**MSBD 6000F Final Project**

**Haojun Chen 12233738**

**Xiling Sheng 20491358**

**The Hong Kong University of Science and Technology**

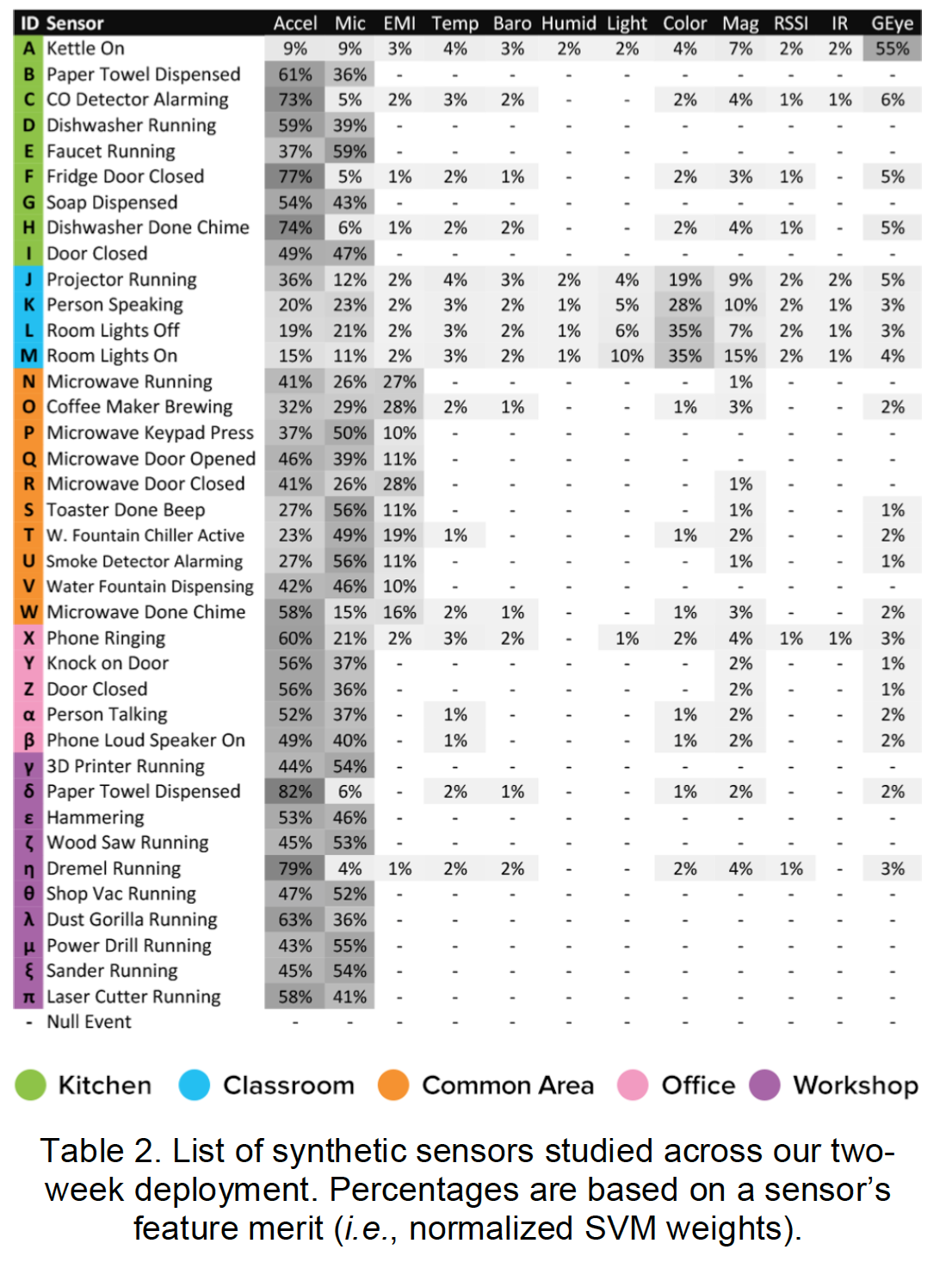
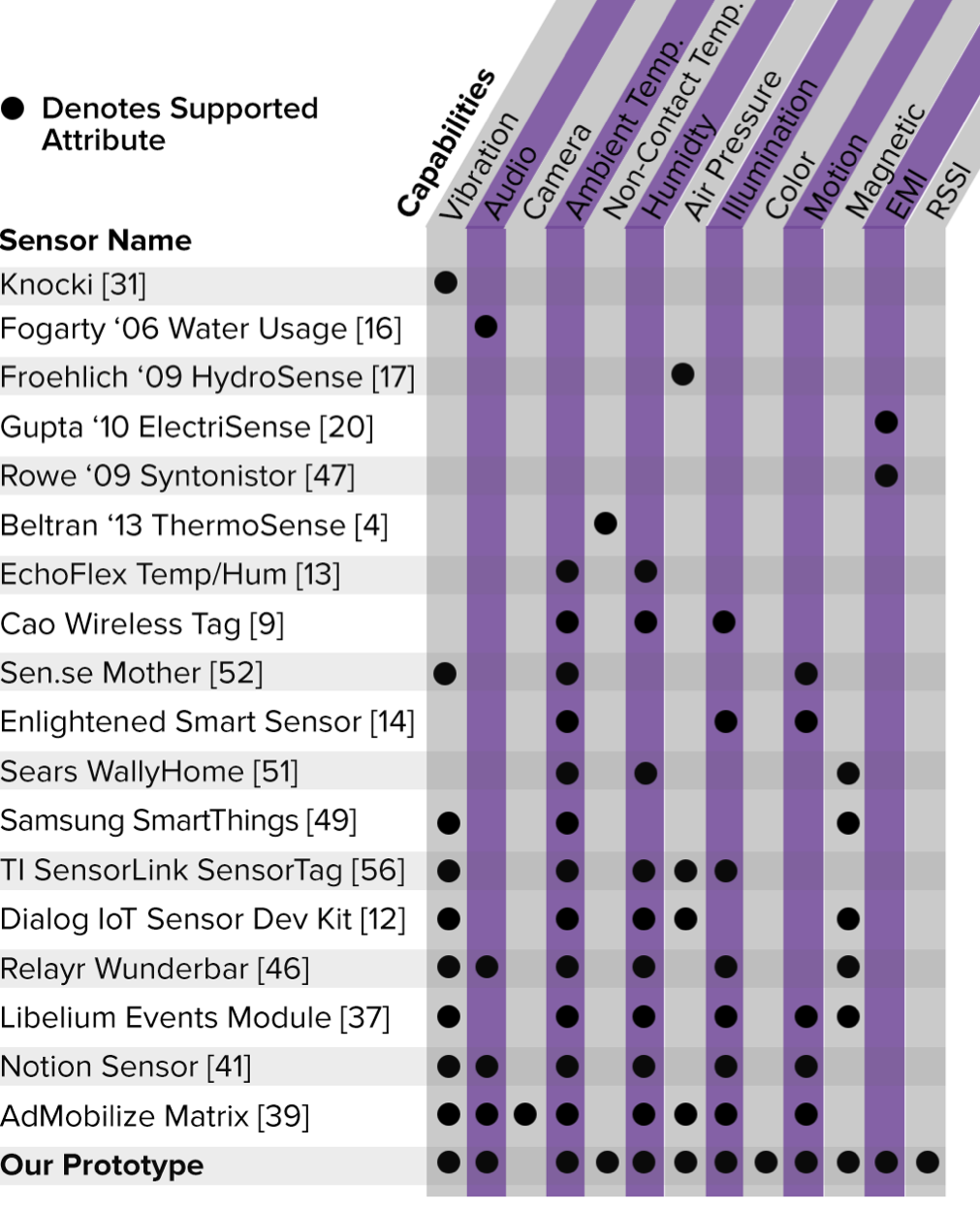
**Abstract**

The promise of smart environments and the Internet of Things (IoT) relies on robust sensing of diverse environmental facets. Traditional approaches rely on direct or distributed sensing, most often by measuring one particular aspect of an environment with special-purpose sensors [12]. What motivates us is the potential to explore the general-purpose sensors that are highly capable of indirectly monitoring a large context, with no direct instrumentation of objects. One of the major general-purpose sensors is the microphone. Despite sound being a rich source of information, computing devices with microphones do not leverage audio to glean useful insights about their physical and social context. For example, a smart speaker sitting on a kitchen countertop cannot figure out if it is in a kitchen, let alone know what a user is doing in a kitchen – a missed opportunity (Figure 1A) [13]. Hence, we are motivated to research on a novel, real-time, sound-based activity recognition system (e.g. Ubicoustics), which takes advantage of the ubiquity of microphones in general consumer electronics such as smart speaker, smart watch, tablets and so on (Figure 1 B-F). In short, as the traditional approaches do not work well enough, we want to explore some potential approaches that can address the above-mentioned issues.

 *Figure 1. Ubicoustics: an activity sensing system*

**Introduction**

Undoubtedly, there is a rising interest in smart environments that enhance the quality of live for humans in terms of e.g. safety, security, comfort, and home care. In order to have smart functionality, situational awareness is required, which might be obtained by interpreting a multitude of sensing modalities including acoustics [1].” The catchy phrases like the “smart home” or “the internet of things” have been addressed a lot but even numerous approaches have been attempted and articulated, none have achieved any popularity up to date. One application is that users can choose to upgrade their environments with new devices such as light switches, kitchen appliances and so on. Though these devices have sensing functionality, their sensing ability is limited to the appliance itself. In addition, the approach to achieve a smart home requires significant upgrade cost. Moreover, extensively constructing an environment that may have dozens of complex environmental facets worth sensing will unavoidably have an aesthetic and social cost too. This gives rise to the awareness of a lightweight, general-purpose, sensing approach that would overcome these issues. In terms of the general-purpose sensing, sensor “boards” as shown in Table 1 contain a wide range of underlying sensors with flexibility that might be considered general purpose because they can be attached to many objects and sense different facets without modification. An ideal sensing approach usually has the characteristics of single, omniscient, and capable of digitizing the whole building [12].



*Table 1. An inventory of research and*

*commercial sensors offering varying degrees of general-purpose sensing.*

We found this problem interesting because acoustics is one of the most important sensing modalities and is gaining increasing popularity in achieving our smart environment nowadays. In addition, there is still a lot of room to improve in terms of the recognition accuracy, power consumption, computational cost and so on. Therefore, our project will mainly focus on the following:

* Use different Machine Learning techniques to analyze the acoustic features and give predictions on what’s happening in the home.
* As the popularity of the chip that can be implemented in neural networks, the local device-run machine-learning algorithm that can recognize sound event has been explored on different types of devices: IoT, mobile phone and smart watch, etc.
* Research the current state-of-the-art techniques that we can leverage in order to come up with a comprehensive solution which not only includes microphones but also EMI, accelerometer etc. By learning the comprehensive pattern through these sensors, a more accurate prediction can be achieved.

Problem Space

Based on the improvement space, we want to use the acoustic feature to detect what happens at home as part of smart home solution, in which smart speakers and vocal assistants such as Amazon Echo, Apple HomePod become increasingly popular and multiple IoTs with microphones can be placed different locations at home. The issue that we are trying to tackle is to potentially find out how to improve the detection rate of acoustic sensor, while maintaining robustness. by leveraging the concept of ‘Smart Home’. The reason that we choose Acoustics Sensor is that it has the advantages as follows:

* Unlike cameras, not involved in any invasion of instrumentation
* Effectiveness of detection of majority activities compare to other sensors (Table 2)
* Smart Speaker is of widespread use
* Microphone as a commonly-used device, can be applied on smart watch, smartphone, computer and IoT devices

We will then come up with different observations after researching on existing work and try to shed light on the disadvantages or limitations that the previous work may not have addressed or handled well enough.

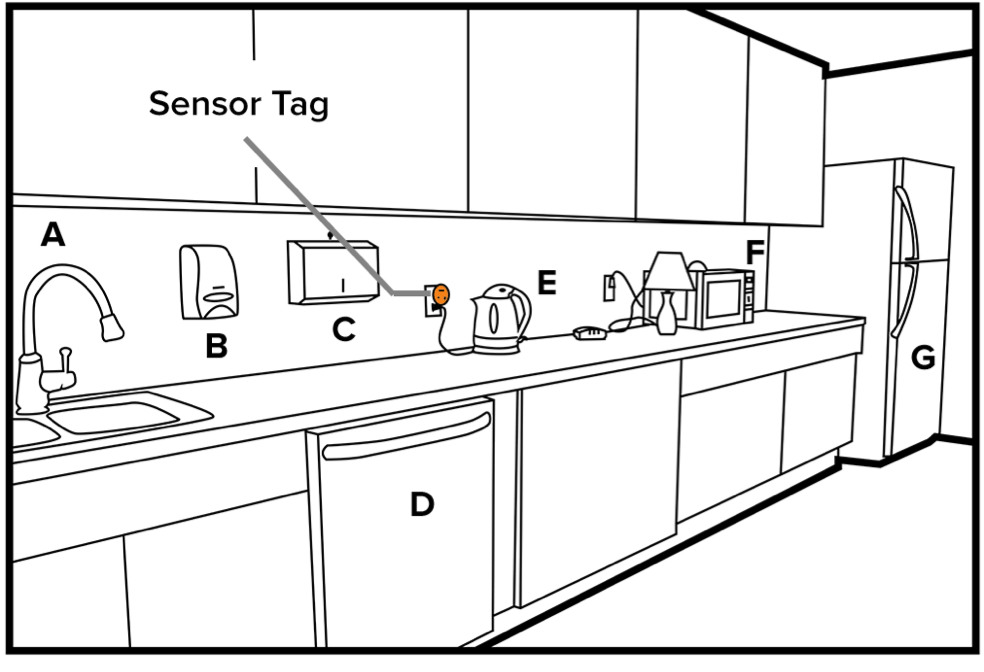
**Related Work**

Currently, computer vision has made its way to achieve the goal of digitizing an entire building as cameras can provide rich, indirect data that can be processed through machine learning to yield sensor related feeds. Since there are a considerable amount of body of work in video-based sensing, achieving human-level abstractions and accuracy is a concrete challenge. Zensors [15], which use a commodity camera, can digitize a wide range of environmental facets, e.g., “how many dishes are in the sink”. Though this CV-based function is rather powerful and may have other advantages, cameras do have been recognized for their higher possibility of privacy invasion and social intrusiveness and thus has hindered their use in many ripe scenarios such as homes, schools, work environments and other settings. This is a serious drawback for camera-based sensors, Hence, we will explore acoustic sensors with much of the same sensing versality and accuracy without camera use.

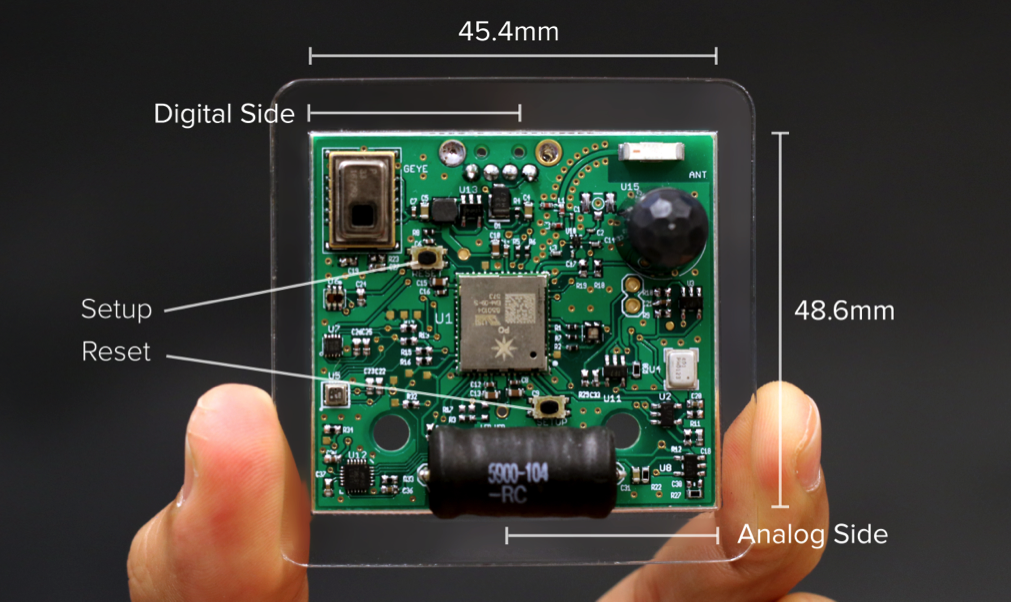
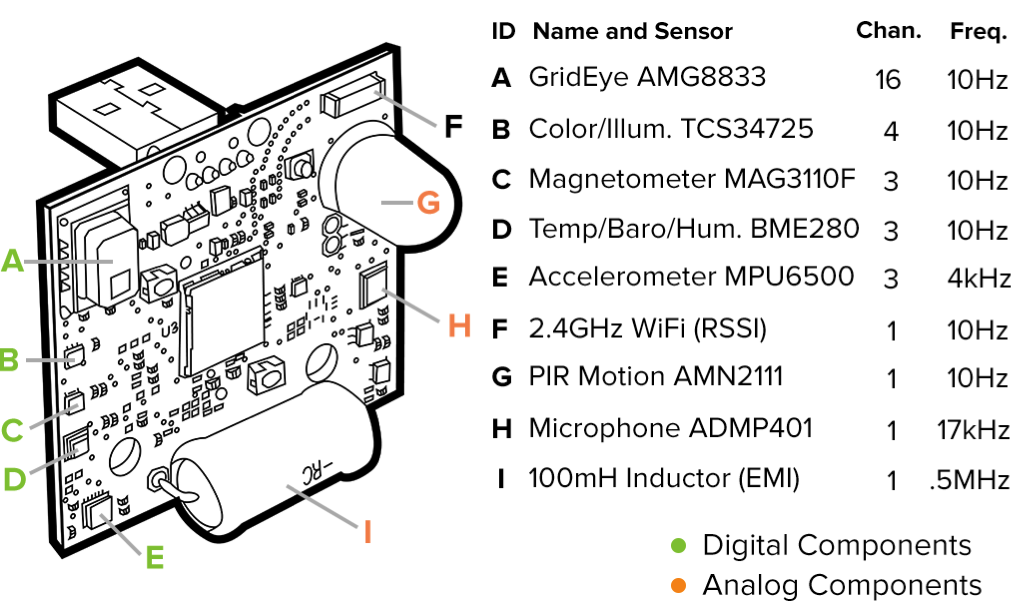
While acknowledging that Smart Houses (Home) are feasible and cost-effective to help improve comfort for the user, Veralia G et al. [2], nevertheless, addressed the ethical challenges that are needed to overcome – it is essential to study the ethical impact of cameras. In the case where the cameras are used, they should be used properly and the privacy of the user must be respect. Computer vision methods are one of the methods to detect the context of Smart House via generating predictive models to monitor human activity and behavior from the sensor data collected. Whereas, some limitations are present on computer vision in a Smart House System. Besides the sentiment that some users may have considered to be invasive, additionally, cameras may produce noisy images causing data generate more uncertainties and thus limiting the accuracy of the results. The authors also point out that even though image processing techniques can improve the image quality collected from the cameras with the computer vision doing the recognition and learning task thereafter, a disadvantage of this method is that processing a heavy number of images can be computationally costly.

Fortunately, from the list of synthetic sensors (Table 2), we can see the weighted breakdown of merit calculated by SVM weights. In particular, the most helpful three sensors are: microphone (17 kHz sample rate), accelerometer (4 kHz) and EMI (500 kHz) – the three highest sample rate sensors on our board. It is also reassured by the author that this result should not be overgeneralized so using these three sensors alone suffice for general purpose sensing.

Synthetics Sensor is a type of sensor be used to virtualize raw sensor data into actionable feed, in the meanwhile, mitigating immediate privacy issues. In the exploratory studies, even though the low-level sensor data has high fidelity, it does not reflect the users’ true intent. Most users do not care about the spectrogram of EMI emissions from coffee maker but they want to know when their coffee is ready or their laundry is done. For this kind of needs, acoustic recognition would be more fit for prediction of such status change. ‘Virtualization’ of low-level data is needed for semantic representations. Synthetic Sensors are then introduced because they are versatile, user-centered and are capable of general-purpose sensing [12]. Moreover, Synthetic Sensors demonstrated hardware featuring many discrete sensors – including a high-speed accelerometer for vibration sensing – that could detect a wide range of activities across an entire room from a single instrumented point.



*Figure 2. This kitchenette example illustrates the concept of general-purpose sensing, wherein one sensor (orange) enables the detection of many environmental facets, including rich operational states of a faucet (A), soap dispenser (B), paper towel dispenser (C), dishwasher (D), kettle (E), microwave (F) and refrigerator (G).*



*Figure3. sensor tag features nine discrete sensors Figure4. General-purposed sensor map*

*- able to capture twelve unique sensor dimensions*

A novel sensor tag is purposed (Figures 3 and 4). It integrates the union of the sensing capabilities across

all of the devices in Table 1, minus a camera. Not only does this serve as an interesting vehicle for investigation (e.g., what sensors are most accurate and useful?), but also an extreme embodiment of board design using many low-level sensors because it is expected to achieve the versatility of camera-based approaches but without the stigma and privacy implications. Nine physical sensors are used to capture twelve distinct sensor dimensions (see Figure 3). The heart of our sensor tag design is a Particle Photon STM32F205 microcontroller with a 120MHz CPU. We strategically placed sensors on the PCB to ensure optimal performance (e.g., ambient light sensor faces outwards), and spatially separated analog and digital components to isolate unintended electrical noise from affecting the performance of neighboring components.

Real-time, sound-based classification of activities and context is not new. There have been many previous application specific efforts. For example, Ward et al. [41] developed a microphone-equipped necklace related to accelerometers mounted on arms that could distinguish between nine shop tools. In these types of constrained uses, the training data for machine learning is mostly domain-specific and captured by the researchers themselves. Thus, Laput et al [13] proposed a more general-purpose and flexible sound recognition pipeline that could be used on the existing device for update and work instantly without the need of collecting end-user or in situ data. This system is called “plug-and-play”. For example, when you plug in your Alexa, it will immediately detect the activities of your kitchen appliances by sound. However, such task is challenging and there are not many sound-based recognitions system capable of achieving the usable end-user accuracies even with the help of pre-trained models that applied in applications like Youtube-8M and SoundNet.

Furthermore, Alexander et al [17] addresses some examples application that can benefit from implementing the wireless acoustic sensor networks (WASN) such as the hands-free telephony, acoustic monitoring techniques in the context of acoustic monitoring environment, for example, vehicle classification/tracking, surveillance, etc., ambient intelligence, which is an intelligent environment that is aware of where a person is and is responsive to his/her needs.

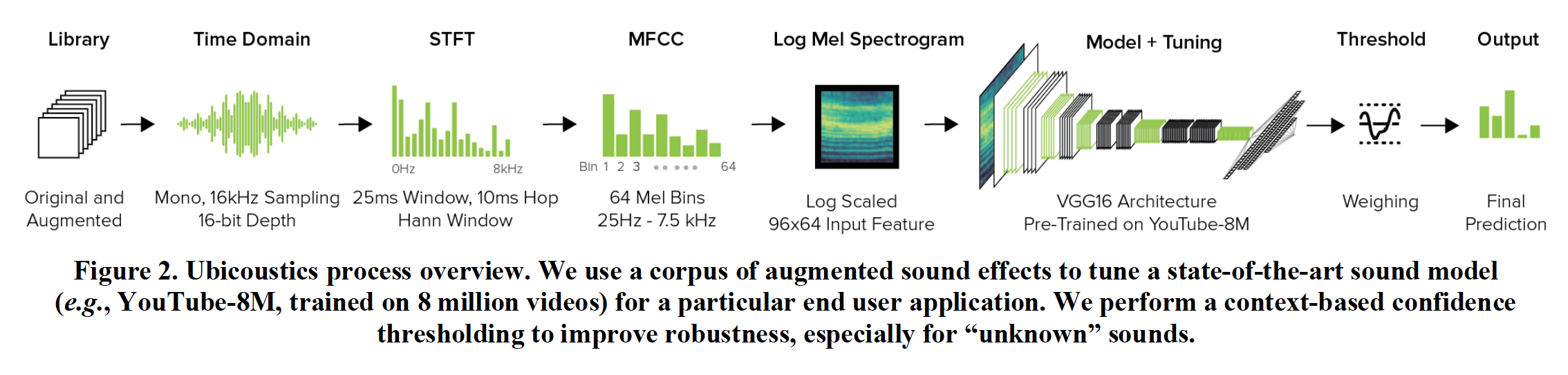
Since acoustic signal are used in those sensors above, the accuracy of sensing of environmental activity largely depended on the how well the model can interpret the acoustic data. That's why the DCASE competition and the Ubicoustic [13] try to improve the effectiveness of model through state of art machine learning techniques and manipulation of professional sound data.

In our project, we introduce the Gierad [13] uses the professional sound effect libraries as corpus to train high-level sound recognition. The professional sound effect usually is ideal for training sound recognition models due to following the property can be obtained from the professional sound effect library. First, atomic: The sound effect has one sound clip that represent the sound event. second, pure. Most of the sound clips are recorded in the studio with relatively high quality without artifacts. Third, diversity. The professional sound effect library usually contains a variety of sound clips belongs to the same effect category. From these three points, the author believes the professional sound effects library is perfectly to be used for machine learning towards sound recognition.

To achieve a better machine learning result, the property of sound needs to be considered. Changing the property of the sound can more reflect the real sound propagation. Foremost is the amplitude and duration, which reflects the loudness and time length/tempo of the sound when we hear. The second property is the transformation of the sound, which means we can add reverb or equalizer (frequency response) and damping to modify its timbre and location context. For example, we could turn a dry sound ‘siren’ into a hall room and make sound faraway with damping its high-frequency spectrum. (the higher frequency is more easily drop during transmitting), the third property is multiple sounds can be added on top of each other. For example, we could put a hair dryer sound on top of street noise or living room noise.

For these abovementioned properties, we observe that multiple corpus would reflect different realistic variations of sound from one sound effect so we could build up a large quantity of corpus for machine learning. The author has showed obvious higher accuracy outcome of prediction when using data augmentation including the changing the property of the sound.

The workflow

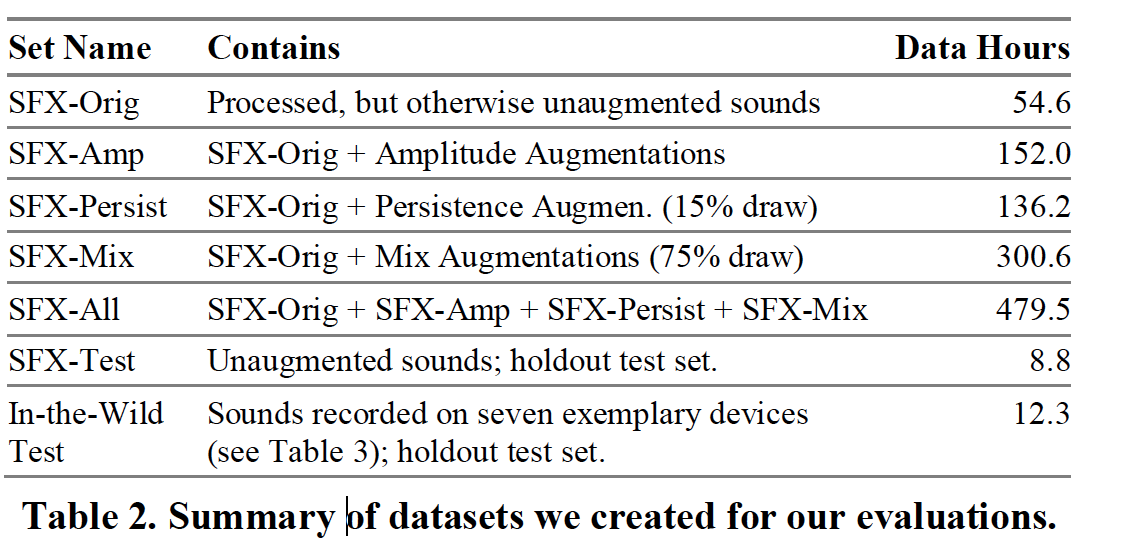


The following would generally introduce the workflow to achieve machine learning of author ‘ubicoustics’ model.

Firstly, we first group the sound clips in the library and make them labeled as we need to classify. For example, ‘the running water ‘and ‘faucet’ can belong to the same category or context. And we could put different sound effect within the same context from our interest.

Second, sound data pre-processing. For the convenience, the sound file needs to be converted to the single format.  The author uses 16khz as the sampling rate and 16-bit and mono-channel as the original sound data for training.

Third, data augmentation. The augmentation can be divided into three parts based on the property changes including amplitude and persistence change, and mixing change. Amplitude is changed by preprocessing the audio data louder and quieter. The persistence changes through applying the reverb and damping model on the data and thus different versions of the same sound are reflected in realistic environments such as ‘bathroom’ and ‘atrium’ and etc. The mixing change is done by adding sound effect on top of different background noise to reflect the real situation in indoor or outdoor listening situation: such as coffee shop and bedroom. Lastly, we can combine those augmentation among amplitude, persistence and mixing to create a variety of versions to for learning.

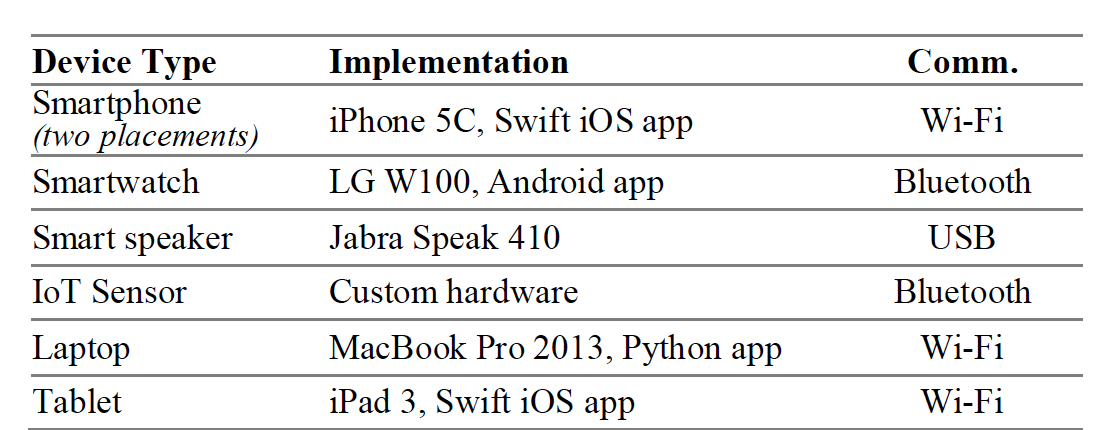


Fourth, featurization. For capturing feature from the dataset, the author used the method coming from Hershey et al. [16]. First, we cut clips into segments of 960ms and compute the STFT (short-time Fourier transforms) for each segment. The window size is 25ms and step size is 10ms. The spectrogram of 96-length is produced, which later the 64-bin log-Mel spectrogram is converted from. After this, we generated 96 \*64 input frame as input feature for our classification model.

The classification model the author use is YouTube-8M VGG-16 [16], which is a variant of the VGG 16 architecture which has trained on 8 million YouTube videos. The author uses this model as our pre-trained model which removed the last layer of the fully connected layer but replacing with our own fully-connect model using a sigmoid activation function. The work left is to tune the parameters based on our owned sound effect corpse.

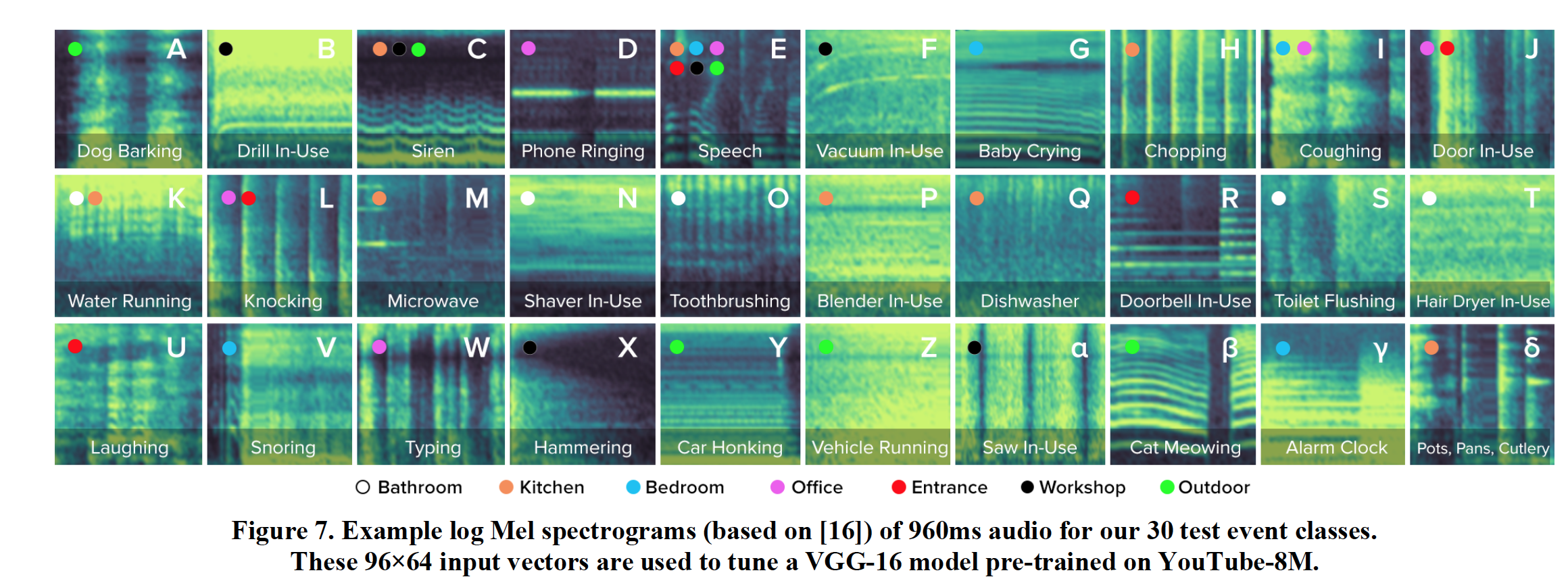
After training the model, we put our model run on local different types of devices, instead of putting

model on the cloud which requires uploading audio data to the cloud which may raise the concern of data privacy. The testing device include iPhone 7, iWatch, IoT device (Raspberry Pi Zero W with a ReSpeaker dual mic shield [34], MacBook Pro 2017. Due to the different computation power these devices have, we run the model with different frame rate respectively, a total of 7 devices have been evaluated for the recognizing performance.



**Evaluation**

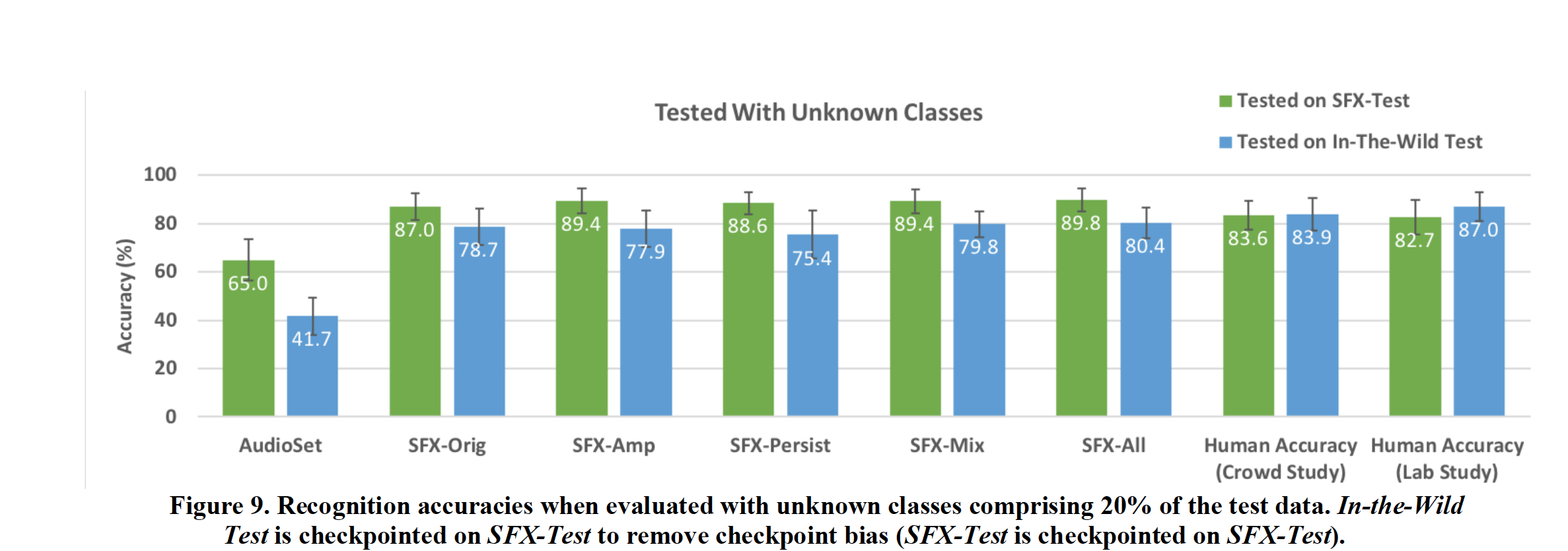
The final test class output contains 30 class including the most used sound event happened in our daily life with certain 7 context. These 30 class and 7 context the can be found through the following graph. There are some principles when choosing the class and context: first, context should offer the realistic scenarios with constraint event class. Second, the sound events should frequently happen in that certain context with proper loudness and prior knowledge. Based on these two principles, the author can tune the model depending on the number of class over particular context. The context and the sound event can be bundled like the following graph.



For the test data, besides we have the test portion from the professional sound effect library, the author will record the sound data through real devices’ mic, such as from iPhone, iPad, MacBook Pro, IoT and smart watch. We call this data from self-recorded as ‘in-the-wild’ sound. Since these would be the real situation that the trained model would put the test on. The tester would take the devices to the different 50 rooms. The way they place the mic way also follows the realistic fashion that IoT Sensor will be plugged into the near-by power outlet, and phone will be placed on the desk and in the pocket situation. To assure certain SNR, the device should not far away more than 3 meters from the source.

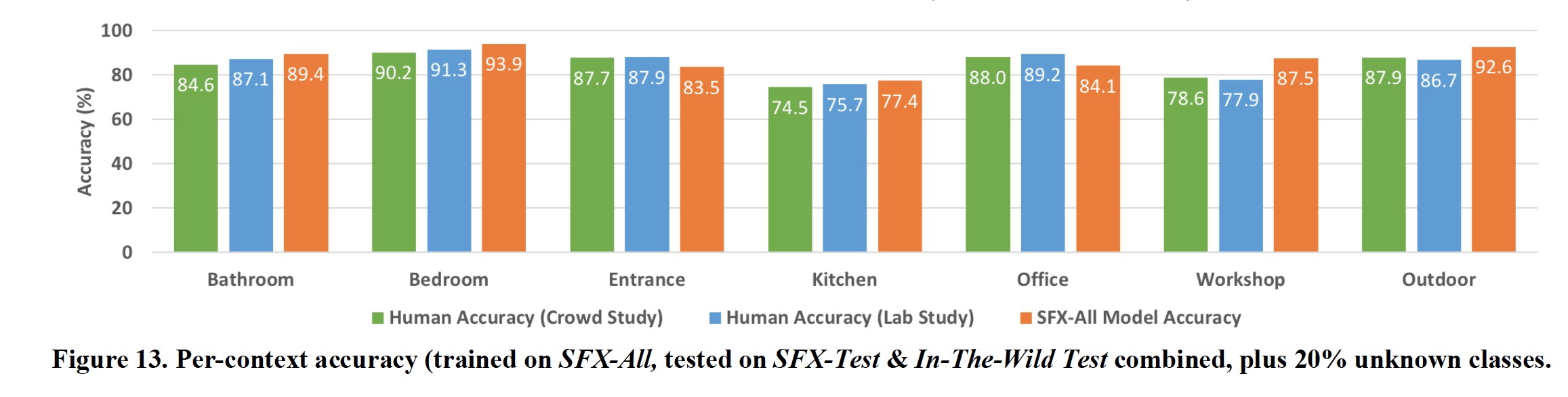
**Result and Discussion.**

The author shows us great promising result in accuracy that the model can achieve compared to the human annotators. From the following figure we could see that even incorporated with unknown class which the sound not even fall in any of the 30 class, the model can nearly compete as human did. The human accuracy still higher than the machine listening which above 3% - 7%.

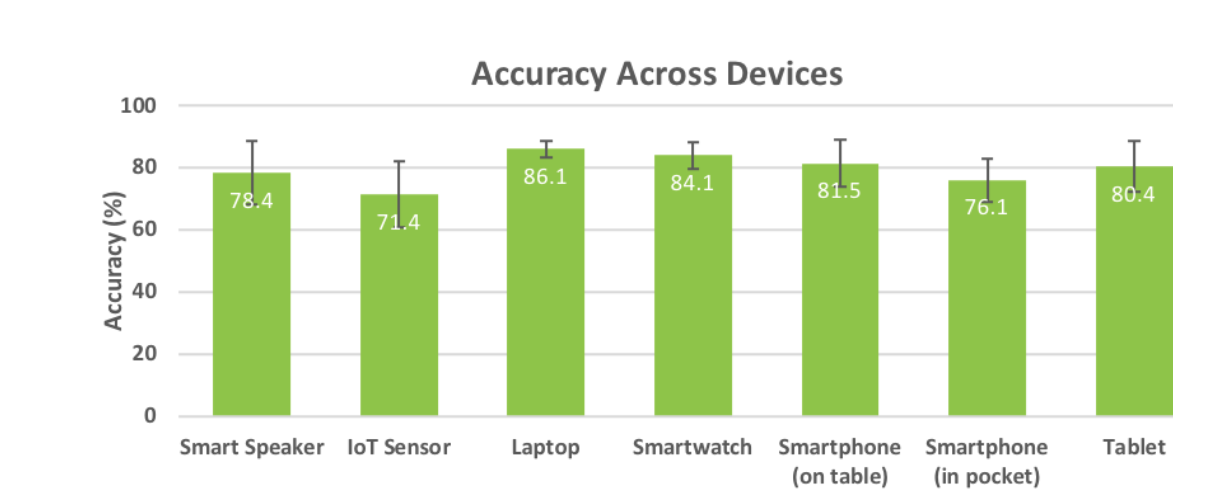
****

Another interesting test is not limited to classify the sound event, but to infer the location context through the sound classification. This is done to be using event classification as proxy for the original context. For some times, the device itself won’t know where it has been placed, but recognizing the sound event it listened, so the model then infers the context. For example, if there is running water sound, the model should more believe this event happen in the bathroom or the kitchen but not in the bedroom.

To simulate this experiment, the author choose ten random sound clips for each context and feed them into the model of 30 class of SFX-all , than the classification outcome would vote for a certain context, some of it using special logic as human’s prior logic: some sound presumed to be happen in certain context or some sound can be happen in anywhere which don't cast a vote (context free event). So, the number with the highest vote with the certain context is chosen and validated against the ground truth context. And this turns out to be a pretty good result compared to human annotation. As following graph shows, model exceeds human judge sound in a certain context are ‘bathroom’, ‘bedroom’, ‘kitchen’, ‘workshop’ and ‘outdoor’. Even for other context like the ‘entrance’ and ‘office’, the model is pretty near to the human judging.



The author also compared the performance across different devices. As the following graph shows that the accuracy that model SFX-all predict ‘in -the -wild’ sound, we can see that IoT sensor achieves the lowest and the laptop get the highest. This is due to the fact that IoT usually is further away with the source but attached to the power socket. Also, the we could see that in-pocket smartphone achieves about 5% lower accuracy than the smart phone on the table as the pocket usually block the mics through cloth and introduced chafing noise.

****

**Potential Improvement (Proposed Solution)**

As we find that acoustic sensor could predict some activities, there are still a lot of room to improve. Firstly, if we look this point from a more general sensing aspect of view, acoustic sensing for activity is only one of the very important method that we could deploy. However, for certain activities, for example (faucet running, microwave running), sound sensing may not be the best option. but other sensor such as Accelerometers or EMI could tell better story with more integrity. So, a more general sensing IoT which include multiple sensors should be explored towards specific situation.

What’s more, from accuracy point of view, the model for recognizing different sound events could be more improved by different machine learning techniques and advanced signal processing. The author has mentioned that the ML techniques could be incorporated with deeper and better model architectures, such as ResNet [44] and those can model the temporal property of sound such as LSTMs and CRNN [14]. Besides, better signal processing techniques such noise suppression, plus acoustic spatial properties would increase the SNR which will benefits more accuracy rate.

The simultaneous sound event also poses a challenge for the model to recognize. As we live in the real-world full of noise, how we interpret the complex sound event while our model is more confined to clean sound effect without overlapping sound effects. This could be a direction worth further exploring.

Privacy is always be a sensitive issue when we capture audio data without notifying. As author using the low resolution of 64 bins of Mel spectrograms discarding the phase for feature detection, which is good enough for model to classify sound at the same time hard to recover the speech directly, the audio data is never need to be transmitted or stored.

Further work including applications could be further explored towards more tailored use case. For example, if we install our model onto the smart speaker, then we could enable the speaker to monitor certain sound events than inform us. For example, we would like our smart speaker to inform other devices to start when we come home by open the door. Or we could get alarm as for the ‘laundry is done’, IoT sensor can detect how long we take the shower by listening the water running and later tell us how much water has been used. Smart device could also track people’s cough or snoring through the model and give out the possible indication of sick or recommendation to doctor.

**Conclusion**

After analyzing the applications of acoustic, we have now shown that Ubicoustics enables real-time activity recognition by leveraging one of the most common sensors found in consumer electronics today – microphones – bringing the promise of smart devices and environments closer to reality. By leveraging existing state-of-the-art sound classification models and tuning them with sound effects, a general-purpose and flexible sound recognition pipeline that requires no in situ data collection is purposed and enabled, yielding innovative end user experience. We evaluated the robustness of our approach across different physical contexts and hardware platforms, and show that our system can achieve superior accuracies over prior work, both in terms of recognition accuracy and false positive rejection. [13] Moreover, Synthetic Sensors, which allows everyday locations to become ‘smart environments’ without invasive attack, shows that general-purpose sensing can be flexible and robust and be applied in variety of applications. [12] For more information about the Ubicoustics, you may kindly refer to the links below to find out more about the Future Interfaces Group of Human-Computer Interaction Institute at Carnegie Mellon's School of Computer Science (1), (2).

**Summary of Member’s Contribution**

Haojun Chen: Abstract, Introduction, Problem Space, Related Work, Conclusion

Xiling Sheng; Evaluation, Result and Discussion, Proposed Solution, Conclusion

**References**

[1] Gert Dekkers, DCASE 2018 CHALLENGE - TASK 5: MONITORING OF DOMESTIC ACTIVITIES BASED ON MULTI-CHANNEL ACOUSTICS

[2] Veralia Gabriela Sanchez Carlos F. Pfeiffer and Nils-Olav Skeie: A Review of Smart House Analysis Methods for Assisting Older People Living Alone

[3] Lode Vuegen, Toon van Waterschoot, Bart Vanrumste, Peter Karsmakers - DCASE 2018 CHALLENGE - TASK 5: MONITORING OF DOMESTIC ACTIVITIES BASED ON MULTI-CHANNEL ACOUSTICS

[4] Tadanobu Inoue, Phongtharin Vinayavekhin , Shiqiang Wang , David Wood , Nancy Greco , Ryuki Tachibana - DOMESTIC ACTIVITIES CLASSIFICATION BASED ON CNN USING SHUFFLING AND MIXING DATA AUGMENTATION

[5] F. Erden, S. Velipasalar, A. Z. Alkar, and A. E. Cetin, “Sensors in Assisted Living: A survey of signal and image processing methods,” IEEE Signal Processing Magazine, vol. 33, no. 2, pp. 36-44, Mar.2016.

[6] M. Yu, A. Rhuma, S. M. Naqvi, L. Wang, and J. Chambers, “A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment,” IEEE Trans. Inf. Technol. Biomed., vol. 16, no. 6, pp. 1274-1286, 2012.

[7] N. K. Suryadevara, S. C. Mukhopadhyay, “Determining wellness through an ambient assisted living environment,” IEEE Intell. Syst., vol. 29, no. 3, pp. 30-37, 2014.

[8] E. Principi, P. Olivetti, S. Squartini, R. Bonfigli, F. Piazza, “A floor acoustic sensor for fall classification,” in Proc. 138th Audio Engineering Soc. Convention, pp. 949-958, 2015. [5] G. A. Koshmak, M. Linden, A. Loutfi, “Evaluation of the android-based fall detection system with physiological data monitoring,” in Proc. Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society, pp. 1164-1168, 2013

[9] K. Ozcan, A. K. Mahabalagiri, M. Casares, and S. Velipasalar, “Automatic fall detection and activity classification by a wearable embedded smart camera,” IEEE J. Emerg. Sel. Top. Circuits Syst., vol. 3, no. 2, pp. 125-136, 2013.

[10] A. K. Bourke, S. Prescher, F. Koehler, V. Cionca, C. Tavares, S. Gomis, V. Garcia, and I. Nelson, “Embedded fall and activity monitoring for a wearable ambient assisted living solution for older adults,” in Proc. Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society, pp. 248-251, 2012.

[11] Jeremy Chew, Yingxiang Sun, Lahiru Jayasinghe, Chau Yuen, DCASE 2018 CHALLENGE: SOLUTION FOR TASK 5

[12] Gierad Laput, Yang Zhang, Chris Harrison: Synthetic Sensors: Towards General-Purpose Sensing

[13] Gierad Laput Karan Ahuja Mayank Goel Chris Harrison: Ubicoustics: Plug-and-Play Acoustic Activity Recognition

[14] Gierad Laput, Walter S. Lasecki, Jason Wiese, Robert

[15] Xiao, Jeffrey P. Bigham, and Chris Harrison. 2015. Zensors: Adaptive, Rapidly Deployable, Human-

Intelligent Sensor Feeds. In Proceedings of the 33rdAnnual ACM Conference on Human Factors in Compu-

ting Systems (CHI '15). ACM, New York, NY, USA,1935-1944. DOI:

[16] Jamie A. Ward, Paul Lukowicz, Gerhard Troster, and Thad E. Starner. 2006. Activity recognition of assembly tasks using body-worn microphones and accelerometers. In IEEE transactions on pattern analysis and machine intelligence, 28, no. 10: 1553-1567.

[17] Alexander Bertrand: Applications and trends in wireless acoustic sensor networks: a signal processing perspective

[18] Veralia Gabriela Sanchez Carlos F. Pfeiffer and Nils-Olav Skeie: A Review of Smart House Analysis Methods for Assisting Older People Living Alone

**Useful Links**

<http://www.figlab.com/#/ubicoustics/> (1)

<https://www.youtube.com/channel/UC1TLWwu-sGVuN51AhTjeSZg> (2)