, we report they are associated.

2.4.6 Use Case Study

We implement the two aforementioned uses cases on the public dataset [60] due to the availability of the identity and activity labels. We consider the use case scenario of three participants co-habit in a house. We investigate three association conditions: 1) Ideal association (ground truth). The pair of IMU and vibration data are of their true associations. 2) *CMA* association. The pair of IMU and vibration data are based on the *CMA*'s output. 3) Random association (baseline). The pair of IMU and vibration data are randomly assigned. For learning models, we randomly select 80% data for training, and the rest for testing.

Occupant Identification In scenarios of vibration-based in-home elderly or patient monitoring [52, 100], it is challenging to acquire the identity labels of each occupant's vibration signals to bootstrap the learning model in the real-world deployment. We envision a temporary setup with the IMU sensor could be used with our *CMA* association scheme to provide initial identity labels for the learning model for a household of three people. We run *CMA* to acquire the identity label of the structural vibration signal segments, and then we train an SVM model [52] on these segments with pseudo label from the association. We report the identification accuracy values over the three association scenarios.

Multimodal Human Activity Recognition (HAR) In this use case, we conduct the multimodal human activity recognition (HAR) [60] to depict the cross-modal association's importance. Instead of directly fusing two types of sensor data with random association, or provide manual labeling of this association (ideal), we leverage *CMA* to provide this information. Then this association will determine the input IMU-vibration signal pair to the multimodal learning training and testing for activity recognition. We use a same fully connected neural network as a classifier to recognize the occupant activity [60]. The model is trained with a cross entropy loss and the Adam optimizer. We report the accuracy of nine activities recognition over the three association scenarios.

2.5 Results and Analysis

In this Section, we first introduce the overall performance (Section 2.5.1) of *CMA*, the impact from data (Section 2.5.2), and *CMA* configuration (Section 2.5.3) over the public dataset. Then we further show the performance in the continuous uncontrolled dataset (Section 2.5.4). Finally, we demonstrate the performance of *CMA* in two user cases (Section 2.5.5).

2.5.1 Overall Performance

In the overall performance experiment, we randomly select three pairs of segments out of the full set (1048 pairs) to conduct the overall performance evaluation with the experiment procedure introduced in Section 2.4.2. Figure 2.6(a) shows the ROC

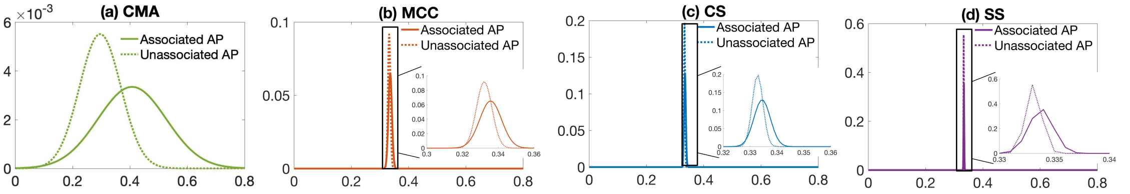


Figure 2.7: The distribution of associated and unassociated AP of *CMA* and baselines.

curve of *CMA* and baseline methods. The solid line presents the average value of the ROC curve, and the area around the line presents the standard deviation of the 10-repetition experiments. We observe that the ROC curve of *CMA* is always above those of the baseline methods, which indicates a better association accuracy. If we consider a tolerable false positive rate of 0.2, the average true positive rate for *CMA* can achieve 0.63, which is up to 1.5× (50% improvement) of the baselines (MCC 0.39, CS 0.40, SS 0.41). The average AUC value of *CMA* achieves 0.80, which is up to 30% improvement compared to the baselines (MCC 0.63, CS 0.62, and SS 0.64). Figure 2.6(b) and (c) show the F1 score and accuracy calculated from the circled data points in Figure 2.6(a). The average F1 score of *CMA* and baselines achieve 0.64, 0.49 (MCC), 0.50 (CS), and 0.49 (SS), respectively. *CMA* achieves 1.3× F1 score value of the baseline methods (up to 31% improvement). The average accuracy of *CMA* and baselines achieve 0.72, 0.58 (MCC), 0.59 (CS), and 0.57 (SS), respectively. The accuracy of *CMA* achieves up to 26% improvement than the baseline methods.

AP Distribution We also demonstrate the distribution of associated and unassociated AP to further analyze the performance of *CMA* and baselines. For the baselines, we adopt the softmax function to convert the metric values between two cross-modal segments to association probabilities (Equation 2.6). Figure 2.7 shows the AP distribution of *CMA* and baselines. We can observe that the distribution of associated and unassociated AP of *CMA* has less overlap than the baselines, which indicates the estimated AP value of *CMA* is more separable.

Table 2.2: Types of activities and cross-modal association levels.

Assoc. Levels	Activities
direct	cutting food, stir-fry, open/close drawer
indirect	keyboard typing, handwriting, vacuuming
semi-direct	using mouse, wiping countertop, sweeping

2.5.2 Impacts of Data: Activity Category and Association Levels

One potential factor that may impact the association performance is the type of activities, because the association level varies for different activities. For some activities, the motion measured by the wearable also directly induces structural vibration. For example, when people cut food, their wrist motion (measured by the IMU) **directly** causes the knife to impact the cutting board (measured by vibration sensors). On the other hand, for some activities, the motion measured by the wearable does not directly associate with the structural vibration. For example, vacuuming the floor causes the floor to vibration due to motor vibration, which does not directly but **indirectly** causes structural vibration via wrist motions. Therefore, we categorize the nine types of activities into three levels of association – direct, indirect, and semi-direct – in Table 2.2.

Activity Category Combinations

To demonstrate *CMA*'s robustness over types of activities with different association levels, we randomly select four pairs of segments out of subsets of pairs with different types of activities – direct associated activities, indirect associated activities, semi-direct associated activities, and mixed activities. Then we follow the same experiment

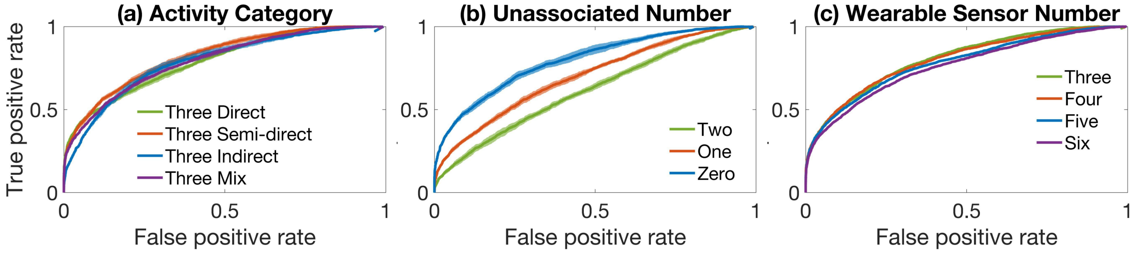


Figure 2.8: Roc curve of *CMA* in the public dataset analysis. (a), (b), and (c) show the performance of *CMA* for impact of activity category, unassociated number, and wearable sensor number, respectively.

procedure in Section 2.4.2.

Figure 2.8(a) depicts the average ROC curve on 10 repetition experiments. The average AUC value of *CMA* is 0.79 (direct), 0.82 (semi-direct), 0.79 (indirect), 0.80 (mix). Three baselines depict overall lower than 0.7 AUC values. *CMA* achieves the best performance in all activity category combinations. Furthermore, *CMA* demonstrates robustness over different activity categories, while the baselines have inconsistent performance with the AUC value varying between 0.6 and 0.7.

Unassociated Combinations

To better understand how *CMA* performs in the real scenario, we further evaluate when some of the vibration signals are generated by occupants without an IMU sensor. We randomly select three pairs of signals ($VibSig_i$ and $IMUSig_j$) from the full set of pairs (1048) and investigate the scenario where for 0/1/2 of them $i \neq j$ and the rest $i = j$. Then we follow the same experiment procedure in Section 2.4.2 and compare the AUC values when there are different numbers of unassociated pairs among the three.

Figure 2.8(b) shows the average ROC curve of *CMA* on 10 repetition experiments. Overall, when the number of unassociated pairs increases, the AUC value decreases.

This could be because the prediction of the unassociated infrastructural signal is done with multiple IMU signals equally not associated, which results in similar APs that is not efficient for distinguishing the association relationship. When there is one unassociated signal pair, *CMA* achieves an AUC value over 0.7, while the baselines only achieve 0.57, 0.56, and 0.58, respectively (random selection's AUC value is 0.5). *CMA* also achieves the best performance than baselines.

Wearable Sensor Number

To better understand the scalability of *CMA*, we further evaluate *CMA* when the number of wearable devices N is larger than 3. In this experiment, we first randomly select three pairs of signal segments from the full set of pairs. Then we further randomly select extra numbers of $IMUSig$ and apply *CMA* to associate $M = 3$ number of $VibSig$ and N number of $IMUSig$, where $N = 3, 4, 5, 6$. Then we follow the same experiment procedure in Section 2.4.2.

Figure 2.8(c) shows the average ROC curve of *CMA*. When the number of wearable sensor N increase from 3 (the same as the number of vibration sensors M) to 6, the average AUC values decreases slightly (≤ 0.05) with the number of wearable increase. This is because the difficulty of finding the associated IMU segment increases when the number of IMU segments N increases. In summary, *CMA* also works for the scenario that is more than three people.

2.5.3 Impacts of *CMA* Configuration

We further explore the impact of the hyper-parameter configuration of *CMA* on the performance. As introduced in Section 2.3.2, *CMA* contains three hyper-parameters: 1) hidden layer number L, 2) receptive field F (adjusted by kernel size K), and 3) input AU length η. The default values for these hyper-parameters are shown in Table 2.3. We randomly select three pairs of segments from the full set (1048 pairs) and

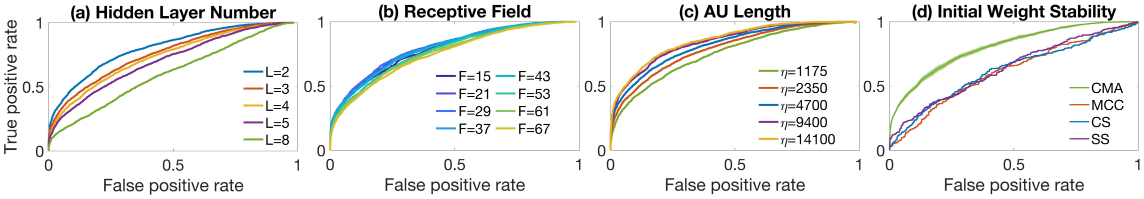


Figure 2.9: Impact of *CMA* configures and repeatability of *CMA*. (a), (b), and (c) show the performance of *CMA* under the different configurations of hidden layer number, receptive field, and AU length, respectively. (d) shows the ROC curve of *CMA* and baseline methods when we repeat *CMA* 10 times on the same experiment set.

conduct experiments with the procedure introduced in Section 2.4.2 with varying *AD-TCN* hyper-parameters.

Hidden Layer Number L

Hidden layer number directly impacts the complexity of the neural network. Therefore we investigate how the model acts at different levels of complexity for the cross-modal time series prediction. We increase L from 2 to 8, and demonstrate the average ROC curve of *CMA* in Figure 2.9(a). The average AUC value of each configuration are 0.81, 0.76, 0.74, 0.71, 0.61, respectively. We observe that *CMA*

Table 2.3: *CMA* hyper-parameters

Parameters	Default	Controlled Experiment Values
L	2	2,3,4,5,8
F	29	15, 21, 29, 37, 43, 53, 61, 67
η	2350 to 3055	1175, 2350, 4700, 9400, 14100

achieves the highest AUC value when the L is set to 2. This result indicates that a shallow architecture is more suited for the cross-modal association task. It could be because the association discovery task is fundamentally a binary classification task, and the model can be presented with a simple network architecture sufficiently. A large L value may cause the network to overfit [12]. When the overfit occurs, the network cannot generalize to test data, hence is not able to make accurate prediction [49]. Under this circumstance, the calculated association score is not reliable for the association discovery.

Receptive Field F

The receptive field F is determined by both the hidden layer number L and the causal convolutional layer's kernel size K as $F = (K - 1) \cdot L + 1$ [95]. It describes how 'far' the model can 'see' to predict the current samples [77]. For example, Figure 2.4(b) shows an example of a causal convolutional layer with a kernel size $K = 2$. If layer number $L = 2$, then receptive field $F = (2 - 1) \cdot 2 + 1 = 3$.

Figure 2.9(b) shows the average ROC curve in different receptive field configurations. When F increases, the average AUC value first increases then decreases (0.79, 0.80, 0.81, 0.80, 0.79, 0.78, 0.76, and 0.75 for F from 15 to 67). *CMA* demonstrates a stable performance and achieves the highest average AUC value when F is 29. One explanation for why *CMA* achieves the highest AUC with $F = 29$ is that the time duration for 29 samples is approximately 0.1 second, which is approximately the duration for an arm motion to cause an impulsive vibration signal. Therefore, this amount of 'history' data is most helpful for the prediction of current sample value.

Input AU Length η

The input AU length η determines how much data is available to calculate AP and determine the association relationship. Intuitively, the longer the observation data is, the more accurate the time-series prediction model is, and hence the network parameter that describes the association relationship is more accurate.

Figure 2.9(c) shows the average ROC curve of *CMA* when the input AU length η varies from 1175 ($\tau = 5$ seconds) to 14100 ($\tau = 60$ seconds). With the increase of η, the performance of all evaluated methods increases. We select η taking into account the trade-off between the prediction accuracy and the data practicality. Since our assumption is that the signal association within η is invariant, it means the higher the η, the more unlikely the assumption holds. For the public dataset, we consider the default value of η is 2350 ($\tau = 10$) because the duration of activity from the public dataset is in the range of 10 to 15 seconds.

AD-TCN Initial Weight Stability

The initial weight assignment can directly impact the neural network model and the performance [47]. Therefore, we also investigate the repeatability of *AD-TCN* with different random initial weights. We randomly select three pairs of segments out of the full set, and conduct the *AD-TCN* training with different initial randomization 10 times. We repeat this random selection 110 times to avoid sampling bias.

Figure 2.9(d) show the average ROC curves of *CMA* and baselines when we train *AD-TCN* on the same dataset 10 times with different random initial weights. The green line shows the average false positive rate and true positive rate, and the green area around the green line shows the standard deviation of 10 times of weight initialization. *CMA* demonstrates a stable performance when the weights of the neural network module are initialized differently.

2.5.4 Robustness in Uncontrolled Deployment

Figure 2.10 (a) shows the average ROC curve and the standard deviation of false positive rate and true positive rate of *CMA* and baselines in five houses dataset.

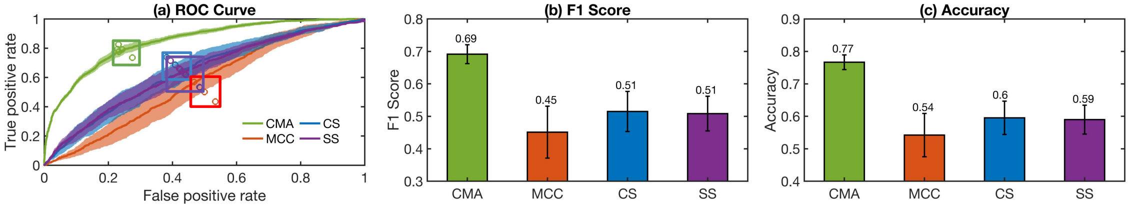


Figure 2.10: Overall performance with the uncontrolled dataset. (a) shows the average ROC curve and the standard deviation (width of the curve) of false positive rate and true positive rate in different houses dataset. The circle on the curve indicates the false positive rate and true positive rate when *CMA* operates with the selected threshold. (b) and (c) show the F1 score and accuracy under the selected association threshold, respectively.

We can observe the performance of *CMA* is better than the baselines and the false positive rate and true positive rate is more stable. The average AUC value of *CMA*, and baselines are 0.85, 0.64 (MCC), 0.56 (CS), and 0.64 (SS), respectively. The AUC value of *CMA* achieves 0.85, which is up to 37% improvement compared to the baselines.

The circle marks in Figure 2.10(a) indicate the false positive rate and true positive rate under the selected association threshold (introduced in Section 2.4.5). Figure 2.10(b) and (c)demonstrates the F1 score and accuracy, respectively. The average F1 score of *CMA* and baselines achieve 0.69, 0.45 (MCC), 0.51 (CS), and 0.51 (SS), respectively. The F1 score of *CMA* achieves 0.69, which is up to 53% improvement compared to the baselines. The average accuracy of *CMA* and baselines achieve 0.77, 0.54 (MCC), 0.60 (CS), and 0.59 (SS), respectively. The accuracy of *CMA* achieves 0.77, which is up to 43% improvement compared to the baselines.

We also observe that compared with the performance in the public dataset, the performance of *CMA* in the uncontrolled dataset is 0.05% better (AUC value 0.8 vs. 0.85, F1 score 0.64 vs. 0.69, accuracy 0.72 vs. 0.77). This might be because in the

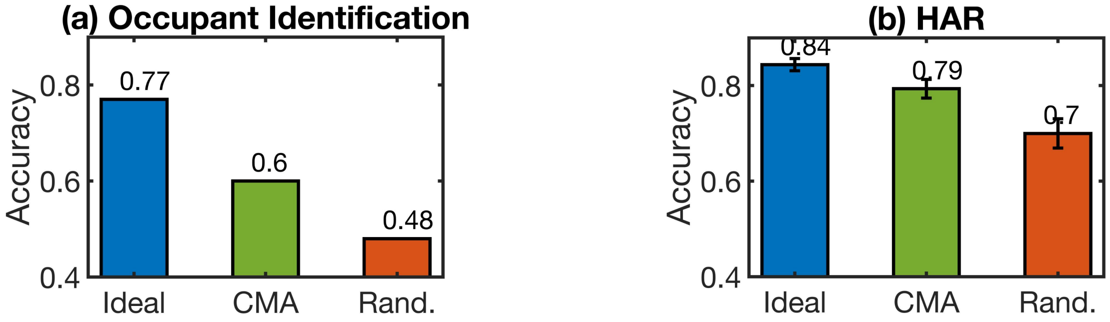


Figure 2.11: The performance of two use cases with cross-modal association information provided by *CMA*.

uncontrolled dataset, the three human subjects are more likely to conduct different types of activity at the same time than in the public dataset. Finding the association relationship from the same type of activity is more difficult since the IMU segments of the same type of activity are more similar to each other.

2.5.5 Use Case Performance

Figure 2.11 shows the accuracy of *CMA* compared with baselines for two use cases. The blue, green and red bars represent of ideal association (ground truth), *CMA* association, and random association (baseline), respectively. We observe that with the association provided by *CMA*, both use cases demonstrate an improvement in accuracy compared to the baseline. For occupant identification, the system achieves a 12% accuracy increase with the pseudo label provided by *CMA* without any manual label. For HAR, *CMA* achieves approximate 10% accuracy improvement compared to without the association information, and it is only 5% lower than the accuracy with ideal association. Such improvement is promising, considering that it is made with leveraging the pervasive wearable IMU data, and without requirements of any label data.

2.6 Discussion

Temporal Overlapping and Activity Segmentation In this work, we focus on the cross-modal segment-level association problem with the assumption of no temporal signal overlapping of multiple sources at one structural vibration sensor. If one structural vibration sensor captures overlapped signals from multiple activities, the implicit shared context can be learned for association purposes will be more constrained than what has been investigated in this work and therefore more challenging. In the future, we plan to explore either leveraging hierarchical temporal information over different time resolutions, or combining frequency domain analysis to tackle the signal temporal overlapping challenge.

Activity segmentation is another important aspect of indoor occupant sensing. In this work, we adopted the lumping algorithm [51]. Our uncontrolled experiment result inherits the segmentation error from the lumping algorithm. In the future, we will explore incorporating other activity segmentation schemes. Furthermore, we will explore jointly conducting the separation and segmentation with *CMA* to further improve the robustness.

Association-Aware Multimodal Learning With the segment-level association learned for each segment, we can further use this learned information to enhance the existing multimodal learning. For example, the association can be used as a dynamic sensor selection criteria to allow the inference models to adapt to input channels, as well as a regularization to reduce the chance of learning a spurious relationship between input channels and data labels. For graph neural network-based models, this association may be used as the prior knowledge to establish the graph, ensuring a more efficient and robust inference [82].

Modality Generalizability In this work, we evaluate *CMA* with the combination of structural vibration sensing and wearable on-wrist IMU sensing. *CMA* is designed for general time series sensing modalities, and in the future, we plan to explore more modalities (e.g., acoustic, event camera, electricity load, physiological sensors) combination to further understand its limitation and generalizability. For the high-dimension sensing data, we can build an encoder to convert the high-dimension data to one-dimension sequences, such as data2vec [10]. On the other hand, association learning is more challenging for modalities with a latent and longer dependency. For example, when the occupant turns on the heater, the indoor temperature becomes warmer, and the occupant's heart rate will slowly go higher [19]. In this case, the association between the electricity load sensor and physiological sensors (heart rate monitor) data is latent and potentially requires a new framework for association learning.

Computational Requirements of *CMA* In our experiment, the time consumption of *CMA* for one AU is around 10 seconds in an Apple MacBook Pro 2022 using CPU only. In this work, we focus on providing a data-driven method to discover the association relationship between two modalities without the requirements of label data. However, the time consumption can be decreased by optimizing multiple factors, such as the code implementation framework, and adopting parallel computing. The current computation is on the server side, and in the future, we would also consider offloading the computation to the nearby devices with an event-driven design on the embedded platform side.

2.7 Chapter Summary

In this Chapter, we present *CMA*, a cross-modal signal segment association scheme between wearable and structural vibration sensors. We introduce *AD-TCN*, a TCN-based framework, to calculate the amount of shared context between signal segments from two modalities. After training the network, we calculate the association probability based on the weights of the trained *AD-TCN*, and determine the pairwise segment association. We evaluate *CMA* via a public multimodal dataset for systematic evaluation, and we collect a continuous uncontrolled dataset for robustness evaluation. *CMA* achieves up to 37% AUC value, 53% F1 score, and 43% accuracy improvement compared to baselines.

Chapter 3

Cyber Augmentation for Robust Data Acquisition in Real-world Environment

3.1 Introduction

IoT systems are becoming more and more pervasive in people's daily life. Due to their increasing applications and advantages in deployment (e.g., sparse, privacy preserving), many non-intrusive indirect sensing techniques are developed for indoor human information acquisition, including RF-, vibration-, light-based methods. However, the indirect sensing mechanisms of these systems also induce large variances of the acquired data quality over deployment environment conditions and configurations, which reduces the system performance. We focus on structural vibration-based indoor sensing due to its passiveness, non-intrusiveness, room-level sensing range enabling extraction of fine-grained information [13, 32, 96]. The system's information inference performance (e.g., detection rate, learning accuracy) is impacted by the deployment environment. To systematically understand these deployment environment impacts, we define **sensing quality** as a series of measurable factors/models reflecting how they impact a given information inference task. Quantifying sensing quality allows further enhancement of deployment efficiency to improve IoT sensing systems' performance.

Compared to prior works on signal quality assessment, which are mainly used in the domains of communication [9, 16, 64, 119] and computer vision [75, 125, 132], our proposed *AutoQual* reflects effects of deployment environment on the sensing task performance. The data quality assessments [21, 104] target the evaluation of the existing dataset and provide comparisons between multiple acquired datasets. However, they do not quantify the environmental impacts to acquired data characteristics. Our prior work on application-oriented sensing signal quality (SSQ) proposes a system-level signal quality assessment scheme with a set of metrics [151] and demonstrate the possibility of using measurements of these metrics for optimal sensor placement selection [145]. However, this sensing signal quality assessment requires manual calibration with known excitation of dense coverage. As a result, the approach is labor-intensive and impractical for large-scale sparse deployment assessment. In addition, the SSQ model combines different factors' measurements heuristically, which makes the approach difficult to generalize.

In this Chapter, we present *AutoQual*, an autonomous sensing quality assessment framework to quantify impacts of deployment environment on IoT sensing system performance. We take the structural vibration-based indoor human sensing system as an example and apply *AutoQual* for an autonomous sensing quality assessment on multiple sensing tasks. The main **challenges** include 1) how to identify and quantify environmental characteristics that impact the system performance, 2) how to integrate these AFs to assess the system for a given sensing task – different sensing tasks may be sensitive to different AFs, and 3) how to achieve 1) and 2) autonomously without manual efforts. We tackle these challenges by 1) utilizing domain knowledge on wave propagation and structural properties to identify a set of AFs and design the measuring method accordingly (Section 3.3.1), 2) adopting a data-driven approach to estimate the relationship between measured AFs and sensing tasks' performance (Section 3.3.3), 3) automating AF measurements using human-induced vibration signals extracted by co-located mobile devices (Section 3.3.2). The contributions of this work are as follows:

- We present *AutoQual*, a framework of autonomous task-oriented sensing quality assessment that predicts the IoT system performance utilizing the mobility of ambient occupants.
- We identify a set of measurable environmental factors that determine the sensing quality.
- We propose an auto-assessment scheme via human-induced signals enabled by co-located mobile sensing context.
- We evaluate *AutoQual* through real-world experiments at 48 deployments in 11 environments on multiple sensing tasks.

The rest of the Chapter is organized as follows. First, Section 3.2 introduces related works on IoT system quality assessment. Next, Section 3.3 presents the auto-assessing system design leveraging the co-located mobile devices. Then, Section 3.4 and Section 3.5 explain the evaluation experiments and result analysis. Furthermore, we discuss the limitation of this work and future directions in Section 3.6. Finally, we summarize the Chapter in Section 3.7.

3.2 Related Work

We consider the following aspects of related work including cross-modal system with infrastructural and mobile sensing (Section 3.2.1), sensing data quality measuring metrics (Section 3.2.2), and CPS-IoT system performance quantification methods (Section 3.2.3).

3.2.1 Infrastructural and Mobile Cross-Modality Sensing

Prior work has been done on combining infrastructural and mobile sensing to acquire target information, such as human activity recognition [41,58], and air quality monitoring [27,38,43,65,135,136]. The infrastructure- and mobile-based subsystems often provide complementary data for each other to achieve a higher accuracy or a finer granularity of information [97].

Although they combine the infrastructural and mobile sensing to improve system performance, they are usually focus on a specific sensing task or application instead of assess sensing quality or cross-modal system characterization. To the best of our knowledge, our work is the first sensing quality assessment framework that fuses the infrastructure and mobile sensing to assess sensing quality. Our work utilize the event label information from the mobile data to achieve autonomous sensing quality assessment for structural sensing.

3.2.2 Signal Quality Metrics

Prior work on signal quality measurements mainly focuses on measuring general signal properties, such as signal-to-noise ratio (SNR) [94], or signal structural similarity (SSIM) [75]. When it comes to different types of signals for particular tasks, specific metrics have been developed. For example, wireless communication quality measurements include but are not limited to Received Signal Strength Indicator (RSSI) [119], Carrier-to-Noise Ratio (CNR) [64], Signal-to-Interference-plus-Noise Ratio (SINR) [9], and Link Quality Indicator (LQI) [16]. Similarly, computer vision quality measurements include and not limited to image quality index [125], Structure SIMilairty (SSIM) [75], and universal image quality index [132]. These metrics, when applied to structural vibration-based approaches inferring information indirectly, do not model the deployment characteristics that impact the performance of tasks with different requirements. In this work, we focus on modeling 1) the deployment characteristics via ambient sensing signal and 2) the relationship between these deployment characteristics and different sensing tasks (which we refer to as the sensing quality metrics).

3.2.3 CPS/IoT Performance Quantification

Various frameworks or metrics have been proposed to quantify CPS/IoT systems' performance by measuring their data quality. For example, Karkouch et al. define the IoT data quality with a multi-dimensional definition including accuracy, confidence, completeness, data volume, and timeliness [69]. While Banerjee et al. define the application-driven IoT quality from the computer system perspective [14]. Other CPS/IoT performance assessments include communication performance and RFID-based health care application [116, 127]. However, none of these prior works systematically evaluate the sensing (or data acquisition) process and identify the environment characteristics that impact this process, and eventually determine the CPS/IoT application performance.

3.3 *AutoQual* System Design

We present *AutoQual*, a cyber-physical system sensing quality assessment framework that leverages indoor occupant mobility for automating the assessment of structural vibration-based indoor human sensing systems. Figure 3.1 shows an overview of *AutoQual*, consisting of three major components 1) vibration sensing assessment factor measurements, 2) mobile sensing context extraction for auto-assessment, and 3) Task-oriented Sensing Quality (TSQ) scoring model.

When occupants walk in a target sensing area, they are sensed by both vibration- and mobile-based systems, assuming a set of sensing systems are deployed on various indoor environments for data acquisition. This ambient sensing data of occupants is then used to train a model that describes the deployment environment characteristics

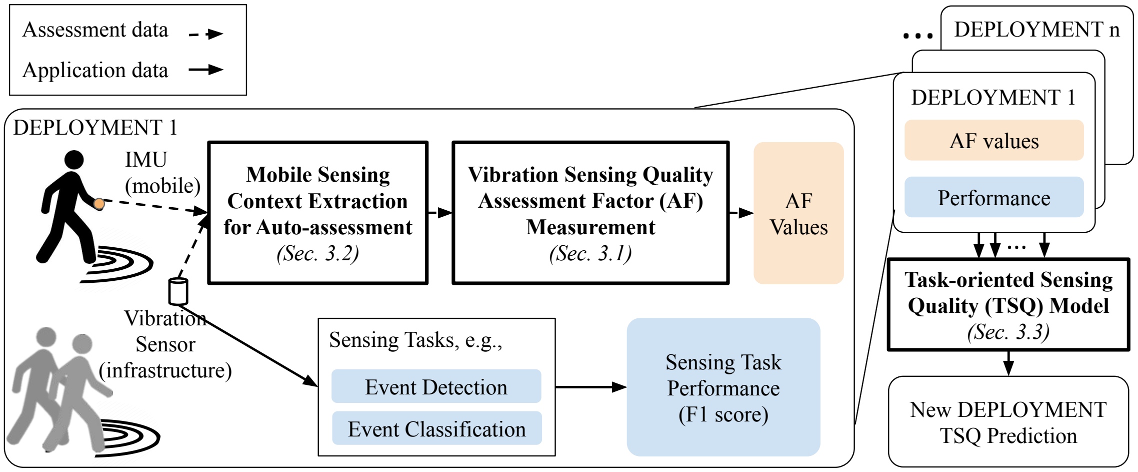

Assessment data
Application data
DEPLOYMENT 1
IMU (mobile)
Vibration Sensor (infrastructure)
Mobile Sensing Context Extraction for Auto-assessment (Sec. 3.2)
Vibration Sensing Quality Assessment Factor (AF) Measurement (Sec. 3.1)
AF Values
Sensing Tasks, e.g.,
Event Detection
Event Classification
Sensing Task Performance (F1 score)
DEPLOYMENT n
DEPLOYMENT 1
AF values
Performance
Task-oriented Sensing Quality (TSQ) Model (Sec. 3.3)
New DEPLOYMENT TSQ Prediction

Figure 3.1: *AutoQual* overview. The dash line arrows represent assessment data (mobile + infrastructural) and the solid line arrows represent sensing task data (infrastructural only).

and predicts sensing tasks' performance under a new deployment (i.e., assessing the sensing quality of the new deployment).

During the assessment data collection, footstep-induced vibrations are detected and grouped by traces based on the shared-context from mobile sensing. These selected assessment data are then sent to the AF measurement module and used to calculate the system's key impact factors – AF values. For the training deployments, the AF values and the system performance are sent to the TSQ scoring model module to establish a data-driven model that predicts the sensing tasks' performance of new deployments.

3.3.1 Vibration Sensing Quality Assessment Factor Measurement

In order to quantify environmental impacts to sensing task performance, we use three physics models to describe the deployment environment characteristics. These models are 1) the attenuation model (AF_1, AF_2), 2) the structural homogeneity model (AF_3), and 3) the structure-sensor coupling model (AF_4, AF_5). *AutoQual* autonomously measures parameters of these models based on the temporal association between mobile- and vibration-based sensing data.

Attenuation Model

When a vibration wave propagates through a solid, its attenuation is determined by the combined effects of geometric wave spreading, intrinsic attenuation, and scattering attenuation [39, 117, 120]. These effects are modeled as a function of propagation distance d

$$Amp(d) = \frac{Amp_0}{\sqrt{d}} e^{-2\alpha d} \tag{3.1}$$

where the decay rate $\alpha = \alpha_s + \alpha_a$, α_a is the coefficient of absorption and α_s is the coefficient describing mean-field attenuation due to scattering [117]. Amp_0 is the initial amplitude of the vibration wave created by an excitation.

AutoQual measures the decay rate α and the initial amplitude Amp_0 of a deployment. However, Eq. 3.1 only describes the ideal attenuation without taking into account ambient noise, which cannot be directly measured in practice. We subtract the logarithmic amplitude of the background noise $\lg Amp_N$ on both side of E.q. 3.1 and get

$$\lg\left(Amp(d)\right) = \lg Amp_0 - 2\alpha d \lg e - \frac{1}{2} \lg d \tag{3.2}$$

$$20 \lg\left(\frac{Amp(d)}{Amp_N}\right) = -\,40\alpha d \lg e + 20 \lg\left(\frac{Amp_0}{Amp_N}\right) - 10 \lg d \tag{3.3}$$

The term $20 \lg\left(\frac{Amp(d)}{Amp_N}\right)$ is the Signal-to-Noise Ratio (SNR) [94] of the signal generated by excitation at distance d to the sensor, which can be directly measured. We estimate $AF_1 = -40\alpha$, and $AF_2 = \lg\left(\frac{Amp_0}{Amp_N}\right)$ by conducting a linear fitting with SNR values measured at locations with multiple excitation-sensor distance d.

Structural Homogeneity Model

The homogeneity of the structure directly impacts the data distribution of sensing signals. When a vibration wave propagates through a solid, the waveform distortion occurs and can be represented as

$$\mathbf{Y} = \mathbf{H}(d, \mathbf{dir})\mathbf{X} \tag{3.4}$$

where $\mathbf{X}$ is the input force spectrum and $\mathbf{Y}$ is the vibration frequency representation (i.e., the spectrum of the acquired signal). The function $\mathbf{H}$ is a frequency response function of the structure [84]. The function is impacted by 1) distance d due to the dispersion effects of the Rayleigh-Lamb waves [128], and 2) the signal propagation direction $\mathbf{dir}$ due to the structural homogeneity difference. In a homogeneous structure, the structural distortion effects at different directions are the same, i.e.,

$$\mathbf{H}(d) = \mathbf{H}(d, \mathbf{dir}).$$

A sensing system deployed in a homogeneous structure has a higher data efficiency because we can use signals propagated from any direction to establish the model $\mathbf{H}$.

AutoQual measures the structural homogeneity of a deployment as the similarity of signals frequency response $\mathbf{Y}(d, \mathbf{dir})$ from different propagation directions dir with controlled (same) input force spectrum $\mathbf{X}$ and sensor-excitation distance d. *AutoQual* calculates this signal similarity as,

$$AF_3 = \max \mathrm{xcorr}(\mathbf{Y}(d, \mathbf{dir1}), \mathbf{Y}(d, \mathbf{dir2})),$$

where signals $\mathbf{Y}(d, \mathbf{dir1})$ and $\mathbf{Y}(d, \mathbf{dir2})$ are normalized by their signal energy. AF_3 reflects the directional distortion of signal propagation media.

Structure-Sensor Coupling Model

The structure-sensor coupling condition varies over different deployments, especially for different surface materials. When there is a tight coupling between the

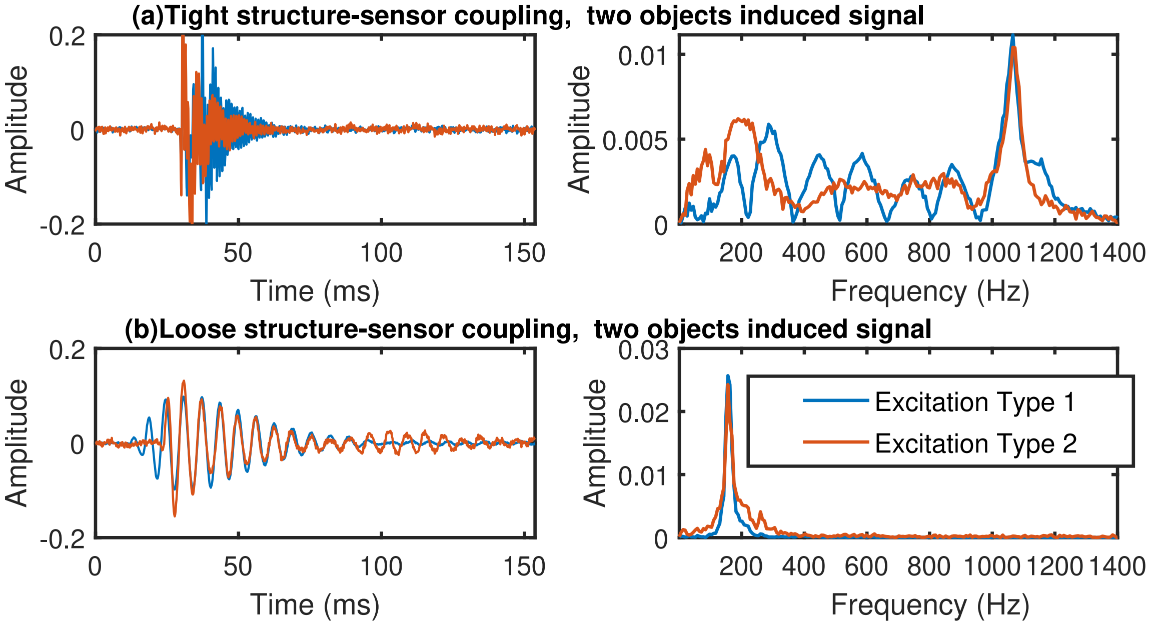


Figure 3.2: Frequency response over different structure-sensor coupling conditions. (a) and (b) show the vibration signals (normalized by signal energy) induced by different types excitation with different structural-sensor coupling condition. The excitation frequency responses in tight structural-sensor coupling environment are more distinguishable.

structure and the sensor, the sensor captures distinct structural response frequencies caused by different excitation. However, for surfaces with a loose structure-sensor coupling, the excitation's frequency response is dominated by the interaction between the structure and the sensor showing a less distinguishable waveform for different excitation. Figure 3.2 demonstrates an example of structural response frequency over different structure-sensor coupling conditions. The signals shown in Figure 3.2 (a) are acquired with a tight structure-sensor coupling condition, and the signals in (b) are acquired with a loose structure-sensor coupling. A tight coupling condition ensures that the acquired signal is dominated by the effects of structure-excitation interaction. These types of signals are less impacted by effects induced by structure-sensor interaction. Therefore, a sensing system with a tight structure-sensor coupling is more informative of different excitation sources, giving better sensing quality in event classification task.

AutoQual measures the excitation vibration signal's frequency component distribution as $x\%$ energy concentration bandwidth (ECB).

$$ECB(x) = \arg\min_{b} \left(\sum_{f=b_0}^{b_0+b} PSD_{norm}(f) \geq x\% \right) \tag{3.5}$$

where the PSD_{norm} is the power spectral density (PSD) normalized by signal energy. b_0 is traversed from 0 to $Fs/2 - b$. Fs is the sampling rate. Specifically, *AutoQual* measures $AF_4 = ECB(75)$, $AF_5 = ECB(50)$ to reflect structural modes that can be excited by the excitation.

3.3.2 Mobile Sensing Context Extraction for Vibration Sensing Assessment

To reduce human efforts for collecting data for assessing environmental impacts [151], *AutoQual* utilizes occupants' mobility and shared-context between mobile- and vibration-sensing to achieve autonomous assessment.

The properties that allow pedestrian's footstep-induced signals to substitute manually generated standard excitation are twofold: 1) the same person's footfall generates consistent excitation, which is equivalent to standard excitation for assessment in prior work, 2) when a pedestrian passing by, their footfall locations change, generating signals with different sensor-excitation distances needed for AF measurements. The challenges of using ambient human footstep-induced vibration signals to measure AFs are twofold: 1) there are ambient non-footstep events that can be detected and should not be used for AF measurement, 2) the human footfall, compared to the standard excitation, are less consistent (e.g., left/right foot difference) and more complicated (e.g., toe push-off induces damped free vibrations).

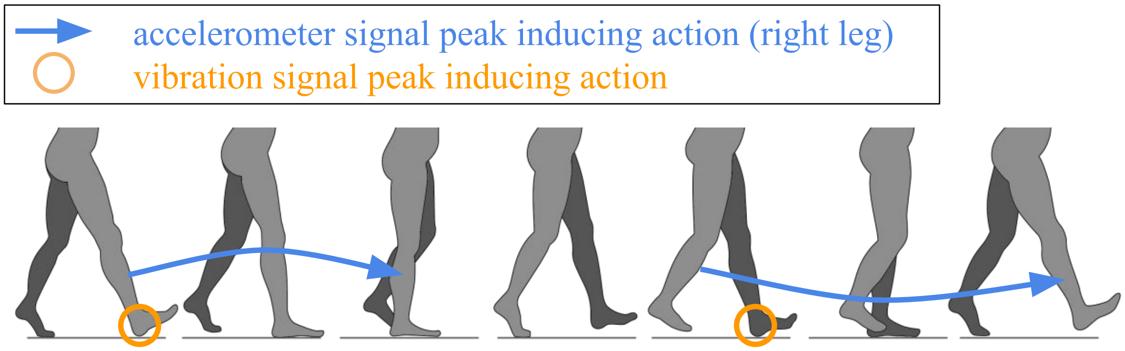

accelerometer signal peak inducing action (right leg)
vibration signal peak inducing action

Figure 3.3: Cross-modal gait-cycle-based timing association.

Gait-based Mobile-Infrastructural Sensing Signal Temporal Association

When people walk in their natural form, their gait/footstep can be detected by mobile devices carried on them [53] through the Inertial Measurement Unit (IMU). *AutoQual* leverages co-located mobile device to detect the timing of the footfall, which has been explored for position and gait cycle detection [50, 88, 89]. In this work, we use a three-axis accelerometer sensor on the calf to measure the pedestrian's footstep timing. *AutoQual* conducts peak detection on the accelerometer signal by finding local maxima as footfall timestamp. This is done on the axis with the highest signal amplitude. For vibration signals, *AutoQual* conducts the anomaly detection to extract the excitation event signal segments [96].

The challenge to associate the footstep timing between the structural vibration and accelerometer data is that the detectable gait patterns have a *cycle offset* between these two sensing modalities, illustrated in Figure 3.3. The *initial strike* of the investigated leg (marked in the orange circle) would induce a footstep-induced vibration signal event at time T_{vib}. However, the calf motion that induces the cycle phase shift from the *initial strike* to the *mid stance* would induce a detectable peak in the acceleration measurement at time T_{acc}. As a result, the gait phase detected by the IMU is slightly (∼150ms) lagged compared to that of the vibration signal.

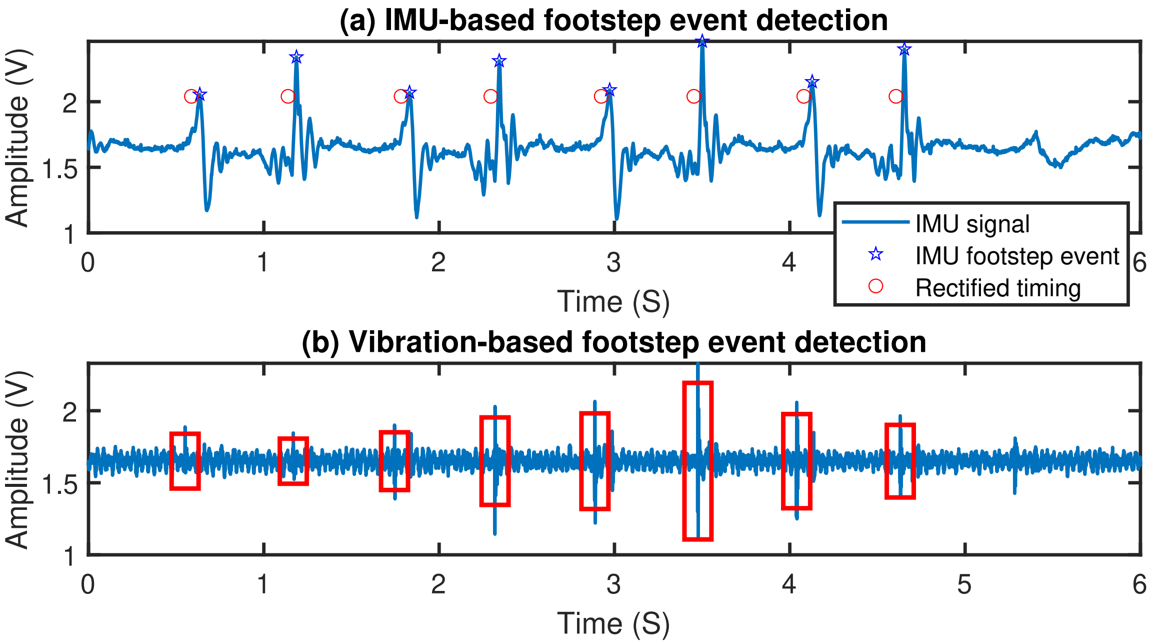


Figure 3.4: Multi-modal gait-cycle-based timing rectification.

To rectify this gait cycled offset for robust event signal association (e.g., when there are detectable ambient events occur at a similar time of the footstep), *AutoQual* utilizes the average of the gait cycle offset between event timing of two modalities within a trace [1]. *AutoQual* first associates the IMU footstep timing to the vibration footstep event with the closest timing. Given q^{th} detected vibration event, we find the p^{th} IMU event $\arg\min_p |T_{acc}(p) - T_{vib}(q)|$ as a pair. For the Q pairs of associated IMU and vibration event in a trace, *AutoQual* estimates the gait cycle offset as

$$T_{offset} = \frac{1}{Q}\sum_{i=1}^{Q} T_{acc}(q) - T_{vib}(p).$$

Then, *AutoQual* rectifies the T_{acc} as $T_{acc} - T_{offset}$. Figure 3.4 shows an example of the two modalities event timing respectively as well as the rectified timing.

[1] We define a trace as a series of footstep events when a person walks by.

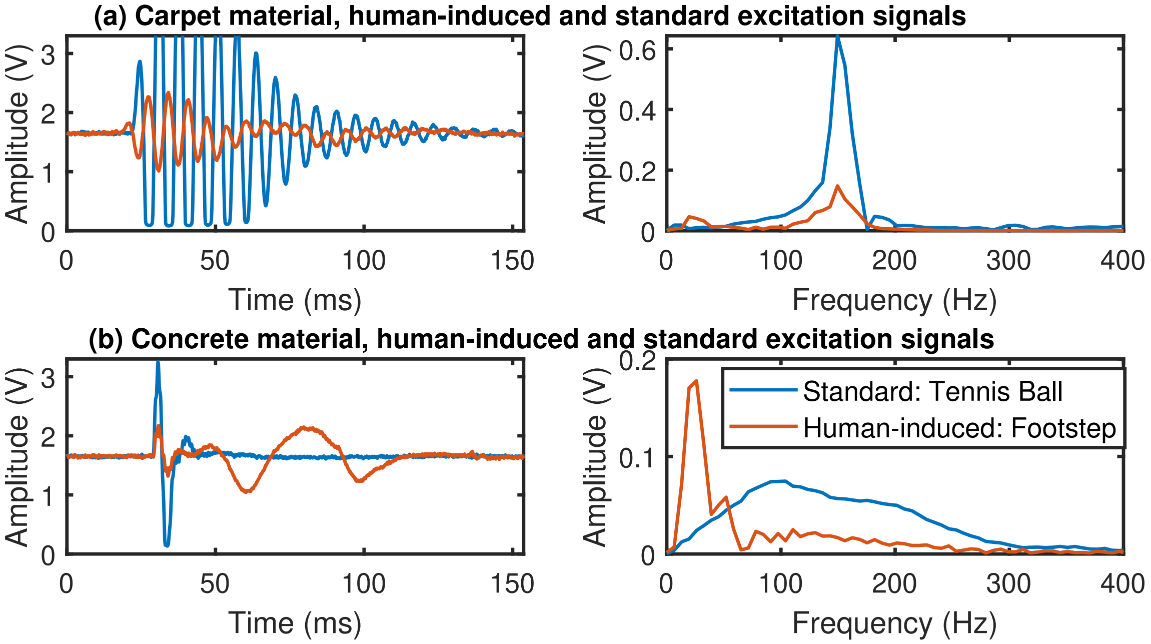


Figure 3.5: An example of the difference between human-induced signal and standard excitation signal. (a) and (b) show the time domain and frequency domain (normalized) of two signals that induced by human and standard excitation in two environments. The surface of environment (a) is a soft carpet and the surface of environment (b) is concrete which has a higher stiffness. We can observe a clear damped free vibration with the footstep-induced signal in (b).

Signal Processing for AF Measurements

Unlike the standard excitation, using footstep-induced vibration signals is challenging because of the following reasons 1) the randomness of human behavior makes the excitation less consistent compared to that of the standard excitation. 2) the footstep-sensor distance d used in the attenuation model is unknown. 3) the footstep-induced vibration signal has a heavy damped free vibration due to the toe push-off motion, which may directly impact the signal's frequency characteristics for the structure-sensor coupling model calculation. As shown in Figure 3.5 (b), the footstep-induced vibration signal contains strong damped free vibration components, i.e., the tail of the signal [118], which makes the ECB measurements ambiguous for different structure-sensor coupling conditions.

To address the first challenge, *AutoQual* leverages the associated information from IMU data to identify corresponding footstep-induced vibration signals to calculate the AFs. To address the second challenge, we consider an average foot stride length *strLen* of people approximately 2 ft at a normal walking speed for men and women between 20-39 years old [93]. When a pedestrian passes by, *AutoQual* first conducts the event detection with an anomaly detection algorithm [96]. Then the system selects the detected footstep signal with the highest signal energy as the reference footstep and assign a fixed reference distance d_{ref}. For a footstep that is k step away from the reference footstep, we consider its distance to the target sensor d_k can be calculated as

$$d_k = \sqrt{d_{ref}^2 + (k * strLen)^2} \tag{3.6}$$

We use the estimated distances and the footstep induced vibration signals to further calculate the attenuation model factors as discussed in Section 3.3.1. For the associated events of a sequence of footstep-induced vibration signals, *AutoQual* calculates the AF values and estimates the TSQ score. For the deployment that no events are associated, *AutoQual* reports as not a valid assessment. To address the third challenge, we only extract the onset of the footstep signal to avoid the push-off induced damped free vibration when calculating the structure-sensor coupling model factors.

3.3.3 Task-Oriented Sensing Quality (TSQ) Scoring Model

Different sensing tasks are sensitive to different environmental factors and have distinct requirements. To ensure the fairness of system sensing quality comparison, a Task-oriented sensing quality score is necessary. Compared to conventional once-for-all models, the TSQ score provides a fine-grained representation of how AFs influence the system performance under different tasks. It further enables a comprehensive understanding of sensing system qualities under different task scenarios and thus achieves a more precise prediction of system performance on a set of tasks. To calculate the TSQ score, we first model the projection from the individual raw AF measurements to a saturation function ranged between 0 and 1 (Section 3.3.3). Then we integrate all AF's saturation functions as the TSQ score (Section 3.3.3).

Saturation Function for Individual AF

The impacts of AFs on the sensing task performance are often constrained. For example, AF_1 is calculated from the signal decay rate $AF_1 = -40\alpha$. For the same excitation, the higher the decay rate (the lower the AF_1 value), the lower the event detection rate. When AF_1 is lower than a threshold, there will be no footstep-induced vibration signal detected, and the sensing task performance's decreasing trend flattens. Similarly, when the value of AF_1 is higher than a threshold, the detection rate increase in the target sensing range is no longer obvious. In both cases, we consider the impact of AF_1 is saturated, which we model with a saturation function – sigmoid.

$$S(\mathrm{AF}) = \frac{1}{1 + e^{(a\mathrm{AF}+b)}} \tag{3.7}$$

As shown in E.q. 3.7, we quantify the impact of individual AF on system performance in the range of 0 to 1. In addition, we further constraint the saturation range by setting an upper and lower threshold (T_u and T_l) on the AF measurement where $S(T_u) = 0.9$ and $S(T_l) = 0.1$.

Task-oriented Integration Model

A holistic model is built to integrate all AFs and provide an overall assessment (TSQ score) of the deployment. Different AFs may have different impacts on the sensing task performance. For example, the attenuation model AFs may play more important roles for the event detection task than the structure-sensor coupling model.

The integration model should configure a unique weight for each AF. On the other hand, the same AF has different impacts for different sensing tasks. For example, the attenuation model AFs may have a more essential influence on event detection than event classification. **The weight of each AF should vary for different tasks**.

To enable a uniform and quantified sensing quality assessment criteria, we build a task-oriented integration model to calculate the TSQ score. In real-world scenarios, there are often a limited number of sensing system deployments as well as noisy AF measurements to train the integration model. Therefore, a model with high capacity and complexity is easy to result in over-fitting with limited training sets and results in low performance of final TSQ score prediction. Meanwhile, the interpretability of the integration model is important for people to understand how different AFs impact the system performance and response accordingly. Therefore, we choose linear regression model [133] on top of the saturation function for individual AFs to build the integration model. Given a deployment environment j, the TSQ score of j is modeled as

$$TSQ^j = \sum_{i=1}^{M} w_i S(\mathrm{AF_i^j}) + \mathrm{c}, \tag{3.8}$$

where $\mathrm{AF_i^j}$ is the ith AF of the deployment environment j, w_i is the weight of $\mathrm{AF_i^j}$, and the c is the estimated constant offset. M is the number of all AFs.

The goal is to predict the real-world sensing system performance. The least squares approach is employed here to estimate the parameters. The objective function for obtaining an optimal TSQ score model is as follows:

$$\min_{\boldsymbol{w}, c, \boldsymbol{T_l}, \boldsymbol{T_u}} \sum_{j=1}^{N} \left[\left(\sum_{i=1}^{M} w_i S(\mathrm{AF_i^j}) + \mathrm{c} \right) - p^j \right]^2 \tag{3.9}$$

where p^j is the sensing task performance (classification accuracy, detection rate, etc.) at deployment j. $\boldsymbol{T_l}$ and $\boldsymbol{T_u}$ is a vector consists of the lower and upper thresholds of the saturation function. N is the number of training deployments.

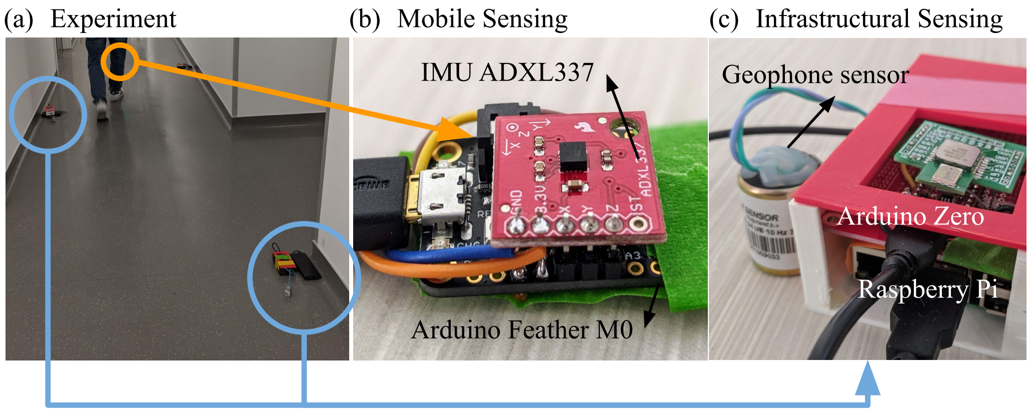


Figure 3.6: Hardware and an example deployment.

We use the gradient descent algorithm [18] to find the optimal solution of the objective function 3.9. To avoid the local minimum, we randomly select the initial point multiple times and select the best loss solution as the final optimal solution. Two stopping conditions are set to exit the iteration, 1) when the difference between two consecutive objectives is less than a given threshold δ_o; 2) when the Euclidean distance between two consecutive solutions is less than a given threshold δ_s. The selection of δ_o and δ_s are described in Section 3.4.2.

3.4 Experiments

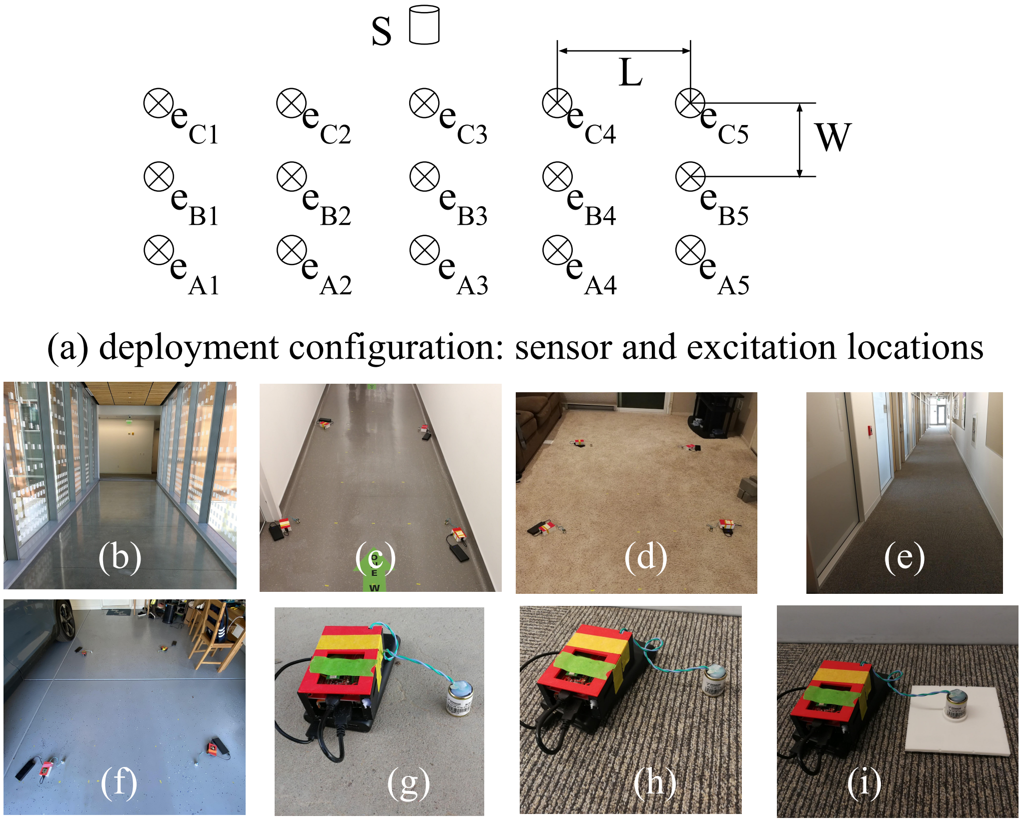
Experiments are conducted to evaluate the introduced assessment framework, including two sensing tasks over 48 deployments at 11 different environments. We obtain sensing task datasets (Section 3.4.1) and sensing quality assessment datasets (Section 3.4.2) separately.

AutoQual consists of two sensing modalities, a structural vibration sensing system and a wearable system. Figure 3.6 shows the two system hardware implementations – (b) demonstrates the mobile sensing system with an three-axis accelerometer and (c) illustrates the infrastructural sensing system with structural vibration sensor geophone SM-24 [63]. The mobile sensing node consists of the accelerometer module ADXL 337 [7] to capture the footfall motion, and an Arduino Feather M0 board [1] to digitize the signal and store the data. We sample the three axes of the accelerometer at 800 Hz to ensure the high temporal resolution of the signals. The infrastructural sensing node consists of an operational amplifier LVM385 [122] and an Arduino board with a 32-bit ARM Cortex-M0+ processor and a 10-12 bit ADC module [2]. We sample the geophone at a sampling rate of 6500 Hz.

We conduct experiments at 11 different environments varying over floor materials (e.g., concrete, carpet, epoxy), and layout (e.g. hallways, rooms, and open space). Table 3.1 lists the characteristics of each deployment environment. We setup in total 48 deployments at 11 environments over different locations. Figure 3.7(a) shows an example deployment configuration, where the sensor and relative locations of excitation used for assessment are marked. Since the narrowest hallway we deploy has a width of 5 ft, we set the relative location of the excitation in the environment with $W = 1$ ft. In addition, because people's strike length is approximately 2 ft, we set $L = 2$ ft. We define a set of excitations across the sensing area as one **trace**, e.g., the set of five excitations at e_{A1}, e_{A2},...,e_{A5} is considered as one trace referred to as trace e_A.

3.4.1 Sensing Task Dataset Collection and Processing

We collected floor vibration signals when people walk through sensing areas to establish the sensing task dataset. For each participant, we collect her/his footstep-induced floor vibration with two types of shoes – sneakers (soft-soled) and hiking shoes (hard-soled) over every deployment. In each deployment, the task dataset contains four types of footstep-induced excitation from two participants with two shoe types each. For each scenario (distinguished by person, shoe type, environment, sensor), a three-minute vibration signal is collected and in total we collect more than



(a) deployment configuration: sensor and excitation locations

Figure 3.7: Experiment deployment environments and configurations. (a) Deployment configuration. The cylinders represent sensor locations, the circles with crosses indicate excitation locations. (b-i) Deployment environment examples.

9 hours (576 minutes) worth of floor vibration data. To demonstrate the task-oriented assessment, we evaluate *AutoQual* over two common sensing task representatives: event detection (Section 3.4.1) and event classification (Section 3.4.1).

Event Detection

The event detection task is to detect footstep-induced vibration signal events from ambient floor vibration. We apply an anomaly detection based algorithm to achieve

Table 3.1: Deployment environment details.

Environment ID	Floor Material	Layout
1	Epoxy	Rooms
2	Carpet (hard)	Hallway
3	Epoxy	Bridge (connect two second floor)
4	Epoxy	Hallway
5	Epoxy	Hallway (with beam)
6	Carpet (soft)	Rooms
7	Carpet (soft)	Rooms (with mounting device)
8	Concrete	Rooms
9	Concrete	Open space
10	Epoxy	Rooms
11	Epoxy	Open space

the event detection [96]. The anomaly detection is done on the sliding window (with a window size of 0.415 seconds and an offset size of 0.015 seconds). For each noise window, we calculate the signal energy and build a Gaussian noise model $\mathcal{N}(\mu, \sigma^2)$. For each testing window, the algorithm compares the sliding windows' signal energy to the Gaussian noise model. If the signal energy is higher than a threshold (here set to 8σ empirically), we consider the testing window contains a detected event.

We acquire the label for each footstep occurred during the experiment with the accelerometer on the mobile device (placed on the calf). A peak detection algorithm [144] is used to detect the timestamp of a footfall with the time series data of accelerometer. In total, more than 27,000 individual footstep events are collected in the task dataset. F1 score is used as the evaluation metric to measure the event detection performance. For each labeled timestamp T_{gt}, if an event is detected within the time window $T_{gt} \pm Th_{gt}$, we consider it as a true positive (TP), otherwise

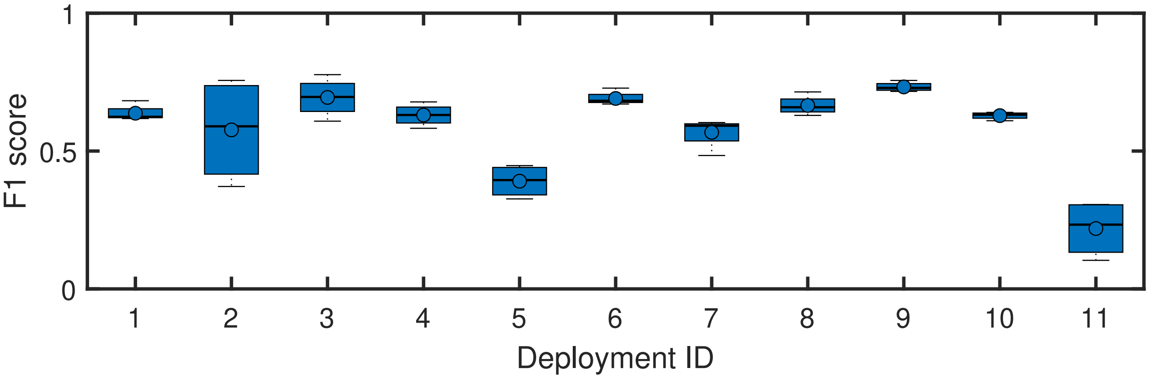


Figure 3.8: F1 scores of event detection in 11 different deployment environments. The x-axis is the deployment environment ID. The y-axis is the event detection F1 score. The event detection task shows varying performances in different environments.

a false negative (FN). If an event is detected outside the time window of $T_{gt} \pm Th_{gt}$, we consider it a false positive (FP). The threshold Th_{gt} is empirically set as 0.0769 seconds (500 samples). The precision and recall are calculated as

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

The F1 score is a function of precision and recall

$$F1\ score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Figure 3.8 shows the detection F1 score in different deployment environments. The F1 score of event detection over 48 deployments vary from 0.044 to 0.777, with an average of 0.600 and a standard deviation of 0.186.

Event Classification

The event classification task is a multi-class classification task to recognize four types of footsteps, i.e., two participants' footstep-induced vibration wearing two types of shoes. In order to mitigate the impact of the classification algorithm and enable a fair comparison between deployments, we apply eight commonly used classification algorithms, including Support Vector Machine (SVM) with Linear Kernel (LSVM) and RBF Kernel (RSVM), Gaussian Naive Bayes (GNB), Random Forest (RF), Extra Trees (ET), K-nearest Neighbors (k-NN), Logistic Regression (LR), and XG-Boost [17, 25, 26, 42, 46, 48, 73, 76, 108]. The classifier inputs are the frequency components ranging from 10 Hz to 400 Hz of a 1000-sample window containing the investigated signal, which is a 391 dimension feature. The model is trained with bootstrap aggregating. The training and evaluation is conducted in a nested 5-fold cross-validation fashion. In each fold, we randomly sample 80% data from the training set each time, and aggregate the learned models to obtain the final model hyperparameters. The ground truth of event classification is the type of excitation distinguished by the pedestrian identities and their footwear types, which are labeled manually. The average F1-score is employed as the metric to evaluate the performance of multi-class classification tasks [115]. Figure 3.9 shows the classification F1 score of each algorithm in different environments. Different algorithms have different performances using the same dataset. To eliminate the performance variation induced by the classification algorithm, the **event classification** performance at each deployment is calculated as the mean value of the averaging F-1 scores of eight different classification algorithms. In this way, we can eliminate the assessment bias introduced by algorithms. The event classification performance over 48 deployments varies from 0.302 to 0.780, with a mean value of 0.555 and a standard deviation of 0.125.

3.4.2 Sensing Quality Assessment

The goal of the sensing quality assessment is to provide a fair and comparable measurement of the deployment environments regardless of the applications/tasks.

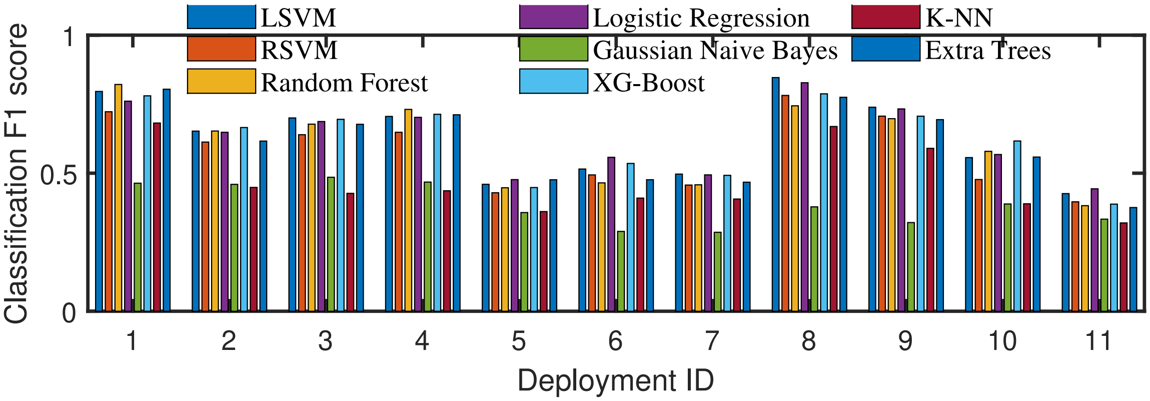


Figure 3.9: F1 scores of event classification using eight learning algorithms in 11 different deployment environments. The x-axis is the deployment environment ID. The y-axis is the event classification F1 score. Different color bars represent results using different algorithms. The event classification task shows varying performance in different environments.

As a result, the data used for assessment should be independent of the sensing tasks, so we collected the assessment dataset separately from the sensing task dataset. In order to compare to prior work on manual sensing signal quality assessment [151], we collect both the manual assessment dataset with standard excitation (tennis ball drops from a consistent height) and *AutoQual* assessment dataset with ambient footstep excitation.

The locations of excitation are marked as circles with crosses in Figure 3.7(a). For the manual assessment dataset, we collected 10 excitation samples at each marked excitation location. For human excitation, we collected 18 traces (people walking by) including six along e_A, e_B, and e_C respectively.

AutoQual **Auto-Assessment via Ambient Footstep Excitation**

In order to calculate a stable TSQ score of a deployment, we use 18 traces from the sensing quality assessment dataset at each deployment to calculate the AFs.

The AFs are calculated as follows. 1) Attenuation Model (AF_1, AF_2): *AutoQual* selects the footstep-induced vibration signal with the highest SNR in each trace as the reference footstep and set the parameter $d_{ref} = 2ft$ in E.q. 3.6 to estimate each signal's sensor-excitation distance. Then we use E.q. 3.3 to conduct linear regression. We remove outliers beyond 2σ of the model expectation, where the σ is the standard deviation of the model variance. 2) Structural Homogeneity Model AF_3: *AutoQual* selects pairs of footstep-induced signal that have the same distance to the reference footstep signal to calculate AF_3. We report the maximum of all the calculated values as the measurement of AF_3 in this deployment. 3) Structure-sensor Coupling Model (AF_4, AF_5): *AutoQual* uses the reference footstep signal in each trace to calculate $ECB(x)$, where $x = 75\%, 50\%$ to calculate the 3rd quartile, 2nd quartile of the energy distribution. We use the average value over multiple traces as the final AF_4, AF_5 of this deployment.

For each training deployment, we calculate its AFs and acquire the corresponding ground-truth system performance p. Given N training deployments, the data-driven model is calculated by solving the objective function in E.q. 3.9. We apply the gradient descent algorithm to obtain the optimal TSQ score model parameters. The stopping conditions are setup as: $\delta_o = 10^{-6}$ and $\delta_s = 10^{-10}$. To avoid overfitting, we set the maximum number of iterations to 200. The gradient descent algorithm exits the iteration if one of the stopping conditions is satisfied. We randomly select 50 initial points and report the solution with the lowest loss as the final solution.

Baseline 1 Manual Assessment v.s. Auto Assessment

To understand the efficiency of *AutoQual* with mobile-enabled assessment, we conduct a manual assessment with standard excitation as a baseline. In each deployed environment, we drop a tennis ball at the marked 15 locations in Figure 3.7(a) as the standard excitation to calculate the AFs. 1) Attenuation Model AFs: we use the mean SNR of all excitations at each locations as the SNR of each location, and then calculate the distance between each location and the vibration sensor. Finally, SNR and distance values to calculate AF_1 and AF_2. 2) Structural Homogeneity Model AF: we select pairwise signals with the same distance to the sensor in each trace and calculate their waveform similarity. The highest value from all traces in a deployment is used as AF_3. 3) Structure-sensor Coupling Model AFs: we utilizes signals with the shortest excitation-sensor distance in a trace ($\mathrm{e_{A3}}$, $\mathrm{e_{B3}}$, or $\mathrm{e_{C3}}$ in Figure 3.7(a)) to calculate AF_4 and AF_5.

Baseline 2 AF Saturation Function: Sigmoid v.s. Piecewise

*AutoQual*utilizes data-driven approaches to determine the saturation bound of each AF, i.e., when an AF value is out of the saturation bound, its impact on the system performance is saturated. To demonstrate the efficiency of the data-driven approach adopted by *AutoQual*, we compare *AutoQual*'s saturation function to a piecewise function as the baseline

$$f(\mathrm{AF}) = \begin{cases} 1 & T_{max} < \mathrm{AF} \\ (\mathrm{AF} - T_{min})/(T_{max} - T_{min}) & T_{min} \leq \mathrm{AF} \leq T_{max} \\ 0 & \mathrm{AF} < T_{min} \end{cases} . \tag{3.10}$$

where T_{min} and T_{max} are the minimum and maximum values of each AF from the training deployments.

Baseline 3 Assessment Factor Selection

To understand the effectiveness of the AFs for representing the sensing quality, we predict the system performance using the same framework with two traditional signal quality metrics: SNR [142] and SSIM [24, 33]. For the SNR, we calculate the mean SNR value of all footstep-induced signals in a deployment. For the SSIM, we calculate the mean value of all signal pairs with equal distance to the sensor in a deployment. The SSIM value of two signals x and y is calculated as

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_x\sigma_y + C_2)(\sigma_{xy} + C3)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)(\sigma_x\sigma_y + C3)},$$

where μ_x is the mean value and σ_x is the standard deviation of the signal x, respectively. σ_{xy} is the the covariance of x and y; $C_1 = (K_1 L)^2$, $C_2 = (K_2 L)^2$, and $C_3 = C_2/2$ are constants that ensure stability when the denominator approaches zero; L is the dynamic range of the digitized signal values. In our implementation, we select SSIM parameters $K_1 = 0.01, K_2 = 0.02$ based on recommended parameter ranges [24]. The dynamic range of the signal values is $L = 2^{10}$ for a 10-bits ADC module. We project the SNR and SSIM values in the training set to the range of 0 and 1 linearly, and predict the TSQ score as $TSQ_{SNR} = f(SNR)$ and $TSQ_{SSIM} = f(SSIM)$, where $f(\cdot)$ is defined in E.q. 3.10, T_{min} and T_{max} are the minimum and maximum of SNR or SSIM values from the training set.

Baseline 4 Integration Methods: Task Oriented v.s. Equal Weight

AutoQual provides a task-oriented integration model by selecting different weights for different AFs. We consider a non-task-oriented and non-data-driven integration model (Equal Weight) as the baseline to evaluate the efficiency of our integration model. For each deployment, we use the mean value of all AFs to calculate the TSQ score, which is calculated as

$$AF_{avg} = \frac{1}{N}\sum_{i=1}^{N} AF_i.$$

The TSQ score is calculated as

$$TSQ = f(AF_{avg}),$$

where $f(\cdot)$ is defined in E.q. 3.10, T_{min} and T_{max} are the minimum and maximum of AF_{avg} from the training set.

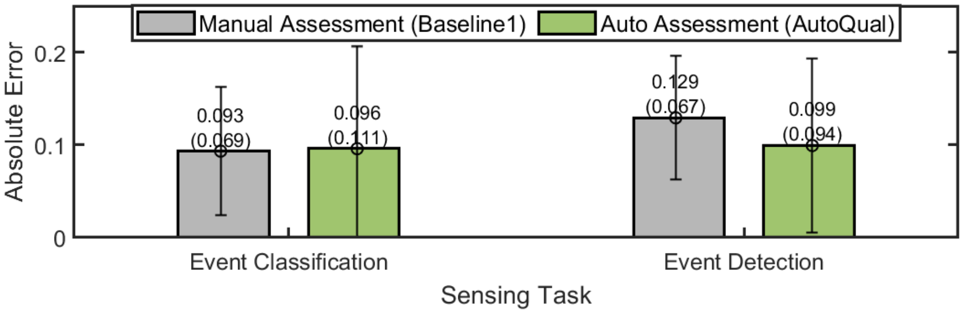


Figure 3.10: *AutoQual* evaluation with manual assessment and mobile-enabled auto assessment. The mobile-enabled auto assessment method has a similar performance comparing with the manual assessment. The x-axis is the sensing task. The y-axis is the absolute error between the predicted sensing quality score and the task performance of the testing deployment.

3.5 *AutoQual* Evaluation

We conduct a detailed performance analysis on *AutoQual* over baselines from multiple perspectives in this section. Implementation details and multiple types of baseline methods are first presented. We further compare our framework with baseline methods and characterize relationships between assessment factors, optimal weights of each factor, and final assessment performance. Finally, a robustness test is conducted to evaluate how the performance varies with multiple human objects.

Since the TSQ score represents the prediction of the system performance at a given deployment, we calculate the absolute error between the predicted TSQ score and the task performance of each testing deployment to evaluate the performance of *AutoQual*. For each analysis, we randomly sample some deployments as the training deployments from the aforementioned datasets (Section 3.4) to train the TSQ scoring model, and use all the other deployments as the testing deployments to evaluate the model performance. The final result of each analysis is the evaluation of 2000 testing deployments. We repeat the random sample of the training and testing multiple times until the testing deployments number is not less than 2000. The default training size is 24.

3.5.1 *AutoQual* Comparisons to Manual Assessment (Baseline 1)

We compare *AutoQual*'s performance to the manual assessment approach using standard excitation as described in Section 3.4.2. Figure 3.10 shows these two approaches performance with absolute error of task performance prediction. The grey bars present manual assessment results using standard excitation, and the green bars present our *AutoQual* results with ambient footstep excitation. The manual assessment achieves an average absolute error of 0.093 and 0.129 for the two sensing tasks – event classification and event detection. *AutoQual* achieves a similar performance compared to the manual assessment using standard excitation, where it achieves an average absolute error of 0.096 and 0.099 for the two sensing tasks.

In order to further illustrate the advantage of *AutoQual*, we compare the time duration needed for manual assessment and using *AutoQual* to achieve autonomous assessment with ambient footstep excitation. For the manual assessment using standard excitation, the user needs to mark the excitation locations on the floor (Figure 3.7 (a)) and generate the standard excitation (e.g., drop a tennis ball) at each location. We time the manual deployment assessment procedure five times, which reaches an average of 622 seconds ($\sim$ 10 minutes) with a standard deviation of 23.4 seconds. On the other hand, *AutoQual* does not require additional time for assessment data collection. In summary, *AutoQual* demonstrates the robustness of the assessment accuracy over different sensing tasks and the efficiency in terms of time consumption.

3.5.2 AF Saturation Function: Sigmoid v.s. Piecewise (Baseline 2)

In order to verify the effectiveness of the non-linear projection for *AutoQual*, we compare our non-linear model to the Min-Max baseline as discussed in Section 3.4.2. Figure 3.11 demonstrates the absolute error of our approach (green bars) and the baseline method (grey bars). The x-axis is the training size for event classification and event detection assessment shown in (a) and (b), respectively. *AutoQual* achieves a lower error rate than the baseline method, especially when the number of training deployments is less than 10. With eight training deployments, the mean absolute error decreased 35% and 38% for event classification and event detection, respectively. When the training size increased from 8 to 24, the mean absolute error of our method decreased less than 0.031, which is 3X better than the decrease of the baseline method (0.081).

To further understand the performance of these two approaches with limited training deployments, we also analyze their error distribution. Figure 3.12 (a) and (b) show the error distribution of the two methods with eight training deployments for event classification and event detection task performance prediction, respectively. We observe that our approach has fewer outliers and presents a more concentrated distribution, which leads to a lower absolute error. The performance is improved due to two reasons. First, when the training size is small, the maximum and minimum measurements of each AF in the training set may not be sufficient to represent the model. Second, the baseline approach assumes that the relationship between each AF and system performance is linear in all ranges, which is not true as discussed in Section 3.3.3.

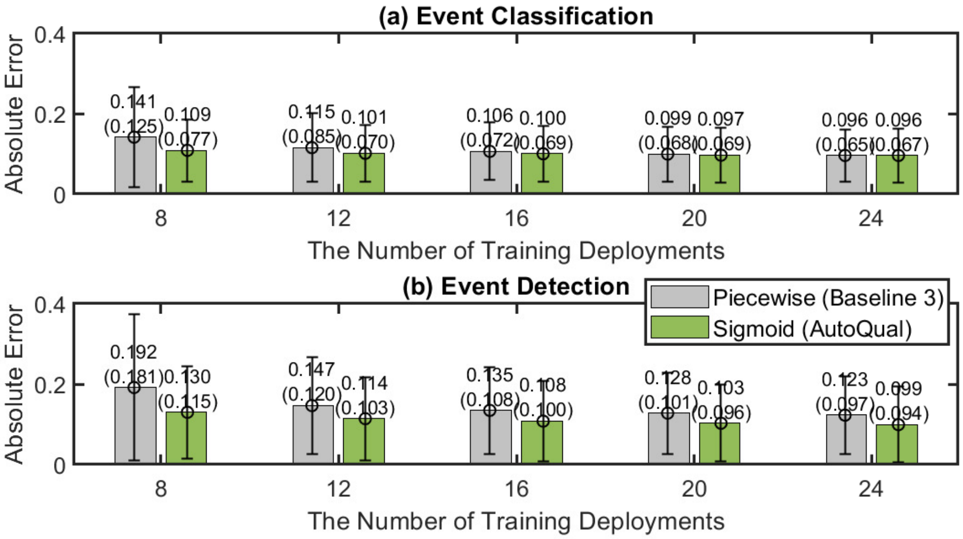


Figure 3.11: AF impact quantification analysis. The x-axis is the number of training deployments that we used to select the parameters. The y-axis is the absolute error between the TSQ score and task performance of the testing deployments. (a) and (b) show the performance of two methods for event classification and event detection, respectively.

3.5.3 Assessment Factor Selections (Baseline 3)

We compare the AFs used in *AutoQual* to the traditional signal quality metrics SNR and SSIM as discussed in Section 3.4.2. Figure 3.13 shows the absolute error of the system performance prediction over different assessment metrics. The results for baseline metrics are marked in grey bars, and the results for *AutoQual* are marked in green bars. We observe that for the event classification, the mean absolute error for *AutoQual* is 0.096, and the corresponding mean absolute error for event detection task is 0.99, which is 2X better than those of the SNR (0.27 and 0.22) and SSIM (0.207 and 0.25). In addition, when using *AutoQual* to predict the system performance, the standard deviation of the prediction is also lower than those of the baselines. In

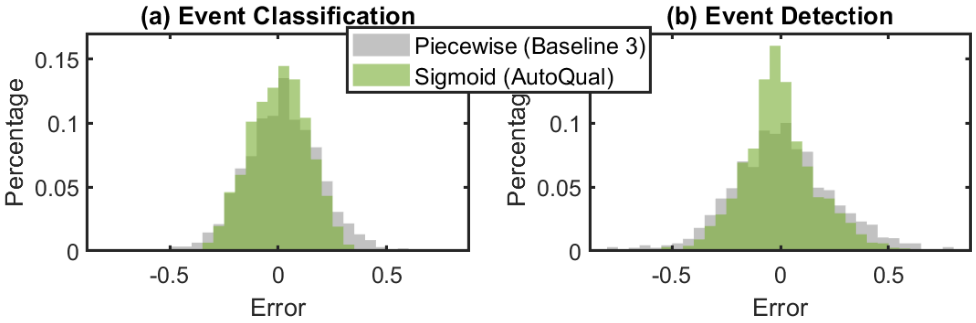


Figure 3.12: Analysis on error distribution of our method and the baseline method when only use eight deployments as the training deployment. (a) and (b) show the result for event classification and event detection, respectively. The x-axis is the error between the TSQ score and the task performance of the testing deployment. The y-axis is the percentage of the error range in all errors. The error distribution of our method is more concentrated than the baseline method for in (a) and (b), which identifies that our method has a better performance than the baseline method.

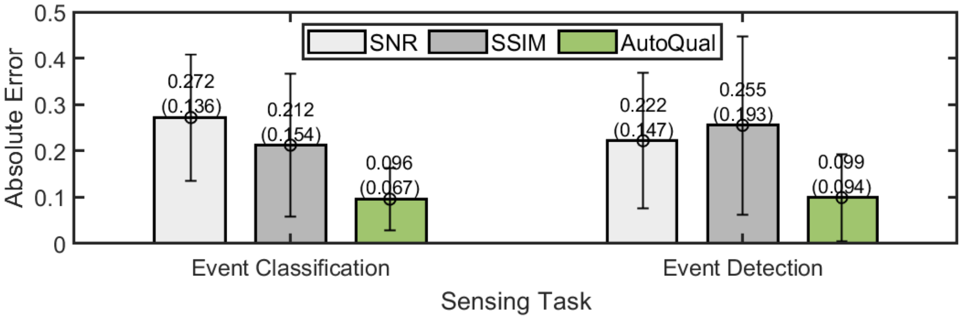


Figure 3.13: Comparing with SNR and SSIM baseline methods, the absolute error of *AutoQual* is at least 2X lower than them for the two sensing tasks. The x-axis is the sensing task. The y-axis is the absolute error between the predicted sensing quality score and the task performance of the testing deployment.

summary, *AutoQual* achieves a more accurate and stable sensing quality estimation compared to the baseline methods.

We further investigate the impact of each AF on the sensing task. To do so, we use single AF to implement *AutoQual*. Figure 3.14 shows the absolute error between the TSQ score and the task performance. We observe that, for the event classification task, AF_2 and AF_3 demonstrate lower error compared to other AFs. For the event detection task, AF_2 achieves the best performance, which indicates that AF_2 has the strongest impact. AF_2 measures the ratio of the initial amplitude of the signal over the amplitude of background noise. The event detection algorithm is based on the signal to noise ratio (SNR). As a result, it is more sensitive to the SNR related AFs, i.e., AF_1 and AF_2. AF_3 measures the directional distortion of signal propagation media. A higher AF_3 value indicates that the same type of signals collected from different directions have less variation. This may lead to less structure variance induced data distribution shift within each class for the event classification task. In addition, when compared between tasks, the event classification task is more sensitive to AF_1, AF_4 and AF_5 compared to the event detection task. This is because the event detection task does not require the signal to contain rich frequency components that reflect the structural mode excitation for fine-grained characterization.

3.5.4 Integration Methods: Task-Oriented v.s. Equal Weight (Baseline 4)

In order to evaluate the efficiency of the AFs integration model, we compare our task-oriented data-driven integration model to the Equal Weight baseline as introduced in Section 3.4.2. Figure 3.15 shows the absolute error of our approach (green bars) and the baseline method (grey bars). For event classification assessment, the mean absolute error of our method (0.092) is near 3X lower than the baseline method (0.306); For event detection assessment, the mean absolute error of our method (0.102) is near 2.5X lower than the baseline method (0.254). On the other hand, the standard deviation of our method also lower than the baseline method for

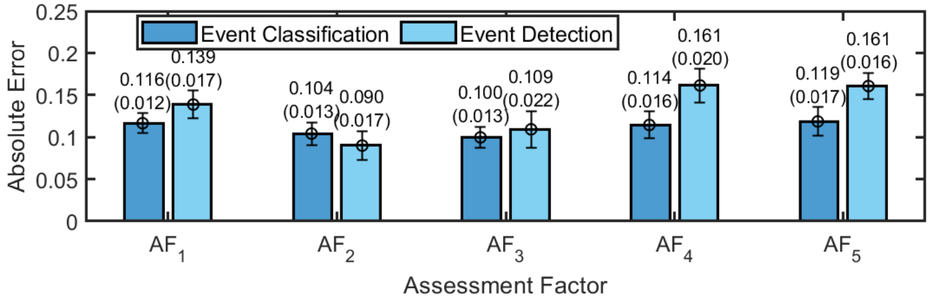


Figure 3.14: Single AF model analysis. The x-axis is the AF and the y-axis is the absolute error between the predicted TSQ score and sensing task F1 score. The models are trained with 24 deployments. AF_2 and AF_3 have a stronger impact for event classification than other AFs. AF_2 has a stronger impact for event detection than other AFs.

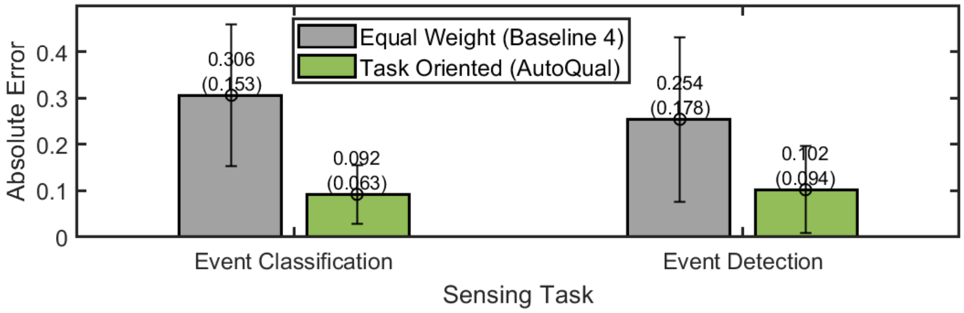


Figure 3.15: Multiple AFs integration model analysis. Our task-oriented data-driven model achieves at least 2.5X lower absolute error than the baseline method. The x-axis is the sensing tasks. The y-axis is the absolute error between the TSQ score and the sensing task performance over the testing deployments.

both sensing tasks, which indicates that our approach is more stable. In summary, the task-oriented data-driven integration model is more stable and at least 2.5X better than the baseline method.

We further analyze the selected parameters of our integration model for these

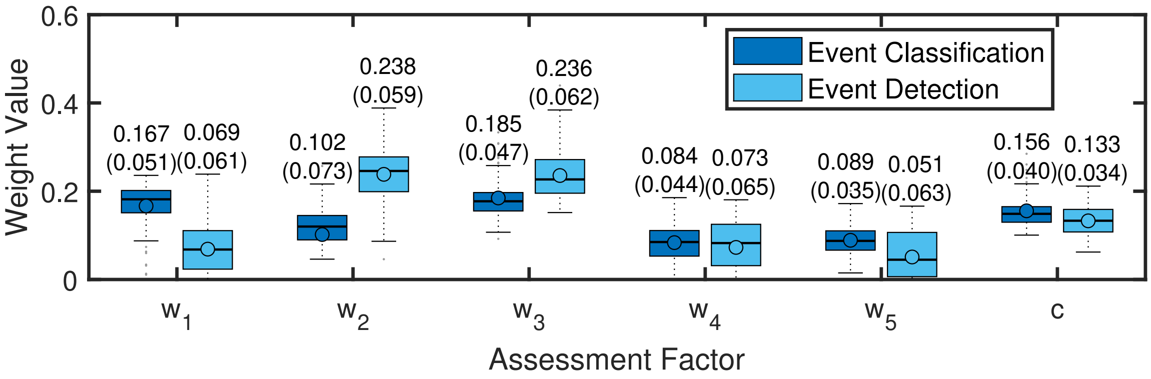


Figure 3.16: Analysis on task-oriented integration model variation over different tasks. The x-axis is the AFs and the y-axis is the weight values. The models are trained with 24 deployments. The weight values of the same AF varies over different tasks, and the different AFs have varying weights for the same task.

two sensing tasks. The weight values of the integration model (Eq. 3.8) reflect the contribution of different AFs to the assessment of sensing tasks. Figure 3.16 presents the weight values of the integration model with a training size of 24. We observe that different AFs have different weights for the same task. This indicates that the deployment characteristics (AF) affect the system performance. On the other hand, most AFs have different weights for different tasks. This indicates that different tasks favor distinct criteria for assessing sensing quality.

3.5.5 Assessment Robustness over Different Users

The human variation experiment evaluates the robustness of *AutoQual* when it utilizes multiple persons' footstep excitation for assessment. Prior work on structural vibration-based human sensing indicates that different people's footstep excitation have distinct characteristics [99, 101]. To validate the robustness of *AutoQual* using different people's footstep excitation, we collect two persons' footstep excitation in each deployment for auto-assessment. Figure 3.17 shows the assessment performance

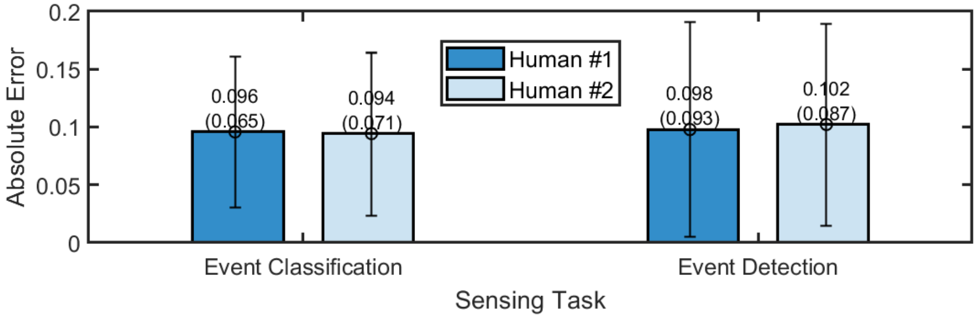


Figure 3.17: Analysis on human variation. The x-axis is the sensing task, and the y-axis is the absolute error between the TSQ score and sensing task performance. *AutoQual* demonstrate similar performance when the assessment is done on Human #1 and Human #2, which indicates robustness over different users.

of *AutoQual* for event classification and event detection with two people's footstep excitation. We observe that *AutoQual* has a similar performance when using different people's footstep excitation as the assessment excitation. The mean absolute error of event classification assessment is 0.0096 and 0.0094 from human #1 and human #2, respectively. The mean absolute error of event detection assessment is 0.0098 and 0.0102 from human #1 and human #2, respectively. The difference of mean absolute error between two people's assessment performance is 0.002 and 0.004 for event classification and event detection, respectively. In summary, *AutoQual* is robust when using different people's footstep excitation to achieve the assessment.

3.6 Discussion

In this section, we further discuss the limitation of this work and possible future directions.

3.6.1 Robust Gait Cycle Detection and Gait-based Temporal Association

In this work, we rely on a mobile device on the calf for gait-based temporal association between different sensing modalities. However, mobile devices or wearables may come in various form factors, e.g., smartwatch [44], earable [105], on-cloth/limbs [88]. Because these form factors are placed on different body parts, they may be sensitive to capturing different phases of gait cycles. In our future work, we plan to explore schemes that allow automatic calibration of this gait cycle phase offset between mobile devices/wearables and infrastructural sensors for robust temporal association. Furthermore, for some form factors, the detecting gait cycle phase may change over different contexts (e.g., smartwatch detecting people walking with/without carrying loads). To ensure robust gait-based temporal association, the system may adapt or selectively conduct temporal association based on the context.

3.6.2 Assessing a Group of Sensors on Collaborative Tasks

In this work, we only investigated applications/sensing tasks done by an individual sensor. However, there are sensing tasks that require multiple sensors to conduct collaborative computation, such as multilateration based localization [39, 86]. We plan to extend the current Assessment Factor (AF) concepts for individual sensors to sensor groups. This can be done by measuring the group factors over multiple devices assigned for the collaborative task. The group factors could be clock synchronization resolution, environment-circuit interference over different locations, etc. The TSQ score model will take both group and individual factors into account for collaborative task's sensing quality assessment.

3.6.3 Integration Model

In this work, we focus on individual AF's modeling, i.e., the saturation function. The individual AF model reflects the specific causes of high/low learning accuracy. However, the linear regression model we adopted is limited in terms of representing the relation and dependency between different AFs. The advantage of our linear integration model is the selected parameters (weight) intuitively reveal the relationship between the AF and the task performance. The limitation of the linear relationship is that it may not be the precise representative for the relationship between individual AF and task performance. When two AFs are dependent, the relationship between the task performance and two AFs is no longer a linear combination, even the relationship between the task performance and the individual AF may be linear. To further improve the robustness of the integrated score, we plan to explore more approaches to estimate the integration models, such as kernel-based regression, general regression neural network, etc. in our future work.

3.7 Chapter Summary

In this Chapter, we present *AutoQual*, a sensing quality assessment framework. We introduce our Task-oriented Sensing Quality (TSQ) assessment model as well as the auto-assessment scheme that utilizes a co-located mobile sensing system to assess the performance of structural vibration sensing systems. Ambient human-induced signals are combined to infer the deployment's environmental and hardware factors that impact system performance. Real-world experiments are conducted with 48 deployments at 11 environments. *AutoQual* achieves less than 0.10 absolute error, which is an 2× improvement to baselines. Our auto-assessment approach predicting multiple tasks' performance at untested deployments shows a ≤ 0.018 absolute error difference to the manual assessment.

Chapter 4

Cyber-Physical Augmentation for Robust Data Acquisition and Scalable Information Inference

4.1 Introduction

Customer-product interaction detection is essential for autonomous retail [112]. Various sensing methods have been proposed, including vision [112, 134], RFID [107, 110], load [112], and pressure (piezo) [57] based approaches. They have various limitations in real deployments. Vision-based methods require line-of-sight (LoS) and might not work if the human body or shelf blocks the view. RFID-based methods often require assigning a tag for each product item, which causes a continuous cost of labor and tags. Load-based and pressure-based methods usually require a dense deployment, which causes a cost increase. To address these limitations, we present a vibration-based single-point sensing solution to allow low-cost monitoring without the requirement of gondola retrofit.

Structural vibration-based sensing methods are emerging and allow various smart

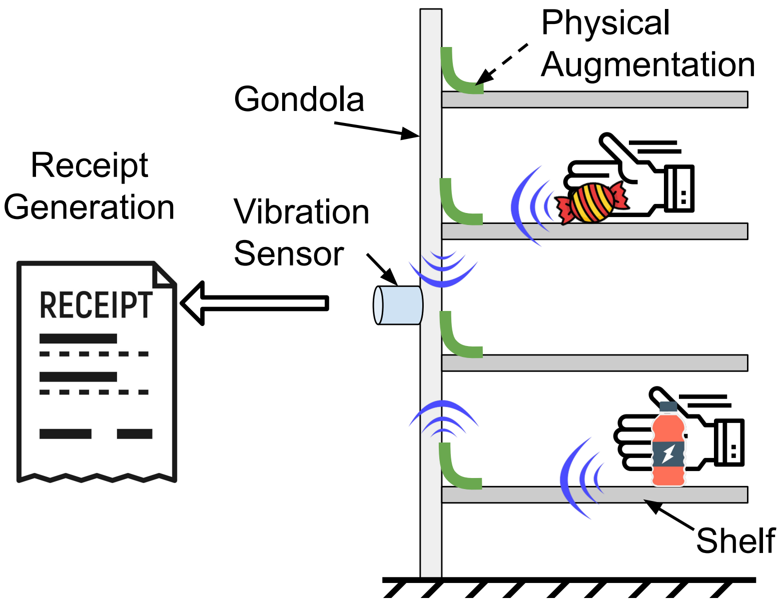


Figure 4.1: Illustration of the vibration-based single-point sensing for customer-product interaction detection. The pickup and/or put down events generate vibrations that propagate through the gondola structure and are captured by the vibration sensor on the back panel of the gondola.

building applications with sparse deployments [100,149]. They can be easily applied to existing infrastructures, such as the retail gondola[1], as shown in Figure 4.1. First, we deploy the vibration sensor on the back panel of the gondola to monitor the vibration of all shelves. The customer-product interaction-induced vibration propagates through the gondola structure and is captured by the sensor. Then, we can infer the interaction information via the vibration signal and enable autonomous retail and/or inventory monitoring. However, this is challenging for traditional vibration-based human sensing systems because of the **extremely low Signal-to-Noise Ratio (SNR) signals** and/**or expensive cost of labeling**. For example, when we use the traditional

[1] A gondola is a freestanding fixture used by retailers, and its shelves are mostly used to display merchandise [4].

vibration-based sensing system to detect events of picking up an energy bar (60 grams), we could only achieve a 5% detection rate (more details in Section 4.4.2). On the other hand, manually labeling the event from the video is impractical for a retail store that contains hundreds and thousands types of items [5].

In this Chapter, we present *CPA*, a cyber-physical augmentation scheme that enhances the traditional vibration sensing system to allow low-cost single-point sensing on each gondola. With one vibration sensor on the gondola's back panel, we are able to monitor customer-product interaction on multiple shelves. *CPA* handles the aforementioned challenge via 1) physical augmentation with a physical arc between each shelf and the back panel to enhance the vibration signal on the back panel; and 2) cyber augmentation with a combination of pseudo labeling and binary classifier to eliminate the need for manual labeling. The **contributions** of this work can be summarized as follows:

- We introduce a vibration-based single-point sensing system for customer-product interaction detection.
- We present *CPA*, a cyber-physical augmentation scheme to enable accurate interaction detection with extremely low SNR signals and without manual labeling.
- We evaluate *CPA* through real-world experiments on a retail gondola.

4.2 Background: Deflection and Vibration

Deflection describes the degree of displacement for the structural element under a load [109]. Figure 4.2(a) shows a cantilever beam end-load deflection model, and the displacement δ at L distance from the support calculates as:

$$\delta = \frac{FL^3}{3EI}, \quad I = mL^2, \tag{4.1}$$

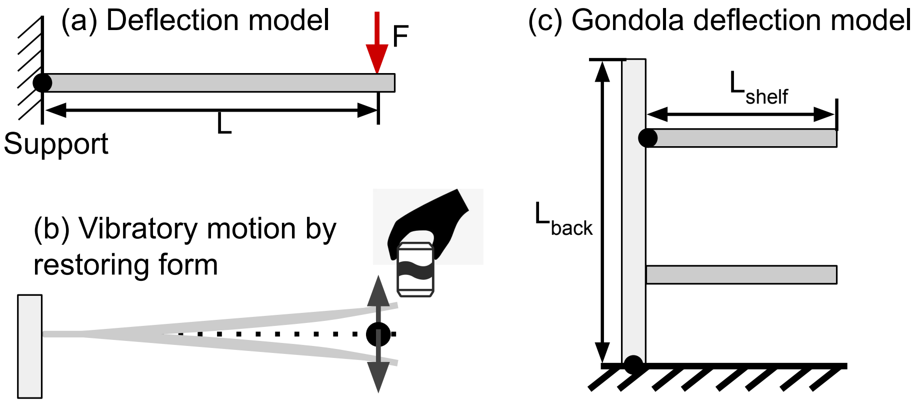

(a) Deflection model
F
L
Support
(b) Vibratory motion by restoring form
(c) Gondola deflection model
L shelf
L back

Figure 4.2: Deflection models. (a) Cantilever end-load deflection model described with the parameter of force F and force-fixed-end distance L. (b) Pick-up event induces vibration by restoring deflection caused by product load. (c) Deflection models applied on the gondola – individual shelf and back wall.

Where I is the moment of inertia [45]. Here, F is the force applied, E is Young's modulus that describes how easily the material would deform and stretch [109], and L is the distance of the beam.

As shown in Figure 4.2(b), when we remove the force F, the cantilever beam will return to its original shape, and the inertia of the beam will cause vibration around that initial location [129]. This vibration is described as a damped oscillation [36], and its amplitude is proportional to the deflection displacement δ. Therefore, the lighter the item is, the less deflection displacement its load will induce, and the vibration induced by interacting with it is of lower amplitude.

Figure 4.2(c) depicts the deflection models to describe the gondola structure: 1) the shelf (L_{shelf}), with the support that connects it to the back panel, and 2) the back panel (L_{back}), with the support that connects it to the ground. The customer-product interaction (pickup and put down) can cause a displacement of the shelf and hence generate vibration waves. When a person puts an item on the shelf, this impact

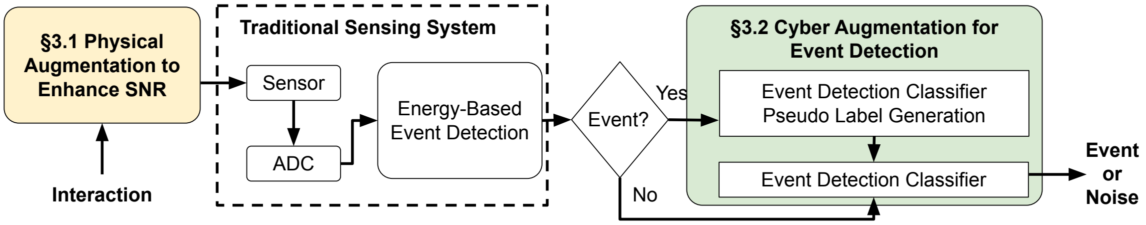

§3.1 Physical Augmentation to Enhance SNR
Interaction
Traditional Sensing System
Sensor
ADC
Energy-Based Event Detection
Event?
Yes
No
§3.2 Cyber Augmentation for Event Detection
Event Detection Classifier Pseudo Label Generation
Event Detection Classifier
Event or Noise

Figure 4.3: System architecture. *CPA* contains physical augmentation to enhance SNR and cyber augmentation for event detection.

deforms the shelf and causes displacement. For a product placed on the shelf, its load induces the deformation of the shelf. When a person picks it up, the restoring force of the shelf induces vibration.

4.3 System Design

We present a vibration-based customer-product interaction detection system for autonomous retails. Figure 4.3 depicts the system overview. When the customer pickup or puts down a product on the shelf, these interactions induce the shelf to vibrate. The vibration propagates through the shelf surface and shelf-gondola connection and is then captured by the sensor placed on the back panel of the gondola shown in Figure 4.1.

However, this interaction-induced vibration signal significantly dissipates at the structure connector between the shelf and the back panel of the gondola, as shown in Figure 4.4(a). Therefore, it is challenging to **capture the interaction-induced vibration on shelves via single-point vibration sensing at the back panel of the gondola**.

In addition, as mentioned in Section 4.2, the pickup interaction induces an extremely low amplitude of vibration signal, which makes event detection with a traditional energy-based approach difficult. On the other hand, the machine learning approach can effectively classify events from noise with time and frequency features. However, it would require intensive labeling, which makes it inapplicable for each shelf. Therefore, it is challenging to **classify these interaction events from noise without label**.

We introduce *CPA*, a cyber-physical augmentation scheme, to enhance its detection performance by augmenting both sensing (Section 4.3.1) and learning (Section 4.3.2) perspectives.

4.3.1 Physical Augmentation to Enhance SNR

Vibration sensing signal quality is determined by both the data acquisition hardware and the signal propagation media [148, 150, 151]. Here we focus on the physical augmentation of the signal propagation media. To monitor multiple shelves via single-point vibration sensing, the sensor is placed on the back panel of the gondola.

Figure 4.4(a) shows the vibration wave propagation from the shelf to the back panel of the gondola. The horizontal grey bar depicts the shelf surface, and the vertical light grey bar depicts the back panel of the gondola. We can consider it as a simplified model of a flat rectangular plate with one edge fixed. The interaction-induced vibration propagates along the plate surface and reaches the fixed edge. Instead of deflecting the plate, the vibration dissipates at the connecting edge, since the vibration particle motion is aligned with (instead of perpendicular to) the surface of the back panel – the deflection along the surface of the back panel of the gondola is almost zero.

On the other hand, the back panel of the gondola can be considered as a plate with a fixed edge on the floor. If the vibration is perpendicular to the plate, it will propagate along the plate and be captured by the vibration sensor. We present a thin

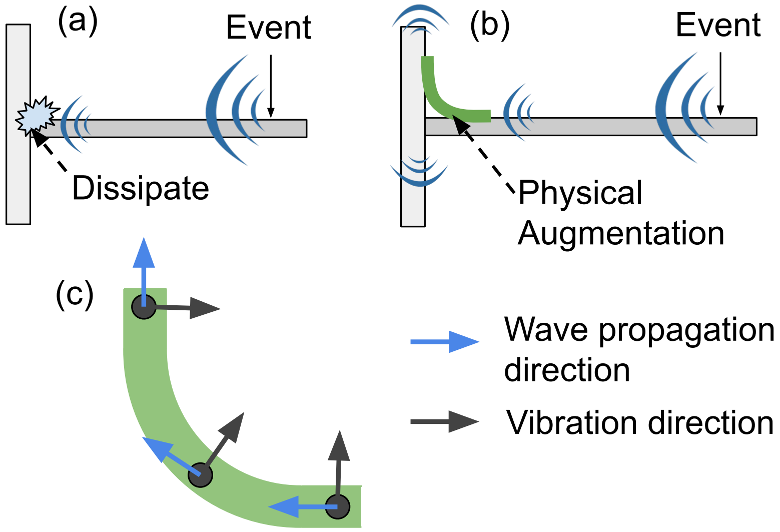


Figure 4.4: Physical augmentation. (a) depicts the wave propagation on the original gondola structure, where the shelf vibration dissipates at the connector to the back panel. (b) shows the wave propagation on the gondola with a physical augmentation structure, where the wave is guided to propagate on the back panel. (c) shows the physical augmentation structure and the wave propagation direction changing over the arc.

arc structure that can align the vibration direction on the shelf and the back panel. We adopt a 90-degree arc architecture with a radius of 25mm and a thickness of 2mm. Figure 4.4(c) depicts the wave propagation along the arc structure, where the blue arrows represent the wave propagation direction, and the black arrows represent the particle motion direction. When the vibration wave propagates through the arc architecture, the vibration direction changes with the arc direction [67,139]. Therefore, our 90° arc architecture can change the vibration direction for 90°, allowing the back panel of the gondola to be used as an efficient vibration signal propagation media for sensing.

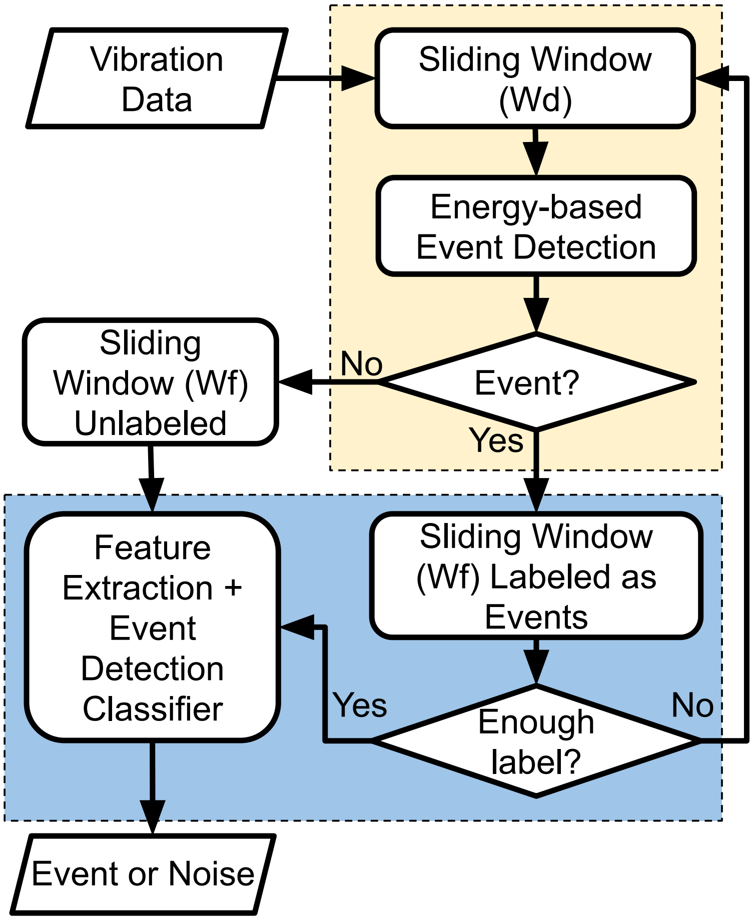

Vibration Data
Sliding Window (Wd)
Energy-based Event Detection
Event?
No
Yes
Sliding Window (Wf) Unlabeled
Feature Extraction + Event Detection Classifier
Sliding Window (Wf) Labeled as Events
Enough label?
Yes
No
Event or Noise

Figure 4.5: Flowchart of the cyber augmentation for event detection. The cyber augmentation contains two main parts: event pseudo label generation (yellow box) and event detection classifier (blue box).

4.3.2 Cyber Augmentation for Event Detection

CPA leverages the target data's heterogeneity and similarity to achieve labeling free accurate event detection. As discussed in Section 4.2, items of different weights induce vibration signals of different amplitude; hence these events' SNR varies. As a result, the interaction with heavier items can be detected with a higher detection rate using signal energy-based approaches [78]. These detected events, which are mostly of heavy items, are then used as pseudo labels of events to train an **event detection classifier**. Figure 4.5 depicts the flowchart of our cyber augmentation components, which contains two main parts: event pseudo label generation (yellow box) and event detection classifier (blue box).

Event Detection Classifier Pseudo Label Generation. We apply the traditional energy-based event detection [100] to generate pesudo-positive labels for the event detection classifier. First, a sliding window (with a window size W_d) is applied to the vibration signals, and the windowed signal energy is calculated. A Gaussian model (μ,σ) is then established based on the windowed noise signal energy. Next, for an incoming sliding window, its energy is compared to the Gaussian model, and if the signal energy is higher than a threshold, e.g., $\mu + 3\sigma$, the window is considered as part of an event. When there are consecutive windows detected as events, we output them as an interaction event IE. These detected IE, which has a low false positive rate and high false negative rate, are labeled as the positive pseudo label. On the other hand, the ambient noise signals acquired when the store is idle or closed, i.e. no customer in the area is labeled as the negative pseudo label.

Event Detection Classifier. When enough pseudo labels are collected, we use them and the corresponding signal segments to train a binary classifier for event detection on each gondola. Because part of the target signals (pickup of small items) often have extremely low SNR values and short duration – therefore are often failed to be detected through the energy-based method. We apply a sliding window with a small window size W_f to effectively capture their features. For each windowed signal, we apply the continuous wavelet transform (CWT) to calculate the component in multiple scales and take the energy of each scale as a feature. Then, the Min-Max normalized [103] energy array of all scales compose the features of this window data.

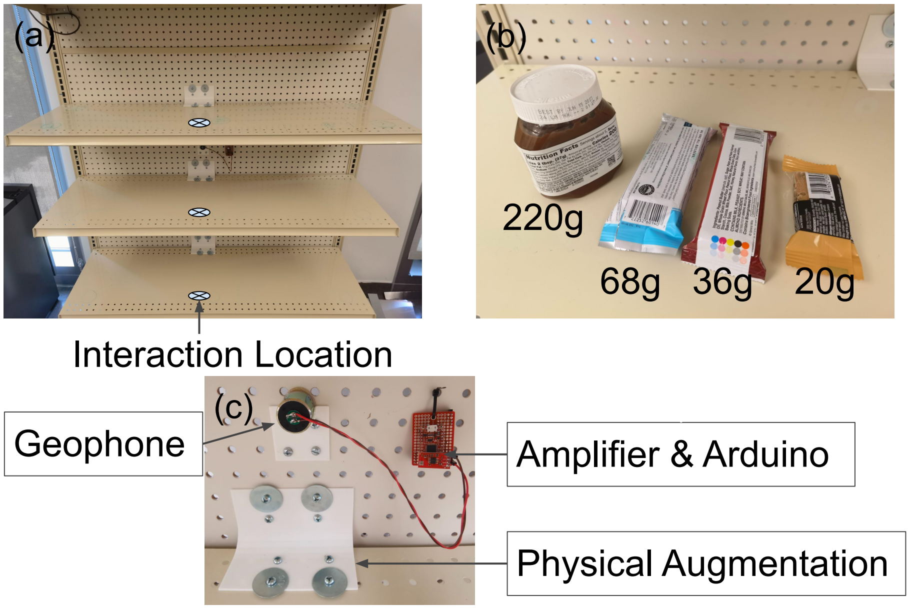


Figure 4.6: Hardware and experiment setup.

Extracted features and pseudo labels are taken as inputs to train a binary classifier to determine if the unlabeled windows are events or noises.

Multiple classifiers can be applied here, such as neural networks, decision trees, and support vector machine (SVM). Exploring these algorithms, however, is not the focus of this work. Here, we adopt SVM due to its generally good performance over a small amount of training data. In addition, we choose the per-gondola design to allow robustness and flexibility over different configurations (e.g., materials, total load, and items) of gondolas in stores.

4.4 Evaluation

We conduct real-world experiments to evaluate *CPA* with multiple product pickup and put down events.

4.4.1 Data Collection and Experiment Setup

Implementation

We 3D print our physical augmentation structure with Acrylonitrile Butadiene Styrene (ABS) material. The width, radius, and thickness of the physical augmentation are 120mm, 25mm, and 2mm, respectively. We use screws to mount the physical augmentation between each shelf and gondola, as shown in Figure 4.6(c). The geophone sensor is horizontally installed on the back panel of the gondola with a 3D-printed mount. To enhance the connection between the sensor and the back panel, we apply resin glue between them. We estimate the cost to retrofit one gondola with our prototype to be lower than \$100, including \$1 for 3D printing, \$4 for OpAmp, \$20 for Arduino, \$40 for Raspberry Pi, \$10 for geophone, which can drop to \$20 or lower if massive produced.

Data Collection

We select four products, including a plastic bottle of chocolate (220g) and three snack bars with different weights (68g, 36g, 20g), to evaluate the event detection performance. We consider them to fall into two categories based on mass: medium ($\geq 100g$) and light ($< 100g$). For each product, we put it down and pick it up 10 times on the centre of each shelf marked in Figure 4.6(a) and record the video as the ground truth of the event timestamp. The materials of the shelf and back panel are steel and cardboard.

Baseline Methods and Evaluation Metric

We compare *CPA* to three baselines: only apply physical augmentation to enhance SNR (Only PA); only apply cyber augmentation for event detection (Only CA); without physical and cyber augmentation (No Augm.). We take the F1 score as the metric to evaluate the performance of event detection. If the timestamp of an event (from the video data) falls into the groundtruth period, we consider this detected event is a true positive case (TP). Otherwise, this detected event counts as a false positive case (FP). If no event is detected at the groundtruth period, we count this as a false negative case (FN). The F1 score is calculated as: $F1 = 2TP/(2TP + FP + FN)$.

Algorithm Implementation

We empirically set the threshold and sliding window length for the energy-based event detection algorithm to be $\mu + 10\sigma$ and 50ms. We set the window length for the event binary classifier to be 20ms. We select the kernel-based SVM model as the classifier to train and detect events. We use the medium-weight product's interaction for autonomous pseudo label generation. We report the detection performance on all products as the overall evaluation of *CPA* and baselines.

4.4.2 Result and Analysis

We first depict the vibration data in Figure 4.7 with and without physical augmentation to demonstrate the challenge of the low SNR problem and the efficiency of our physical augmentation solution. The signal is generated by putting down and picking up a 68g energy bar. Figure 4.7(a)(c) show time domain signal, and (b)(d) show frequency domain signals. The red rectangular boxes mark the time periods where put down events occur, and the black boxes mark pickup events. When the physical augmentation is applied, as shown in Figure 4.7(c)(d), we observe a higher signal energy for both pickup and put down events. Especially for the pickup event, when the physical augmentation is used, both time and frequency domain signals demonstrate distinguishable characteristics compared to the noise. Without the physical augmentation, we do not observe noticeable signal characteristic changes for pickup events, as shown in Figure 4.7(a)(b).

We report the pickup event's detection performance because 1) it is a more

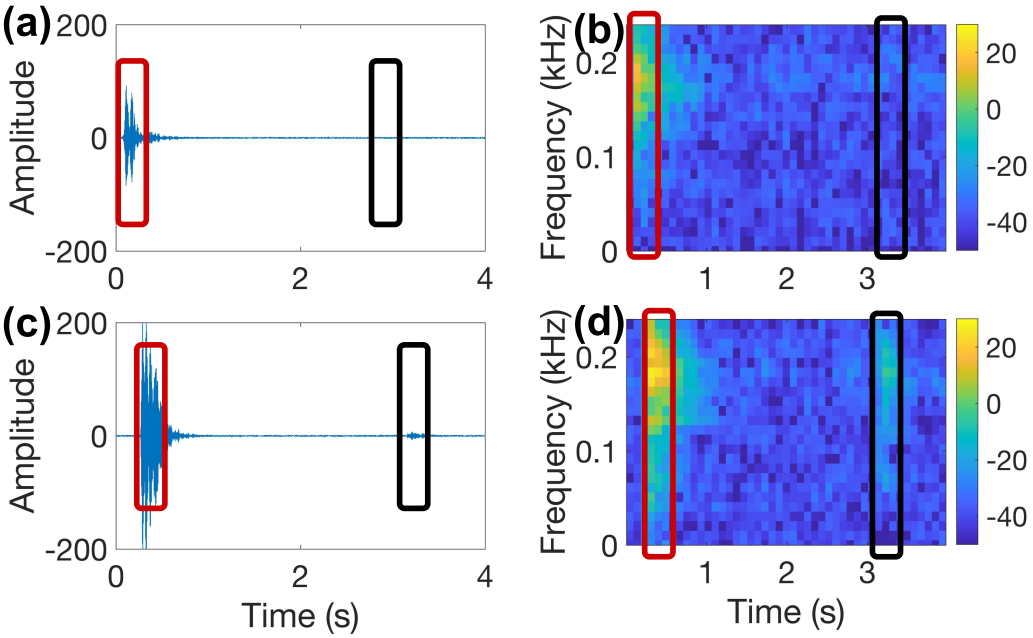


Figure 4.7: An example of an energy bar (68g) put down and pickup events with labels marked in rectangular boxes. (a) and (b) show the time-domain and frequency-domain signals without physical augmentation, respectively. (c) and (d) show signals acquired with our physical augmentation.

challenging problem than detecting put down events, and 2) it is a more frequent event that occurs in retail. Figure 4.8 shows the F1 score of the pickup event detection for each testing product. First, we can observe that the F1 score of *CPA* is higher than no augmentation, especially for the light products, where no augmentation yields a 0 F1 score for 36g and 20g products, and *CPA* still achieves 0.61 and 0.45. Overall, *CPA* achieves a 2.9× improvement compared to the no augmentation baseline across four products' pickup events.

The F1 score of *CPA* for three light-weight products are 0.74, 0.61, and 0.45, respectively, which are the highest in all evaluated methods. Compared with only physical augmentation, *CPA* achieves 0.21 (1.4 ×), 0.19 (1.5 ×), and 0.34 (4.1 ×)

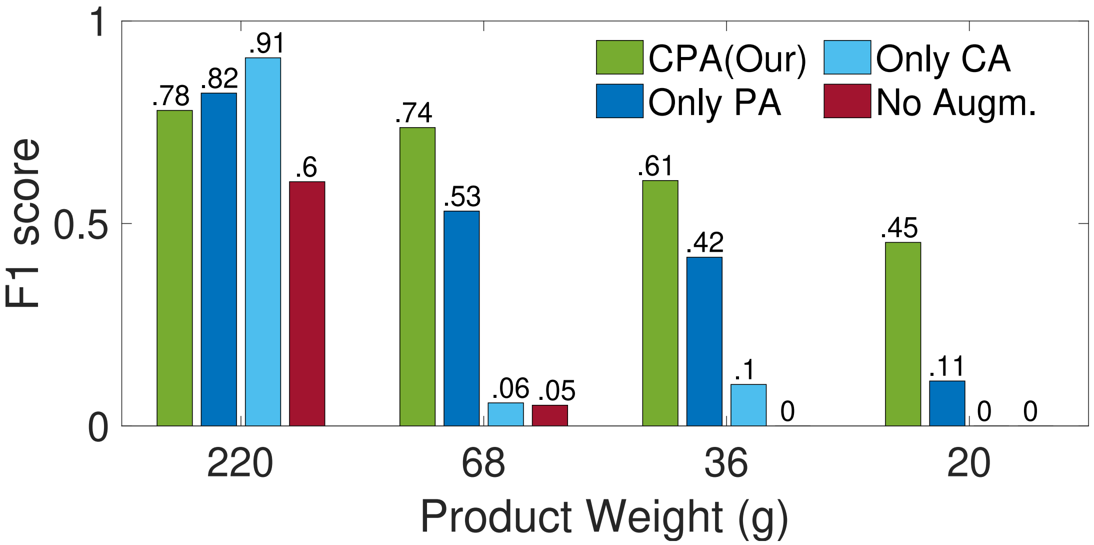


Figure 4.8: The F1-score of *CPA* and baselines for pickup event detection. The colour of the bar represents the augmentation configuration: *CPA* (green), only physical augmentation (dark blue), only cyber augmentation (light blue), and without any augmentation (red).

improvement, respectively. Compared with only cyber augmentation, *CPA* achieves 0.68 (12.9 ×), 0.50 (5.9 ×), and 0.45 (N/A). Physical augmentation demonstrates a significant improvement. The efficiency of cyber augmentation can be further improved by adapting a more representative classifier model or extending the pseudo label set to more products.

4.5 Related Work

Many sensing modalities have been explored for event detection in the autonomous retail setting. The vision is one of the most widely used sensing modalities in autonomous retail stores. Load sensor, RFID, and piezo sensor are the modalities that have also been explored. All these modalities have advantages and disadvantages.

Vision Based Event Detection. The vision-based solutions rely on the video feed from multiple overhead cameras to detect customer activities and pickup and put down events [112]. They require the line of sight, and are susceptible to the occlusion caused by the customer's hand. This makes it hard to detect products that are relatively small in size, such as energy bar. The shelves or the other hardware in stores also cause occlusion in the camera view, especially when cameras are placed on the ceiling. Apart from this, high computation power is required to process the vision data from multiple overhead cameras.

RFID Based Event Detection To overcome the occlusion in the vision, RFID has been explored with RFID tags placed on individual products to track customers' interaction with them [107, 110]. The RFID scanner placed either on the shelf, cart, or the checkout counter scans the item customer interacted with and bills them accordingly. Although this method is highly effective in product identification and inventory monitoring, it brings the continuous labour cost to place RFID tags on the products. Plus, the extra hardware, like shelves or carts with scanners, brings the additional setup cost.

Load Sensor Based Event Detection The load sensors have also been explored for event detection and product recognition in the autonomous retail store setting. In this approach, shelves equipped with multiple load sensors (typically one for each product) are used to detect customers' interaction with different products [112]. This method requires special hardware, which brings the added deployment cost, which makes this solution less desirable for the existing retail stores, as they have to replace all the existing gondola shelves with new smart gondola shelves.

Piezoresistive Smart Textile To convert existing shelves into smart shelves, the piezo sensor base smart shelf-liner are proposed for event detection in an automated retail store setting. In this approach, the smart shelf liner equipped with multiple piezoelectric material-based pressure sensors in the grid configuration are placed on the normal shelf to detect customers' interaction with products [57]. Although this method is highly effective in detecting the events and customers' interaction with the products, this method requires a dense deployment of piezo sensors, making this method expensive and less desirable for retailers.

4.6 Discussion

Cyber-Physical Augmentation Optimization The configuration (material, size, number, location) of the physical augmentation structure may impact the signal augmentation efficiency. For example, vibration wave propagation through different materials varies [66]. When placed at different locations with different densities/numbers, the effective vibration propagation path may vary, which would further impact system performance. In this work, we explore a simple prototype of the arc structure, and the parameters (e.g., material, width, thickness) may not be optimized. Therefore, we plan to explore the optimization of the physical augmentation, including material, size and amount, in the future.

Cyber augmentation also can be further explored by utilizing a more representative classifier model to replace the kernel-based SVM model, such as neural network models. On the other hand, the selected features may also impact the end performance. We plan to explore more data-driven approaches to improve the feature representation for event and noise signal distinction, such as contrastive learning.

Mitigate Impacts from Ambient Vibrations Multiple types of ambient events may induce vibration that can be detected by the vibration sensor, such as customer's footstep, door opening and closing, and the customer touch the gondola. These detected events are false positive cases for pickup/put down item event detection. To distinguish these false positive events, we plan to explore 1) classifier for different event recognition and 2) multi-modal sensing to provide robust labeling for customer-item interaction events.

Prior work on vibration-based human sensing has shown that different human activities would induce signals with distinguishable characteristics [60]. Therefore, it is feasible to recognize signals that are induced by pickup/put down items from those induced by other activities. However, our current cyber augmentation design is not sufficient to achieve pseudo labeling for different types of activities. Since there will be cameras installed in the autonomous retails, we could leverage camera data to achieve fine-grained cross-modal pseudo labeling for cyber augmentation.

System Sensing Range We plan to further study the augmentation design and its impact on the system's sensing range. For example, what is the lower bound of the item that can be detected and how does the augmentation design configuration impact it? In this work, *CPA* achieves 1 and 0.6 recall for 36g and 20g items, respectively. If we focus on the recall performance, we can consider the lower bound weight of the item that can be reliably detected in our experiment is 36 grams. We plan to define the acceptable detection rate based on the target application requirements and conduct fine-grained experiments to explore the lower bound of the system.

Multiple Users Scenarios In the real-world scenario, there might be multiple users interacting with the same gondola at the same time. In the experiment, we observe that the duration of each vibration signal is around 0.1 seconds or less (might varies with gondola material and item), which means it is a low chance that their interaction induced signals will be overlapped. In this work, we check the feasibility of *CPA* when there is no signal overlapping. In the future, we plan to explore methods for signal separation, leveraging the information from other modalities. For example, the cameras in the store can provide complementary information, such as the number of customers interacting with the shame shelf, which could provide prior knowledge for the signal separation.

Robustness to Inventory Changes The total load on the shelf may change over time. The vibration signals of the same event (e.g., pickup a candy bar) may change when the total load and load distribution change [85]. This means the event signals' data distribution may shift with an inferrable physical factor – total load on the shelf. We plan to model this physical phenomenon and design a physical and data-driven approach to mitigate the varying total shelf-load induced data distribution shift. For example, by applying a multi-task learning algorithm to 1) estimate shelf-load and 2) detect events or identify events.

In a real-world scenario, the customer may also return the item to another shelf, which changes the inventory on the shelf. We plan to solve this problem by fusing the vision data, which provides more information on the customer's identity, location, and items. We will also explore how to recognize the item using detected event signals.

4.7 Chapter Summary

In this Chapter, we present a vibration-based single-point sensing system for customer-product interaction detection for autonomous retail stores. To enable the detection of extreme low-SNR signals (light product pickup events), we introduce *CPA*, a cyber-physical augmentation scheme that enhances vibration data SNR with a physical arc structure and provides pseudo labels for event detection classifier leveraging properties of signals generated by different product pickup. We conduct real-world experiments to evaluate the performance of our proposed signal augmentation structure and event detection framework. *CPA* achieves up to 2.9× improvement for the detection of light-weight product pickup events.

Chapter 5

Future Work: Physical Augmentation Enhanced Single-Point Vibration Sensing for Occupant Monitoring

5.1 Introduction

Older adults monitoring is essential for reducing the caregiver's burden and improving older adults' quality of life. Multiple sensing technologies have been explored, including wearable [131], vision [79], radio frequency [68], and vibration [86]. However, they often have limitations when deployed in real-world. For example, wearables require the user to wear them constantly, which may cause discomfort for long-term usage. Vision and audio sensing directly record the image/video or audio data of the user, which rise privacy concerns especially in residential buildings. Radio frequency based methods are non-intrusive solutions, however, they are often susceptible to interference and cluttered environments. Acoustic vibration based

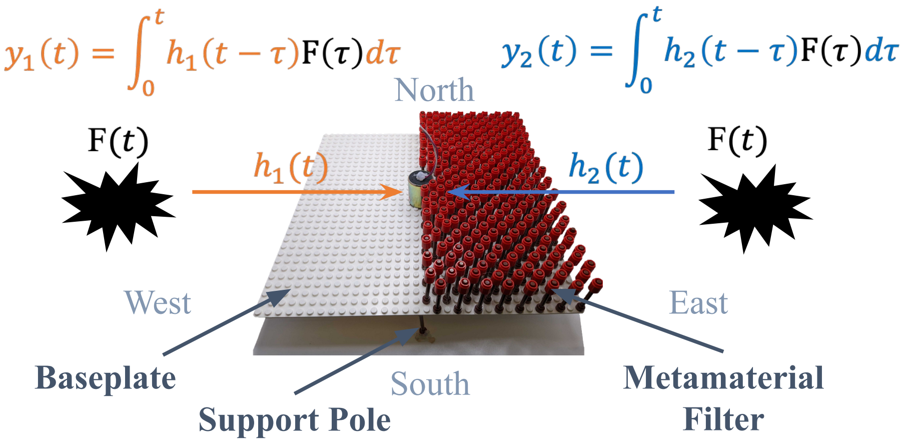

$y_1(t) = \int_0^t h_1(t-\tau)F(\tau)d\tau$
$y_2(t) = \int_0^t h_2(t-\tau)F(\tau)d\tau$
North
F(t)
$h_1(t)$
$h_2(t)$
F(t)
West
East
Baseplate
South
Metamaterial Filter
Support Pole

Figure 5.1: *LEVO* overview. *LEVO* uses configurable bricks to change the structural response function $h(t)$, which embeds a directional signature in the recorded vibration signal $y(t)$.

methods have been explored to acquire fine grained information (e.g., step-level localization, gait balance). However, they require multiple sensor deployment [86] due to their omnidirectional sensing properties.

In this work, we present *LEVO*, a single-point vibration sensing system that leverages configurable LEGO® bricks to embed directional information in mechanical waves for direction sensing. The physical structure of a wave-carrying medium strongly affects the mechanical waves traveling through it, and impacts the frequency distribution of such waves. In particular, the physical structure properties, such as size, shape, and material [39] all affect the wave frequency distribution. By surrounding a sensor with a medium that is drastically different along different directions, the sensor will record waves with different signatures when they come from different directions. As shown in Figure 5.1, *LEVO* leverages this principle to manipulate mechanical waves in a direction-dependent manner – using low cost and re-configurable LEGO® bricks to create a wave-carrying medium that leaves a unique directional signature in the traveling mechanical waves [22]. By carefully designing the structure, it is possible to recognize the wave direction from the vibration data recorded by a single sensor.

5.2 ***LEVO*** Design

For non-intrusive and privacy-preserving occupant tracking, we use a vibration sensor [147, 149] deployed on the structure surface to capture the occupant's activity-induced mechanical waves. To decrease the amount of sensors and computational resources required for directional sensing, we design a physical structure installed between the vibration sensor and the structure surface to embed the directional information into a mechanical wave.

$$y(t) = \int_0^t h(t-\tau)F(\tau)d\tau \tag{5.1}$$

We can model the mechanical wave as a simplified SDOF system as shown in Eq. 5.1 [39], where $F(t)$ is the applied force (i.e., human footstep), $h(t)$ is the impulse response function for the structure, and $y(t)$ is the displacement of the structure. When the structure impulse response h changes, the displacement of the structure y changes with it, even the excitation force F is the same. By building a structure which has a unique impulse response in each direction, the sensor recorded mechanical wave from different directions should be distinguishable. Inspired by this intuition, we designed a physical structure between the sensor and the floor to embed the direction information into a wave.

LEVO contains three parts as shown in Figure 5.1: supporting poles, baseplate and metamaterial filter built with LEGO® bricks [22]. The supporting poles are to separate the base plate from the floor to control the wave propagation paths from the edge to the sensor. The mechanical waves propagates through the designate path and the metamaterial filter. By changing the LEGO® bricks' shape and pattern on the baseplate, we control the metamaterial filter properties [22]. By applying the metamaterial filter at different area of the baseplate, we configure the different metamaterial filter properties on different propagation paths, which embeds the directional information into the mechanical waves before they are captured by the sensor. With these patterns of bricks, *LEVO* represents a passive, low-cost, and re-configurable solution to embedding the direction information in the vibration waveform.

5.3 Preliminary Result

Experiment Setup To verify the design of *LEVO*, we use the excitation from dropping a ball at a consistent location and height as a standard signal. We investigate four directions (shown in Figure 5.1), and collect 20 ball drops induced vibration signal in each direction, which drops at same location from same height. We compare the result of two configurations: with and without metamaterial filter for embedding directional information.

Preliminary Result Figure 5.1 shows the frequency domain (1 Hz to 80 Hz) of the acquired vibration signals of two configurations. We observe that the frequency domain signals of the same excitation propagating to the sensor from different directions showing different level of homogeneity with and without LEGO® bricks built metamaterial filters. Figure 5.1 (a) shows the frequency domain signals without metamaterial filter, where two frequency bands are excited at 22 and 40 Hz for signals from all four directions. While Figure (b) depicts that the frequency domain signal passing metamaterial filter has frequency band excited at 14, 23, and 39 Hz differently over the four directions. This indicates that the LEGO® bricks built metamaterial filter is **capable of embedding directional information into the mechanical waves.** We plan to conduct more studies to control the metamaterial filtering effects

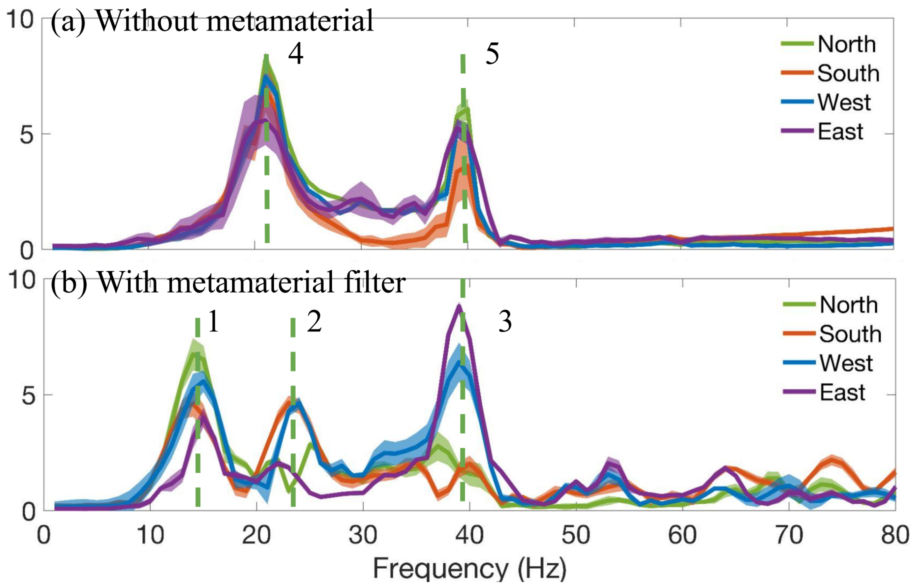


Figure 5.2: The average frequency response and the standard deviation (width of the curve) of the vibration signal without metamaterial filter (a) and with metamaterial filter (b).

on human-induced structural vibration signals for effective single-sensor occupant tracking.

5.4 Chapter Summary

In this Chapter, we introduce *LEVO*, a single-point direction sensing system through a passive, low-cost, and re-configurable metamaterial structure. *LEVO* utilizes the metamaterial filter made of LEGO® bricks to manipulate the frequency components of the mechanical wave, which embeds the direction information into the waveform. We report the preliminary results on direction information embedding and discuss opportunities and future plans.

Bibliography

[1] Adafruit feather m0 bluefruit le. https://www.adafruit.com/product/2995. Accessed: 2021-03-14.

[2] Sparkfun samd21 mini breakout. https://www.sparkfun.com/products/13664. Accessed: 2021-03-14.

[3] Softmax function, May 2019.

[4] Gondola display, 2023.

[5] Our retail divisions, 2023.

[6] Majd Alwan, Prabhu Jude Rajendran, Steve Kell, David Mack, Siddharth Dalal, Matt Wolfe, and Robin Felder. A smart and passive floor-vibration based fall detector for elderly. In *Information and Communication Technologies*, volume 1, pages 1003–1007, 2006.

[7] Analog Devices, Inc. *Small, Low Power, 3-Axis ±3 g Accelerometer*, 2010.

[8] Syed Aziz Shah, Jawad Ahmad, Ahsen Tahir, Fawad Ahmed, Gordon Russell, Syed Yaseen Shah, William J Buchanan, and Qammer H Abbasi. Privacy-preserving non-wearable occupancy monitoring system exploiting wi-fi imaging for next-generation body centric communication. *Micromachines*, 11(4):379, 2020.

[9] François Baccelli, Bartłomiej Błaszczyszyn, et al. Stochastic geometry and wireless networks: Volume ii applications. *Foundations and Trends® in Networking*, 4(1–2):1–312, 2010.

[10] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in

speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022.

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[12] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[13] Dustin Bales, Pablo A Tarazaga, Mary Kasarda, Dhruv Batra, Americo G Woolard, Jeffrey D Poston, and VVN S Malladi. Gender classification of walkers via underfloor accelerometer measurements. *IEEE Internet of Things Journal*, 3(6):1259–1266, 2016.

[14] Tanvi Banerjee and Amit Sheth. Iot quality control for data and application needs. *IEEE Intelligent Systems*, 32(2):68–73, 2017.

[15] Tracy S Barger, Donald E Brown, and Majd Alwan. Health-status monitoring through analysis of behavioral patterns. *IEEE Transactions on systems, man, and cybernetics-part A: systems and humans*, 35(1):22–27, 2005.

[16] Carlo Alberto Boano, Thiemo Voigt, Adam Dunkels, Fredrik Osterlind, Nicolas Tsiftes, Luca Mottola, and Pablo Suarez. Exploiting the lqi variance for rapid channel quality assessment. In *Proceedings of the 2009 International Conference on Information Processing in Sensor Networks*. IEEE Computer Society, 2009.

[17] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[18] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[19] Stewart S Bruce-Low, David Cotterrell, and Gareth E Jones. Heart rate variability during high ambient heat exposure. *Aviation, space, and environmental medicine*, 77(9):915–920, 2006.

[20] Published by Statista Research Department and Feb 22. Average size of households in the u.s. 2021, Feb 2022.

[21] Li Cai and Yangyong Zhu. The challenges of data quality and data quality assessment in the big data era. *Data science journal*, 14, 2015.

[22] Paolo Celli, Behrooz Yousefzadeh, Chiara Daraio, and Stefano Gonella. Bandgap widening by disorder in rainbow metamaterials. *Applied Physics Letters*, 114(9), 2019.

[23] Maojian Chen, Ying Li, Xiong Luo, Weiping Wang, Long Wang, and Wenbing Zhao. A novel human activity recognition scheme for smart health using multilayer extreme learning machine. *IEEE Internet of Things Journal*, 6(2):1410–1418, 2018.

[24] Ming-Jun Chen and Alan C Bovik. Fast structural similarity index algorithm. *Journal of Real-Time Image Processing*, 6(4):281–287, 2011.

[25] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[26] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, and Yuan Tang. Xgboost: extreme gradient boosting. *R package version 0.4-2*, pages 1–4, 2015.

[27] Xinlei Chen, Susu Xu, Xinyu Liu, Xiangxiang Xu, Hae Young Noh, Lin Zhang, and Pei Zhang. Adaptive hybrid model-enabled sensing system (hmss) for mobile fine-grained air pollution estimation. *IEEE Transactions on Mobile Computing*, 2020.

[28] Haneul Choi, Joosang Lee, Yeajin Yi, Hooseung Na, Kyungmo Kang, and Taeyeon Kim. Deep vision-based occupancy counting: Experimental performance evaluation and implementation of ventilation control. *Building and Environment*, 223:109496, 2022.

[29] Haneul Choi, Chai Yoon Um, Kyungmo Kang, Hyungkeun Kim, and Taeyeon Kim. Review of vision-based occupant information sensing systems for occupant-centric control. *Building and Environment*, 203:108064, 2021.

[30] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[31] Seungeun Chung, Jiyoun Lim, Kyoung Ju Noh, Gague Kim, and Hyuntae Jeong. Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning. *Sensors*, 19(7):1716, 2019.

[32] Jose Clemente, Fangyu Li, Maria Valero, and WenZhan Song. Smart seismic sensing for indoor fall detection, location, and notification. *IEEE journal of biomedical and health informatics*, 24(2):524–532, 2019.

[33] WM Clifton, AA Frank, and Ssim-m Freeman. Osteopetrosis (marble bones). *Am. J. Dis. Child.*, 56(Nov.):1020, 1938.

[34] PyTorch Contributors. Conv1d.

[35] PyTorch Contributors. Reducelronplateau.

[36] Philip R Dahl. Solid friction damping of mechanical vibrations. *AIAA journal*, 14(12):1675–1682, 1976.

[37] Lang Deng, Jianfei Yang, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Gaitfi: Robust device-free human identification via wifi and vision multimodal learning. *IEEE Internet of Things Journal*, 2022.

[38] Srinivas Devarakonda, Parveen Sevusu, Hongzhang Liu, Ruilin Liu, Liviu Iftode, and Badri Nath. Real-time air quality monitoring through mobile sensing in metropolitan areas. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, pages 1–8, 2013.

[39] Jonathon Fagert, Mostafa Mirshekari, Shijia Pan, Linda Lowes, Megan Iammarino, Pei Zhang, and Hae Young Noh. Structure-and sampling-adaptive gait balance symmetry estimation using footstep-induced structural floor vibrations. *Journal of Engineering Mechanics*, 147(2):04020151, 2021.

[40] Peter A Flach. Roc analysis. In *Encyclopedia of machine learning and data mining*, pages 1–8. Springer, 2016.

[41] Dany Fortin-Simard, Jean-Sébastien Bilodeau, Sebastien Gaboury, Bruno Bouchard, and Abdenour Bouzouane. Human activity recognition in smart homes: Combining passive rfid and load signatures of electrical devices. In *2014 IEEE symposium on intelligent agents (IA)*, pages 22–29. IEEE, 2014.

[42] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780), 1999.

[43] Yi Gao, Wei Dong, Kai Guo, Xue Liu, Yuan Chen, Xiaojin Liu, Jiajun Bu, and Chun Chen. Mosaic: A low-cost mobile sensing system for urban air quality monitoring. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9. IEEE, 2016.

[44] Vincenzo Genovese, Andrea Mannini, and Angelo M Sabatini. A smartwatch step counter for slow and intermittent ambulation. *IEEE Access*, 5:13028–13037, 2017.

[45] James M Gere and Barry J Goodno. *Mechanics of materials*. Cengage learning, 2012.

[46] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[47] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[48] Jacob Goldberger, Geoffrey E Hinton, Sam T Roweis, and Russ R Salakhutdinov. Neighbourhood components analysis. In *Advances in neural information processing systems*, pages 513–520, 2005.

[49] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[50] Martin Grimmer, Kai Schmidt, Jaime Enrique Duarte, Lukas Neuner, Gleb Koginov, and Robert Riener. Stance and swing detection based on the angular velocity of lower limb segments during walking. *Frontiers in neurorobotics*, 13:57, 2019.

[51] Jun Han, Albert Jin Chung, Manal Kumar Sinha, Madhumitha Harishankar, Shijia Pan, Hae Young Noh, Pei Zhang, and Patrick Tague. Do you feel what i hear? enabling autonomous iot device pairing using different sensor types. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 836–852. IEEE, 2018.

[52] Jun Han, Shijia Pan, Manal Kumar Sinha, Hae Young Noh, Pei Zhang, and Patrick Tague. Sensetribute: smart home occupant identification via fusion across on-object sensing devices. In *Proceedings of the 4th ACM International*

Conference on Systems for Energy-Efficient Built Environments, pages 1–10, 2017.

[53] Michael Hanlon and Ross Anderson. Real-time gait event detection using wearable sensors. *Gait & posture*, 30(4):523–527, 2009.

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[55] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[56] Zhe Hui Hoo, Jane Candlish, and Dawn Teare. What is an roc curve?, 2017.

[57] Gaffar Hossain, Ishtia Zahir Hossain, and Günter Grabher. Piezoresistive smart-textile sensor for inventory management record. *Sensors and Actuators A: Physical*, 315:112300, 2020.

[58] Zhizhang Hu, Tong Yu, Yue Zhang, and Shijia Pan. Fine-grained activities recognition with coarse-grained labeled multi-modal data. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, pages 644–649, 2020.

[59] Zhizhang Hu, Yue Zhang, and Shijia Pan. Multimodal fine-grained human activity dataset, May 2022.

[60] Zhizhang Hu, Yue Zhang, Tong Yu, and Shijia Pan. Vma: Domain variance-and modality-aware model transfer for fine-grained occupant activity recognition. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 259–270. IEEE, 2022.

[61] Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

[62] Qian Huang, Zhenhao Ge, and Chao Lu. Occupancy estimation in smart buildings using audio-processing techniques. *arXiv preprint arXiv:1602.08507*, 2016.

[63] Input/Output, Inc. *SM-24 Geophone Element*, 4 2006. Rev. 3.

[64] AKM Najmul Islam, Elena Simona Lohan, and Markku Renfors. Moment based cnr estimators for BOC/BPSK modulated signal for galileo/GPS. In *2008 5th Workshop on Positioning, Navigation and Communication*, pages 129–136. IEEE, 2008.

[65] Yifei Jiang, Kun Li, Lei Tian, Ricardo Piedrahita, Xiang Yun, Omkar Mansata, Qin Lv, Robert P Dick, Michael Hannigan, and Li Shang. Maqs: a personalized mobile sensing system for indoor air quality monitoring. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 271–280, 2011.

[66] Sylwester Kaliski. *Vibrations and Waves: Part B: Waves*. Elsevier, 2013.

[67] B Kang, CH Riedel, and CA Tan. Free vibration analysis of planar curved beams by wave propagation. *Journal of sound and vibration*, 260(1):19–44, 2003.

[68] Chitra R Karanam, Belal Korany, and Yasamin Mostofi. Tracking from one side: Multi-person passive tracking with wifi magnitude measurements. In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks*, pages 181–192, 2019.

[69] Aimad Karkouch, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications*, 73:57–81, 2016.

[70] TL Kasteren, Gwenn Englebienne, and BJ Kröse. An activity monitoring system for elderly care using generative and discriminative models. *Personal and ubiquitous computing*, 14(6):489–498, 2010.

[71] Jinwoo Kim, Kyungjun Min, Minhyuk Jung, and Seokho Chi. Occupant behavior monitoring and emergency event detection in single-person households using deep learning-based sound recognition. *Building and Environment*, 181:107092, 2020.

[72] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[73] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.

[74] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[75] Chaofeng Li and Alan C Bovik. Three-component weighted structural similarity index. In *Image quality and system performance VI*, volume 7242, page 72420Q. International Society for Optics and Photonics, 2009.

[76] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[77] Yang Lin, Irena Koprinska, and Mashud Rana. Temporal convolutional attention neural networks for time series forecasting. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[78] Dima Litvak, Yaniv Zigel, and Israel Gannot. Fall detection of elderly through floor vibrations and sound. In *2008 30th annual international conference of the IEEE engineering in medicine and biology society*, pages 4632–4635. IEEE, 2008.

[79] Shuo Liu, Le Yin, Weng Khuen Ho, Keck Voon Ling, and Stefano Schiavon. A tracking cooling fan using geofence and camera-based indoor localization. *Building and Environment*, 114:36–44, 2017.

[80] Chris Xiaoxuan Lu, Hongkai Wen, Sen Wang, Andrew Markham, and Niki Trigoni. Scan: learning speaker identity from noisy sensor data. In *2017 16th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 67–78. IEEE, 2017.

[81] Matlab. resample uniform or nonuniform data to new fixed rate - matlab.

[82] Shenghuan Miao, Ling Chen, Rong Hu, and Yingsong Luo. Towards a dynamic inter-sensor correlations learning framework for multi-sensor-based wearable human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–25, 2022.

[83] Marija Milenkovic and Oliver Amft. An opportunistic activity-sensing approach to save energy in office buildings. In *Proceedings of the fourth international conference on Future energy systems*, pages 247–258. ACM, 2013.

[84] Mostafa Mirshekari, Jonathon Fagert, Shijia Pan, Pei Zhang, and Hae Young Noh. Step-level occupant detection across different structures through footstep-induced floor vibration using model transfer. *Journal of Engineering Mechanics*, 146(3):04019137, 2020.

[85] Mostafa Mirshekari, Jonathon Fagert, Shijia Pan, Pei Zhang, and Hae Young Noh. Obstruction-invariant occupant localization using footstep-induced structural vibrations. *Mechanical Systems and Signal Processing*, 153:107499, 2021.

[86] Mostafa Mirshekari, Shijia Pan, Jonathon Fagert, Eve M Schooler, Pei Zhang, and Hae Young Noh. Occupant localization using footstep-induced structural vibration. *Mechanical Systems and Signal Processing*, 112:77–97, 2018.

[87] Frank G Miskelly. Assistive technology in elderly care. *Age and ageing*, 30(6):455–458, 2001.

[88] Frank Mokaya, Brian Nguyen, Cynthia Kuo, Quinn Jacobson, Anthony Rowe, and Pei Zhang. Mars: a muscle activity recognition system enabling self-configuring musculoskeletal sensor networks. In *2013 ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 191–202. IEEE, 2013.

[89] Kee S Moon, Sung Q Lee, Yusuf Ozturk, Apoorva Gaidhani, and Jeremiah A Cox. Identification of gait motion patterns using wearable inertial sensor network. *Sensors*, 19(22):5024, 2019.

[90] Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019.

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

[92] Sandro Nižetić, Nikolina Pivac, Vlasta Zanki, and Agis M Papadopoulos. Application of smart wearable sensors in office buildings for modelling of occupants' metabolic responses. *Energy and Buildings*, 226:110399, 2020.

[93] Tommy Öberg, Alek Karsznia, and Kurt Öberg. Basic gait parameters: reference data for normal subjects, 10-79 years of age. *Journal of rehabilitation research and development*, 30:210–210, 1993.

[94] Alan V Oppenheim and Ronald W Schafer. Digital signal processing(book). *Research supported by the Massachusetts Institute of Technology, Bell Telephone Laboratories, and Guggenheim Foundation. Englewood Cliffs, N. J., Prentice-Hall, Inc., 1975. 598 p*, 1975.

[95] Studio Otwarte. Temporal convolutional networks and forecasting, Feb 2022.

[96] Shijia Pan, Amelie Bonde, Jie Jing, Lin Zhang, Pei Zhang, and Hae Young Noh. Boes: building occupancy estimation system using sparse ambient vibration monitoring. In *Sensors and smart structures technologies for civil, mechanical, and aerospace systems 2014*, volume 9061, page 90611O. International Society for Optics and Photonics, 2014.

[97] Shijia Pan and Phuc Nguyen. Opportunities in the cross-scale collaborative human sensing of'developing'device-free and wearable systems. In *Proceedings of the 2nd ACM Workshop on Device-Free Human Sensing*, pages 16–21, 2020.

[98] Shijia Pan, Carlos Ruiz, Jun Han, Adeola Bannis, Patrick Tague, Hae Young Noh, and Pei Zhang. Universense: Iot device pairing through heterogeneous sensing signals. In *Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications*, pages 55–60, 2018.

[99] Shijia Pan, Ningning Wang, Yuqiu Qian, Irem Velibeyoglu, Hae Young Noh, and Pei Zhang. Indoor person identification through footstep induced structural vibration. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, pages 81–86, 2015.

[100] Shijia Pan, Tong Yu, Mostafa Mirshekari, Jonathon Fagert, Amelie Bonde, Ole J Mengshoel, Hae Young Noh, and Pei Zhang. Footprintid: Indoor pedestrian identification through ambient structural vibration sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–31, 2017.

[101] Shijia Pan, Tong Yu, Mostafa Mirshekari, Jonathon Fagert, Amelie Bonde, Ole J Mengshoel, Hae Young Noh, and Pei Zhang. Footprintid: Indoor pedestrian identification through ambient structural vibration sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–31, 2017.

[102] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga,

et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[103] S Patro and Kishore Kumar Sahu. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*, 2015.

[104] Leo L Pipino, Yang W Lee, and Richard Y Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[105] Jay Prakash, Zhijian Yang, Yu-Lin Wei, and Romit Roy Choudhury. Stear: Robust step counting from earables. In *Proceedings of the 1st International Workshop on Earable Computing*, pages 36–41, 2019.

[106] Juhi Ranjan and Kamin Whitehouse. Object hallmarks: Identifying object users using wearable wrist sensors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 51–61, 2015.

[107] Yacine Rekik, Evren Sahin, and Yves Dallery. Analysis of the impact of the rfid technology on reducing misplacement errors at the retailer. *International Journal of Production Economics*, 112:264–278, 03 2008.

[108] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[109] F Rohde. Large deflections of a cantilever beam with uniformly distributed load. *Quarterly of Applied Mathematics*, 11(3):337–338, 1953.

[110] G. Roussos. Enabling rfid in retail. *Computer*, 39(3):25–30, 2006.

[111] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[112] Carlos Ruiz, Joao Falcao, Shijia Pan, Hae Young Noh, and Pei Zhang. Autonomous inventory monitoring through multi-modal sensing (aim3s) for cashier-less stores. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 395–396, 2019.

[113] Carlos Ruiz, Shijia Pan, Adeola Bannis, Ming-Po Chang, Hae Young Noh, and Pei Zhang. Idiot: Towards ubiquitous identification of iot devices through visual and inertial orientation matching during human activity. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 40–52. IEEE, 2020.

[114] Carlos Ruiz, Shijia Pan, Adeola Bannis, Xinlei Chen, Carlee Joe-Wong, Hae Young Noh, and Pei Zhang. Idrone: Robust drone identification through motion actuation feedback. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–22, 2018.

[115] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[116] Carmine Sellitto, Stephen Burgess, and Paul Hawking. Information quality attributes associated with rfid-derived benefits in the retail supply chain. *International Journal of Retail & Distribution Management*, 35(1):69–87, 2007.

[117] SA Shapiro and G1 Kneib. Seismic attenuation by scattering: theory and numerical results. *Geophysical Journal International*, 114(2):373–391, 1993.

[118] Laixi Shi, Mostafa Mirshekari, Jonathon Fagert, Yuejie Chi, Hae Young Noh, Pei Zhang, and Shijia Pan. Device-free multiple people localization through floor vibration. In *Proceedings of the 1st ACM International Workshop on Device-Free Human Sensing*, pages 57–61, 2019.

[119] Kannan Srinivasan and Philip Levis. Rssi is under appreciated. In *Proceedings of the third workshop on embedded networked sensors (EmNets)*, volume 2006. Cambridge, MA, USA., 2006.

[120] Seth Stein and Michael Wysession. *An introduction to seismology, earthquakes, and earth structure*. John Wiley & Sons, 2009.

[121] Fred Szabo. *The linear algebra survival guide: illustrated with Mathematica*. Academic Press, 2015.

[122] Texas Instruments Incorporated. *LMV3xx Low-Voltage Rail-to-Rail Output Operational Amplifier*, 5 2020.

[123] Northwestern University.

[124] Vincent Van Asch. Macro-and micro-averaged evaluation measures. *Belgium: CLiPS*, 49, 2013.

[125] CJ Van den Branden Lambrecht. Special issue on image and video quality metrics. *Signal Processing*, 1998.

[126] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *SSW*, 125:2, 2016.

[127] Remko van der Togt, Piet JM Bakker, and Monique WM Jaspers. A framework for performance and data quality assessment of radio frequency identification (rfid) systems in health care settings. *Journal of Biomedical Informatics*, 44(2):372–383, 2011.

[128] Igor Aleksandrovich Viktorov. *Rayleigh and Lamb waves: physical theory and applications*. Plenum press, Berlin, Germany, 1970.

[129] Enrico Volterra and Eleftherios Charalampos Zachmanoglou. *Dynamics of vibrations*, volume 1. CE Merrill Books, 1965.

[130] Jin Wang, Mary She, Saeid Nahavandi, and Abbas Kouzani. A review of vision-based gait recognition methods for human identification. In *2010 international conference on digital image computing: techniques and applications*, pages 320–327. IEEE, 2010.

[131] Rui Wang, Weichen Wang, Alex DaSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–26, 2018.

[132] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE signal processing letters*, 9(3):81–84, 2002.

[133] Zhou Wang, Ligang Lu, and Alan C Bovik. Video quality assessment based on structural distortion measurement. *Signal processing: Image communication*, 19(2):121–132, 2004.

[134] Xiu-Shen Wei, Quan Cui, Lei Yang, Peng Wang, and Lingqiao Liu. Rpc: A large-scale retail product checkout dataset, 2019.

[135] Susu Xu, Xinlei Chen, Xidong Pi, Carlee Joe-Wong, Pei Zhang, and Hae Young Noh. ilocus: Incentivizing vehicle mobility to optimize sensing distribution in crowd sensing. *IEEE Transactions on Mobile Computing*, 19(8):1831–1847, 2019.

[136] Susu Xu, Xinlei Chen, Xidong Pi, Carlee Joe-Wong, Pei Zhang, and Hae Young Noh. Vehicle dispatching for sensing coverage optimization in mobile crowdsensing systems. In *2019 18th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 311–312. IEEE, 2019.

[137] Rui Yang and Lingfeng Wang. Development of multi-agent system for building energy and comfort management based on occupant behaviors. *Energy and Buildings*, 56:1–7, 2013.

[138] Zhe Yang, Qihao Zhou, Lei Lei, Kan Zheng, and Wei Xiang. An iot-cloud based wearable ecg monitoring system for smart healthcare. *Journal of medical systems*, 40(12):1–11, 2016.

[139] Zhibo Yang, Xuefeng Chen, Xiang Li, Yongying Jiang, Huihui Miao, and Zhengjia He. Wave motion analysis in arch structures via wavelet finite element method. *Journal of Sound and Vibration*, 333(2):446–469, 2014.

[140] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[141] Ehsan Yavari, Hsun Jou, Victor Lubecke, and Olga Boric-Lubecke. Doppler radar sensor for occupancy monitoring. In *2013 IEEE Topical Conference on Power Amplifiers for Wireless and Radio Applications*, pages 145–147. IEEE, 2013.

[142] Ting-Hua Yi, Hong-Nan Li, and Xiao-Yan Zhao. Noise smoothing for structural vibration test signals using an improved wavelet thresholding technique. *Sensors*, 12(8):11205–11220, 2012.

[143] Jie Yin, Son N Tran, and Qing Zhang. Human identification via unsupervised feature learning from uwb radar data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 322–334. Springer, 2018.

[144] Hong Ying, Carmen Silex, Andreas Schnitzer, Steffen Leonhardt, and Michael Schiek. Automatic step detection in the accelerometer signal. In *4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2007)*, pages 80–85. Springer, 2007.

[145] Tong Yu, Yue Zhang, Zhizhang Hu, Susu Xu, and Shijia Pan. Vibration-based indoor human sensing quality reinforcement via thompson sampling. In *Proceedings of the First International Workshop on Cyber-Physical-Human System Design and Implementation*, pages 33–38, 2021.

[146] Jin Zhang, Bo Wei, Fuxiang Wu, Limeng Dong, Wen Hu, Salil S Kanhere, Chengwen Luo, Shui Yu, and Jun Cheng. Gate-id: Wifi-based human identification irrespective of walking directions in smart home. *IEEE Internet of Things Journal*, 8(9):7610–7624, 2020.

[147] Yue Zhang, Zhizhang Hu, Uri Berger, and Shijia Pan. Integrating on-and off-body sensing for young adults failure to launch (ftl) behavior profiling. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*, pages 320–321, 2023.

[148] Yue Zhang, Zhizhang Hu, Susu Xu, and Shijia Pan. Autoqual: task-oriented structural vibration sensing quality assessment leveraging co-located mobile sensing context. *CCF Transactions on Pervasive Computing and Interaction*, pages 1–19, 2021.

[149] Yue Zhang, Shijia Pan, Jonathon Fagert, Mostafa Mirshekari, Hae Young Noh, Pei Zhang, and Lin Zhang. Occupant activity level estimation using floor vibration. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 1355–1363, 2018.

[150] Yue Zhang, Susu Xu, Laixi Shi, and Shijia Pan. Using mobile sensing to enable the signal quality assessment for infrastructure sensing systems. In *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*, pages 102–102, 2020.

[151] Yue Zhang, Lin Zhang, Hae Young Noh, Pei Zhang, and Shijia Pan. A signal quality assessment metrics for vibration-based human sensing data acquisition. In *Proceedings of the 2nd Workshop on Data Acquisition To Analysis*, 2019.