<Comparative Study of Passive and Active Acoustic Sensing for Indoor Room Recognition>

<Michael Chan[1]>

Supervisor: <Dr. Qun Song[1]>

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: <Michael Chan>
Final project course: CSE3000 Research Project
Thesis committee: <Dr. Qun Song>, <J.A. Martinez Castaneda>

Abstract

The ability to accurately determine the location within indoor settings is crucial for various applications such as indoor navigation, interactive floor plans, and room-specific services. While GPS technology has revolutionized outdoor positioning, it falls short in providing precise location information within buildings due to signal blockage. To address this limitation, specialized indoor positioning systems utilizing acoustic sensing have been explored, leveraging deep learning models. This paper presents a comparative study of passive and active acoustic sensing systems for room recognition.

Passive sensing involves capturing existing background noise in a room and using it as an identifier, while active sensing emits acoustic signals and analyzes the resulting echoes. Previous research has primarily focused on active sensing, achieving high classification accuracy but facing challenges related to device orientation and the presence of multiple individuals. Moreover, the emission of high-frequency chirps used in active sensing may cause discomfort to pets.

The results indicate that passive sensing achieves an accuracy of 73.7%, slightly outperforming active sensing at 63.5% in baseline conditions. However, in the presence of constant background noise, passive sensing accuracy drops to 21.7%, while active sensing exhibits better resilience with an accuracy of 59.7%. Furthermore, when the device orientation is altered by 90 degrees, active sensing results in a lower in accuracy (45.5%), while passive sensing maintains better performance at 71%. The impact of multiple individuals in the room had a relatively minor effect on passive sensing systems, achieving an accuracy of 72.2%. Active sensing was shown to be not as resilient, reaching an accuracy of 44.2%

1 Introduction

Since the introduction of smartphones, determining one's location through GPS has become more accessible than ever before and is currently an indispensable feature of a mobile phone. While GPS works well to determine the general location in outdoor areas, it leaves much to be desired in indoor settings where GPS signals are blocked by the building itself, providing insufficient accuracy to differentiate between different rooms and floors[5].

The recognition of rooms in a building has countless useful applications ranging from indoor tours to interactive floor plans. In such scenarios specialized indoor positioning systems are required. Recently, various research has been conducted to solve this problem by using the acoustic systems of the smartphone in conjunction with a deep learning model. In such systems acoustic data is transformed into a spectrogram and then used as a fingerprint for classification by the model.

Acoustic sensing can be performed either passively or actively. Passive sensing collects the existing background noise that is present a room and uses that as the identifier. Active sensing on the other hand involves sending acoustic signals of which the echoes are than captured by the device. The spectrogram of the echoes is than used as a fingerprint to identify the room.

2 Related work

Indoor positioning systems using acoustic sensing have gained significant attention in recent years due to their potential for accurate infrastructure-free indoor localization. Numerous studies have explored the use of both passive and active acoustic sensing techniques for room recognition and indoor positioning. In this section, the key findings and approaches from relevant literature are discussed.

Recent work that is most similar to this research include ”Batphone”[9] and ”RoomRecognize”[8]. Batphone uses a passive acoustic sensing system and listens to the acoustic background spectrum in a 0-7kHz frequency band for 10 seconds. Fingerprints are extracted from a power spectrum. To reject transient noises the data points in the power spectrum are sorted by ascending magnitude and only the 5th percentile of each frequency bin is considered for the fingerprint. Batphone achieves an classification accuracy of 69%. Narrower low frequency bands achieved greater accuracy's in the presence of noise but performed worse in silent classification. Classification is done by a nearest neighbour classifier [9]. RoomRecognize and utilizes an active sensing system and relies on a Convoluted Neural Network (CNN) for classification. RoomRecognize emits inaudible chirps for a duration and frequency of 2ms and 20kHz respectively every 100ms and records the echoes in a 19.5kHz to 20.5kHz frequency band, collecting 97.5ms worth of usable data every interval. RoomRecognize achieved accuracy's of over 99% in silent environments and was shown to be significantly more robust than Batphone in the presence of background noise[8].

SoundSignature is a passive acoustic fingerprinting model that utilizes Support Vector Machines (SVM) for room recognition [6]. Similarly to Batphone, SoundSignature extracts the fingerprint from the 5th percentile of the power of each frequency in a spectrogram sorted by magnitude. The duration of each sample is 5 seconds. The features considered include the Mel Frequency Cepstral Coefficients (MFCC) as well as centroid, spread, skewness, kurtosis, slope, decrease and roll-off. Sequential Forward Feature Selection was used to select smaller subsets of the feature space. SoundSignature was able to achieve a 10-fold cross validation score of 90.28% in classifying 16 rooms but was only able to achieve an accuracy of 48.08% on new data collected the next day.

Similarly to SoundSignature, RoomSense utilizes the SVM classifier with MFCC features. RoomSense however, uses active acoustic sensing [7]. RoomSense emits 0.68s long Maximum Length Sequences (MLS) signals, and the echoes are then recorded in a frequency band of [0-24]kHz. RoomSense was able to realize an accuracy of 98% across 20 different rooms. MFCC features were found to be the most effective among various other feature extraction methods for SVM classifcation. RoomSense was shown to be fairly noise robust, achieving a 66.6% accuracy with a signal to noise ratio of 10 dB.

Other MLS-based active acoustic fingerprinting include SoundLoc[4]. SoundLoc combines several features obtained from the echo responses such as the kurtosis, spectral standard deviation and reverberation time. SoundLoc was tested with multiple classifiers, more specifically Multilayer Perceptron (MLP), Random Forrest (RF), Logistic Regression, Naive Bayes, J48, RBF Network and RIDOR. MLP and RF were found to be the most efficient, achieving a 95.33% and 91.67% accuracy respectively with only 10 samples. SoundLoc was able to classify 10 rooms with 97.8% overall accuracy with 1000 samples. SoundLoc was also tested for noise robustness. The data of one of the 10 quiet rooms was exchanged for noisy data. SoundLoc was found to be noise robust after feature re-selection, resulting in a 98% accuracy for classifying the noisy room. Without re-selection however, the accuracy was only 24%.

Problem statement

While many advances were made in room classification using active sensing, passive sensing has not been as fruitful.

”RoomRecognize” achieved excellent accuracy’s but was found to be susceptible to the orientation of the device and to the fluctuation of the number of people in the room [8]. Additionally ”RoomRecognize” uses active sensing, emitting 20kHz chirps with a high volume. While not audible to the human ear, dogs were found to perceive frequencies around 20kHz as especially loud [3], making active sensing solutions undesirable in the presence of pets.

Passive acoustic fingerprinting models that use deep learning methods are yet to be explored. This paper will contribute a CNN-based classifier for both passive and active acoustic fingerprints and explore under which circumstances one sensing mode could be preferred over the other.

In this paper 4 different conditions will investigated. These include a baseline silent condition, a noisy condition, orientation change of the device and having multiple individuals present in the room. A detailed description of the conditions can be found in section 4.1.

3 Measurement Study

Research has been performed to support the study incentive, this section will cover the findings and elaborate on their significance to the study.

3.1 Active sensing responses

Previous works have shown that acoustic data serves as an effective fingerprint for the identification of rooms in both passive and active sensing modes [8], [9]. Measurements were performed and were shown to be consistent with previous findings. Figure 1 and 2 show spectrograms of active sensing measurements taken in room A and B respectively. They are consistent when measured in the same rooms while being distinctive between different rooms.

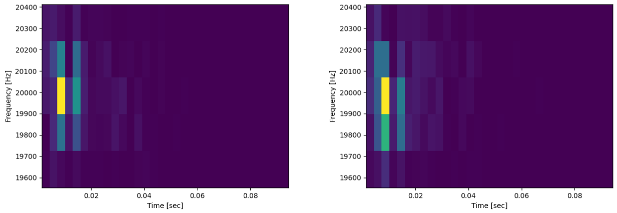

Figure 1: Two Active sensing spectrograms of room A

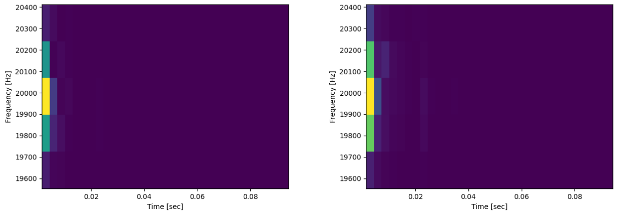

Figure 2: Two Active sensing spectrograms of room B

3.2 Passive sensing responses

When comparing passive sensing fingerprints on the other hand, consistency is not as visibly present in the spectrograms of 0.1 second samples. A small scale test was performed between 4 different rooms with 500 samples each, which resulted in an accuracy of 70% after 50 epochs, after which the accuracy drops down to 61%. The result can be seen in figure 3. While it is still significantly better than random guessing, the accuracy is unsatisfactory. Longer samples seem to be desired. Research has shown increasing performance with longer sample length [9].

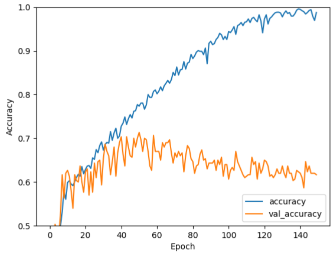

Figure 3: Passive sensing performance across 4 rooms

3.3 The effect of background noise

Active sensing was shown to be a lot more robust than passive sensing in the presence of background noise in [8]. The noise robustness of active sensing was clearly visible in the spectrogram responses in figure 4. Since most common noise is of lower frequencies, it does not significantly impact active sensing as all frequencies outside of the small high frequency band are filtered out.

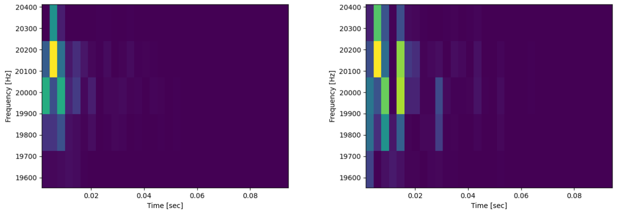

Figure 4: Active spectrogram responses in silent (left) and noisy (right) conditions

3.4 The effect of orientation of the device

The orientation of the phone was found to have an effect in the performance of active sensing [8]. Changes on the spectrogram when the phone is rotated 90 degrees from its original orientation can be observed in figure 5. This is likely due to the change in the echo path. Since passive sensing does not rely on such echoes directly the hypothesis is that passive sensing is more robust to changes in orientation.

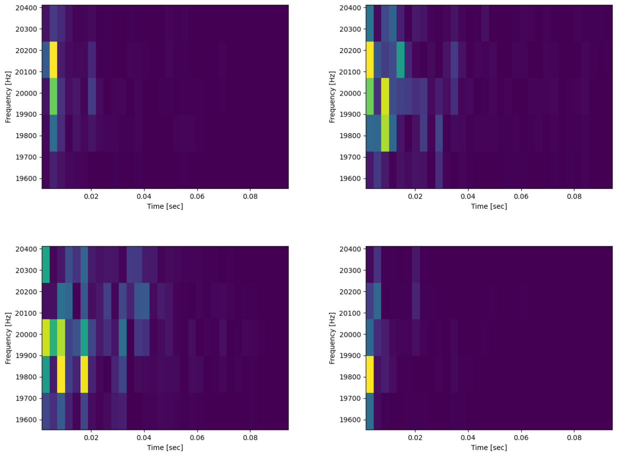

Figure 5: Active sensing spectrograms responses with 0°(top-left), 45°(top-right), 90°(bottom-left) and 180°(bottom-right) rotation.

3.5 The effect of multiple people in the room

It was shown that having multiple people present in a room could affect the results of active sensing [8], presumably due to the human body absorbing echoes. Human bodies were found to absorb low frequency sounds as well with an increasing absorption rate of as the frequency increased [1]. This could impact both performance of passive and active sensing. A test was performed to explore the effects of the number of individuals in the room. Figure 6 shows the effect on the spectrogram when multiple individuals are present in the room.

4 Methodology

In order to compare passive sensing with active sensing both models were implemented and the results of the models were then compared under various situations. The steps taken can be summarised as follows:

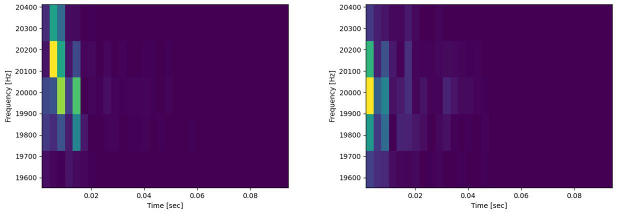

Figure 6: Active spectrogram responses with one person (left) and 3 people (right) present in the room

1. An application supporting the two sensing modes was created. Passive mode exclusively records audio, while active mode emits high frequency chirps and records audio at the same time.
2. A preprocessing pipeline was created to transform and filter data according to the sensing mode to obtain spectrograms which were used as input for the deep learning model.
3. Data was collected for both sensing modes for training and testing of their respective models.
4. The models were tested in various conditions.

4.1 Testing baseline

To find out in which situations passive sensing is favourable over active sensing and vice versa, both sensing modes had to be tested in various conditions. The conditions explored in this study include:

- A baseline in an optimal condition which is most similar to the condition in which the training data was collected in. This implies a silent condition a single person in the room with the phone located in the same position and orientation.
- A condition in which constant acoustic background noise is present. With everything else being consistent with the baseline condition.
- A condition in which multiple people are present at the same time in the room, including one person standing between the microphone and the wall. With everything else being consistent with the baseline condition.
- A condition in which the orientation of the phone is turned 90 degrees around the vertical axis. With everything else being consistent with the baseline condition.

4.2 Evaluation metric

For each sensing mode and condition a confusion matrix was created and the accuracy is calculated as the number of correctly predicted labels over the total number of predictions in that respective condition. By comparing the baseline accuracy's of the two sensing models it was possible to determine which model performs better under ideal circumstances. By comparing the accuracy's of the two sensing modes in suboptimal conditions with their respective baseline accuracy the effect of the different interference conditions were measured and compared.

5 Implementation

This section covers the implementation of the model and the data collection and processing pipeline.

5.1 Data collection

All data was collected in a single building with 6 different rooms on a Pocophone X3 Pro with maximum volume settings. Audio was sampled at a frequency of 44100Hz with a duration of 100ms and active sensing chirps were emitted at a 20000Hz frequency for 2ms as done in [8]. The device was held with a fully stretched arm as far away from the body as possible with the speaker and microphone pointed away from the user. 500 samples of 0.1 seconds each were collected per room for active sensing. Passive sensing used 1 second long samples, this was balanced by reducing the sample size to 50 per room, keeping the total recording duration consistent at 50 seconds. Testing consisted of a sample size of 100 for each room per special condition. Passive and active sensing data was collected separately to avoid potential interference on passive sensing produced by the chirps.

5.2 Preprocessing

Audio data was transformed into spectrograms using hann windows of size 256 with 128 overlapping points as in [8]. Narrow-band frequencies were used for both passive and active sensing. Passive sensing filtered out all frequencies outside of a [0-1000]Hz range. Active sensing only considered frequencies in the [19500-20500]Hz range. This resulted in an image of 5x32 pixels and 5x342 pixels for active and passive sensing respectively.

5.3 Convolutional neural network

A convolutional neural network is used for classification. The model consists of 7 layers as described in [8]. This includes:

1. A convolutional layer of 16 4x4 filters with a stride of 2 and padding to retain the same size.
2. A max pooling layer with a 2x2 filter.
3. A second convolutional layer of 32 4x4 filters with a stride of 2 and padding to retain the same size.
4. Another max pooling layer after which the output is flattened.
5. A dense layer with 1024 nodes and a dropout factor of 0.4.
6. A dense layer with 6 nodes corresponding to the number of rooms.
7. An output layer with a softmax activation function.

6 Results

This section will cover the findings of the experiments. This includes the baseline performance, performance in noisy conditions, the effect of device orientation as well as the impact of the presence of people in the room. A comprehensive table of the results can be found below in table 1. Confusion matrices can be found in appendix A and appendix B

Testing Environment	Active Sensing	Passive Sensing
Baseline	0.6350	0.7367
Noisy Environment	0.5967	0.2167
Orientation Change	0.4550	0.7100
Presence of people	0.4417	0.7218

Table 1: Accuracy of active and passive sensing in the classification of 6 different rooms.

6.1 Baseline performance

The passive and active sensing classifiers performed similarly in classifying new data with passive and active sensing reaching an accuracy of 73.6% and 63.5% respectively. Both models performed extremely well in distinguishing the closet room from the other rooms, likely due to it being considerably smaller in size. The active sensing model was able to classify the living room but was unable to recognize the master bedroom. The passive sensing model was able to recognize rooms more consistently but had some difficulties in distinguishing the master bedroom from the living room.

6.2 Noisy environment

In a noisy environment passive sensing was unable to recognize rooms. Achieving an accuracy of 21.7%, only slightly ahead random guessing. Further proving the unsuitability of passive sensing in noisy conditions. Active sensing seemed largely unaffected by the noise and was still able to reach an accuracy of 59.7%, less than a 4% point drop from the baseline.

6.3 Device orientation

When the device was turned 90 degrees counterclockwise both sensing methods lost considerable performance. Active sensing reached an accuracy of 45.5% while passive sensing was able to reach 71.0%. An 18% and 1.5% point drop respectively from their baselines. Passive sensing was shown to be very robust to any orientation change of the device. While active sensing was not as robust to orientation change, the performance gap was smaller than predicted and the change did not result in a drastic accuracy loss as was seen in the background noise experiment with passive sensing. What is notable however, is the high consistency of active sensing in classifying the new data. It was able to clearly distinguish between the new data but failed to label the hallway correctly due to the orientation change. This is likely due to the shape of the hallway, where an 90°orientation change has a significant impact on the distance between the microphone and the wall.

6.4 Presence of multiple individuals

The effect of the presence of multiple individuals in a room had a relatively small impact on passive sensing. The presence of people and blocking of the signal with the body had a moderate impact on active sensing performance. It significantly increased the likelihood of the master bedroom being predicted. A potential cause for this is the higher sound absorption profile of the master bedroom due to the large bed present in the chamber, absorbing a large amount of the echoes. Passive sensing had no clear visible change in classification behaviour and was overall unaffected.

7 Responsible Research

This section covers the actions taken with regards to the ethical and conscientious conduct of the scientific investigation.

7.1 Ethical Considerations

This research collected audio data in a private residence. In such scenarios privacy of all parties involved is of upmost importance. Prior to the collection of audio data, informed consent was obtained from each resident for the use of audio data recorded in the building. Additionally the audio data will not be made public as an additional precaution.

7.2 Reproducibility

For reproducibility, data integrity is of upmost importance. The collection as well as the handling of data was consistent with the procedure outlined in the methodology and implementation sections. The source code is also made publicly available to aid in the reproducibility of the study [2].

8 Discussion

The results have shown that passive and active sensing have their own set of challenges. Active sensing was found to be significantly more robust to noise compared to passive sensing. However, it was more susceptible to the orientation of the device and change in the number of people present in the environment. Active sensing was also able to classify rooms with 10 times smaller data samples, resulting in a higher degree of privacy as well as faster classification. The overall prediction behaviour was more consistent across samples with the same label.

Passive sensing was shown to be more robust to change in environment while no additional sounds are introduced. Noise had a large impact on the prediction capabilities of the model and could potentially be alleviated through more thorough data preprocessing.

While the overall accuracy of passive sensing in the baseline condition was greater than active sensing in this experiment, this could be attributed to hardware and data preprocessing flaws. Additionally, time invariance was not considered in this experiment. Previous works have shown great decreases in passive sensing performance with testing data captured at a later point in time [9], [6]. Nevertheless, the relative performance of the model can still provide valuable insights.

9 Conclusions and Future Work

Classification through passive and active acoustic sensing are both shown to be a feasible method of room recognition. Both models respond differently to changes in environment, with active sensing being more sensitive to geometrical changes but noise robust and passive sensing being highly susceptible to noise but robust to any geometrical changes.

The models were shown to have different prerequisites. Active sensing systems requires environments in which the emission of high volume ultrasound is permissible. Passive sensing requires longer audio samples to function which can lead to privacy concerns.

Future considerations for both sensing methods may include the fusion of both sensing systems. As both methods were shown to have different strengths and weaknesses, both systems could be able to complementary to each other. As both systems are infrastructure free and audio data recording can theoretically be performed for both models at once, little overhead is introduced by doing so, given that their prerequisites are met. Such a combined system could prove to be more robust overall to changes in environment.

A Confusion Matrices Active Sensing

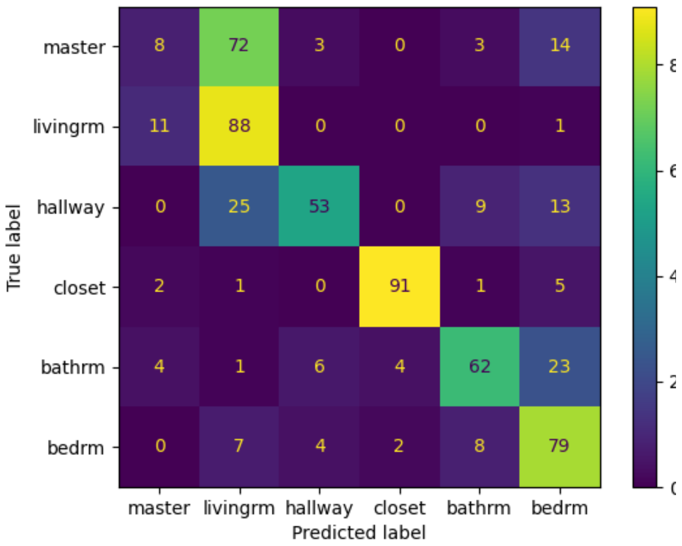

Figure 7: Confusion matrix active sensing baseline

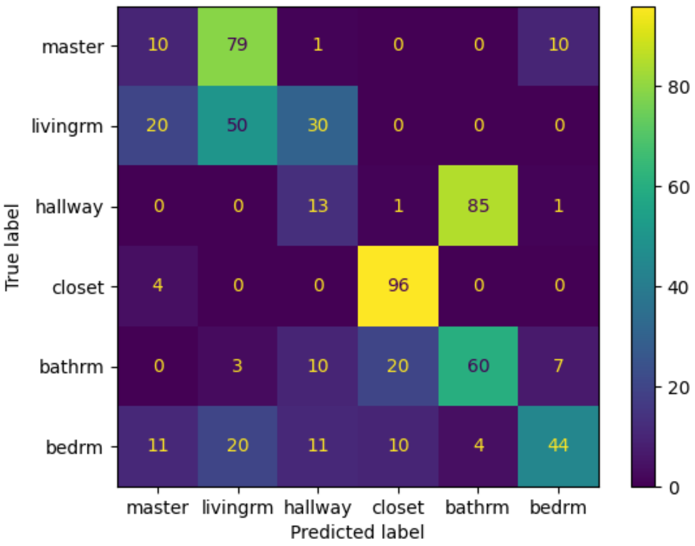

Figure 9: Confusion matrix active sensing rotated

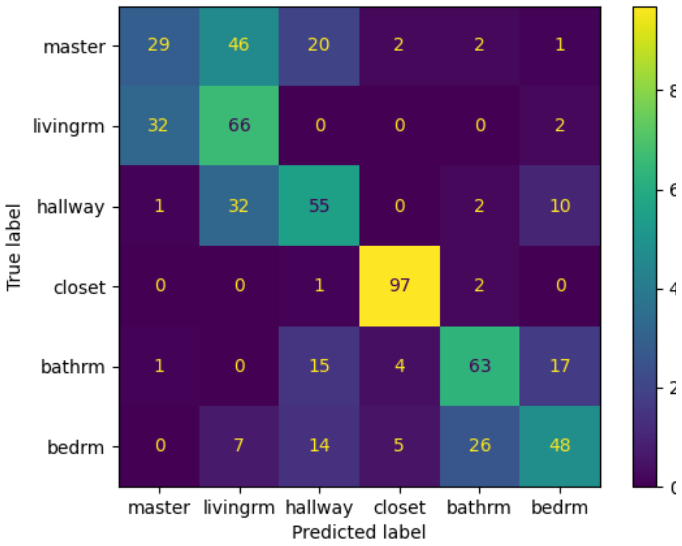

Figure 8: Confusion matrix active sensing noisy

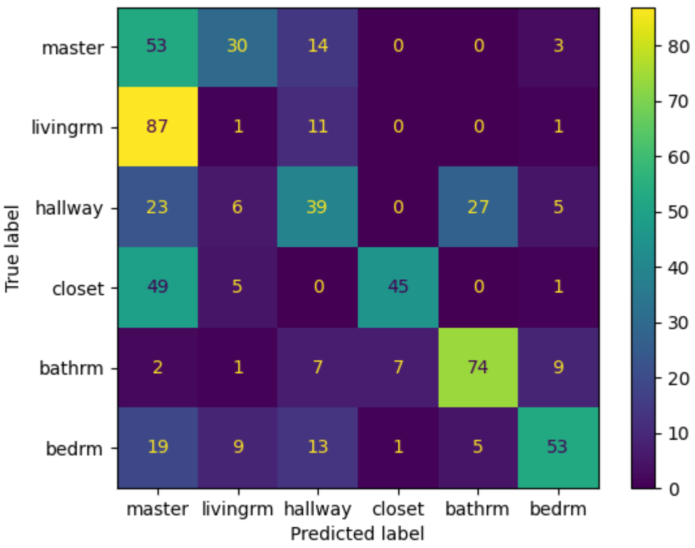

Figure 10: Confusion matrix active sensing multiple individuals

B Confusion Matrices Passive Sensing

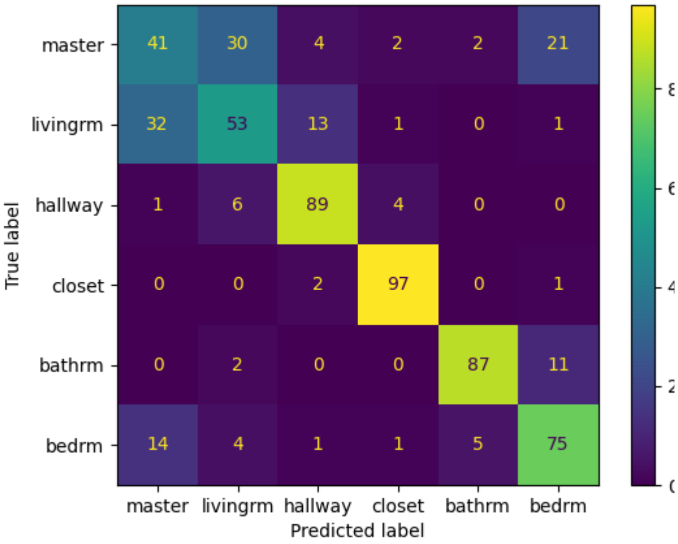

Figure 11: Confusion matrix passive sensing baseline

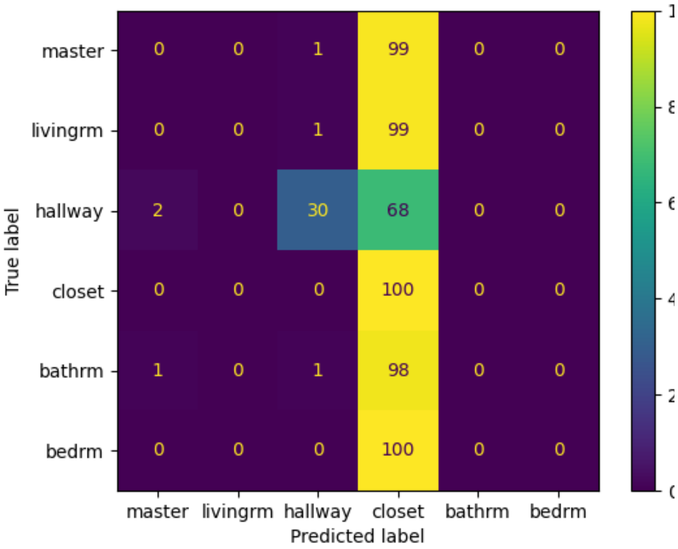

Figure 12: Confusion matrix passive sensing noisy

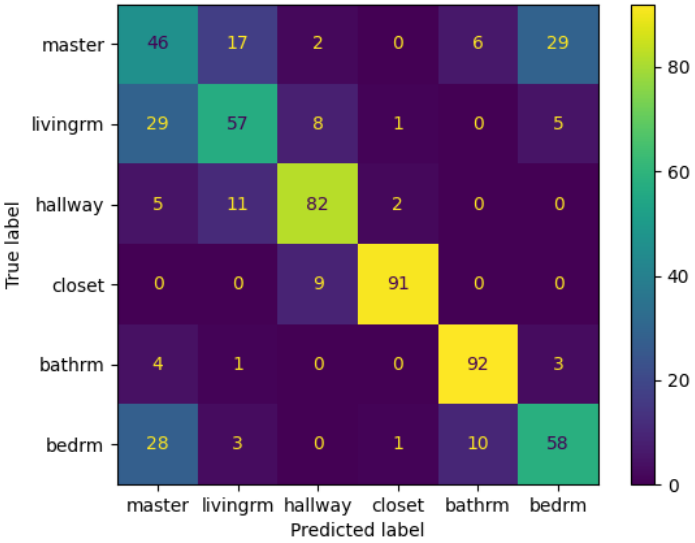

Figure 13: Confusion matrix passive sensing rotated

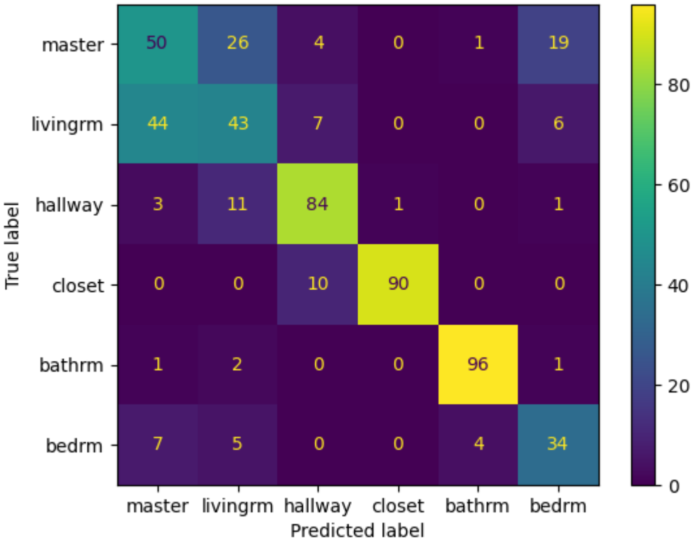

Figure 14: Confusion matrix passive sensing baseline multiple individuals

References

[1] N Chagok, Dinfa Domtau, E Agoyi, and E Akpan. Average sound absorption per person at octave band frequencies between 125hz and 4000hz in an enclosure. *Journal of Natural Sciences Research*, 01 2013.

[2] M. Chan. Source code repository. https://github.com/MichaelChan20/ResearchProject.

[3] Henry Heffner. Hearing in large and small dogs: Absolute thresholds and size of the tympanic membrane. *Behav Neurosci*, 97:310–318, 04 1983.

[4] Ruoxi Jia, Ming Jin, Zilong Chen, and Costas J. Spanos. Soundloc: Accurate room-level indoor localization using acoustic signatures. In *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, pages 186–193, 2015.

[5] Mikkel Kjærgaard, Henrik Blunck, Torben Godsk, Thomas Toftkjær, Dan Lund, and Kaj Grønbæk. Indoor positioning using gps revisited. pages 38–56, 05 2010.

[6] Ricardo Leonardo, Marilia Barandas, and Hugo Gamboa. A framework for infrastructure-free indoor localization based on pervasive sound analysis. *IEEE Sensors Journal*, 18(10):4136–4144, 2018.

[7] Mirco Rossi, Julia Seiter, Oliver Amft, Seraina Buchmeier, and Gerhard Tröster. Roomsense: An indoor positioning system for smartphones using active sound probing. pages 89–95, 03 2013.

[8] Qun Song, Chaojie Gu, and Rui Tan. Deep room recognition using inaudible echos, 2018.

[9] Stephen Tarzia, Peter Dinda, Robert Dick, and Gokhan Memik. Indoor localization without infrastructure using the acoustic background spectrum. pages 155–168, 06 2011.