Optimal data capturing for indoor location sensing

Roald van Heerde[1]

Supervisor(s): 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: Roald van Heerde
Final project course: CSE3000 Research Project
Thesis committee: Qun Song, Jorge Martinez Castaneda

Abstract

In today's world, accurate location sensing is impossible to think away. One of the most prominent and most used techniques for determining location is GPS. In the outside world, GPS is capable of pinpointing a location with only a few meters error. But inside buildings, GPS often fails to deliver the same accuracy. In this paper, a relatively new technique will be presented to solve this problem using acoustic location sensing where a smartphone emits inaudible chirps and records the result. Specifically, this paper will cover what kind of data is needed to train the deep model that will solve this problem.

1 Introduction

Accurate indoor location sensing using the smartphone acoustic system becomes more important everyday. In this paper two key questions are addressed: what is the optimal data set for training the deep model and how should this data be processed to achieve the best results. Just like whales and bats use echolocation to determine their location and the location of objects around them, a smartphone can use its acoustic system (speaker and microphone) to emit short, inaudible chirps akin to sonar signals. The recorded result can then be used to determine its location.

Take for example large buildings with a large amount of rooms, like a hospital, where swift and precise location sensing is vital in case of emergencies. Or a museum where location sensing can be used to improve the responsiveness of an audio tour application. In these situations, acoustic location sensing can provide swifter and more accurate results than existing technologies such as GPS and WiFi. Building on the concept that each room will have its own unique acoustic fingerprint, a deep model can be used in classifying room locations.

Existing studies [1; 2; 4; 5; 6; 8; 10; 11; 12; 13] use the collection of spectrograms as the basis for the training data. However, using spectrograms is only one of several options available for feature extraction that can be used to train a deep model. In this paper, alternative strategies are investigated for efficient processing of acoustic signals. The goal is to find methods that can deliver the highest accuracy in recognizing different locations.

2 Related work

While this technique is relatively new, a lot of papers are already out there with a variety of attempts at creating an implementation for location sensing applications. Applications like RoomRecognize [11] and EchoLoc [10] manage to successfully implement an application that can near perfectly determine which location the smartphone is. The research by B. Zhou et. al. [13] shows that chirps can even be used to create floorplans. This indicates that there is sufficient data within chirps and their corresponding echoes to extract detailed information about a room and the location within that room. However, while chirps seem to be extremely handy in location sensing, S. P. Tarzia et. al. [4] show that just using passive location sensing without using any chirps also shows information about the location that can be used to create a classifier.

While standard spectrograms are often used for location sensing, it is not the only feature extraction technique that has been applied. The research by P. Seetharaman et. al. [7] uses chromagrams to identify cover songs. While this does not directly apply to acoustic location sensing it is an interesting view on audio recognition that might aid in the recognition of different locations. Mfcc has also been used for audio recognition as described by C. Ittichaichareon et. al.[9]. Here, speech recognition is investigated using Mel-Scale Frequency Cepstral Coefficients (MFCC) showing that it can also be used to identify different audio features. This paper will attempt to find out which of these techniques is most fit for recognizing different locations across multiple rooms.

3 Measurement Study

Many existing implementations employ spectral features to train machine learning models for location or room recognition. This approach is logical because the response time and response intensity are key factors for identifying a room with chirps. Spectrograms effectively display this data, providing a comprehensive view of both temporal and frequency information. Therefore they are well suited to train a machine learning model since machine learning models (especially deep models) have the ability to recognize subtle patterns that can be found within these spectrograms. In this paper, other techniques are investigated to show whether spectrograms are the most efficient choice or whether there might be a way to improve the accuracy of the deep models.

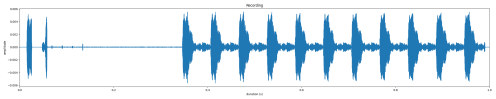


Figure 1: Example of a classic time graph. Here the recording can be seen including multiple chirps after a high pass filter has been applied. The area between the chirps show smaller peaks that correspond to the echos that are produced by these chirps.

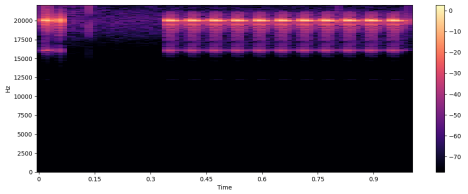


Figure 2: Example of a spectogram of the same recording with same preprocessing.

Figure 1 shows a time graph of a recording. The larger peaks correspond to the chirps that are directly recorded by the phone's speaker while the smaller peaks that can be seen between each of the larger peaks, correspond to the echoes

received by the speaker. This data specifically is what will be used to train a deep model to recognize different locations. Figure 2 shows more detail about the specific chirp response than figure 1 since figure 2 shows the difference across multiple frequencies. This is a key advantage in training a deep model since figure 2 contains more direct information about the echoes that the chirps produce.

In this paper, 6 different feature extraction techniques (FET) will be investigated. Appendix A shows an example image of each technique. Figure 7 and 8 show an example of a spectrogram capped between 19.5k Hertz and 20.5k Hertz where the spectrogram in figure 8 is created with its values converted to a linear scale.
Figure 9, 10 and 11 show examples of a chromatogram, mel frequency ceptrogram and a mel-scaled spectrogram respectively. In these examples, a 2 second audio recording is used with 2ms chirps of 20k Hertz. The mel-scaled spectrogram shows the clearest image which is a promising feature in recognizing rooms. The chromagrams and mfcc images show a vastly different image compared to the other FET's which could either mean that their performance will probably be much better or much worse than the other FET's.

4 Methodology

To investigate different types of feature extraction, 7 different locations will be chosen to run the same experiment for each of the feature extraction techniques. Using chirps with a duration of 2ms and an interval of 100ms, a set of images will be created for every room and processed using each FET.

- First, the data is collected for a duration of 50 seconds twice to gain approximately 1000 samples.
- The data is then saved 6 times using different FET's combined with different preprocessing
- During recording, the smartphone will be held in the same position without rotating and moving it too much. This is done to make sure the recordings from every location has the same recording quality.
- After all the data is collected and processed, the same deep model is created 6 times using each instance of the saved data
- Finally, the results will be stored in a history graph generated during the training of the deep mode and a confusion matrix showing how well the model performs on new input data.

The confusion matrices are created by feeding new data into the trained model and comparing each prediction with the actual label. This test data is created from the initial training data. To create this test set along with the other data sets, the initial training data is divided into three parts.

- **The training set** consisting of 80% of the training data. This is the data that is used to train the deep model
- **The evaluation set** consisting of 10% of the training data. This data set is used to evaluate the trained model.
- **The test set** consisting of 10% of the training data. This set will be used for the confusion matrices

5 Implementation

The application follows a client-server architecture where the client is developed as an Android application using Java and Android Studio. It handles tasks such as emitting and recording audio data, as well as sending and receiving data to and from the server.

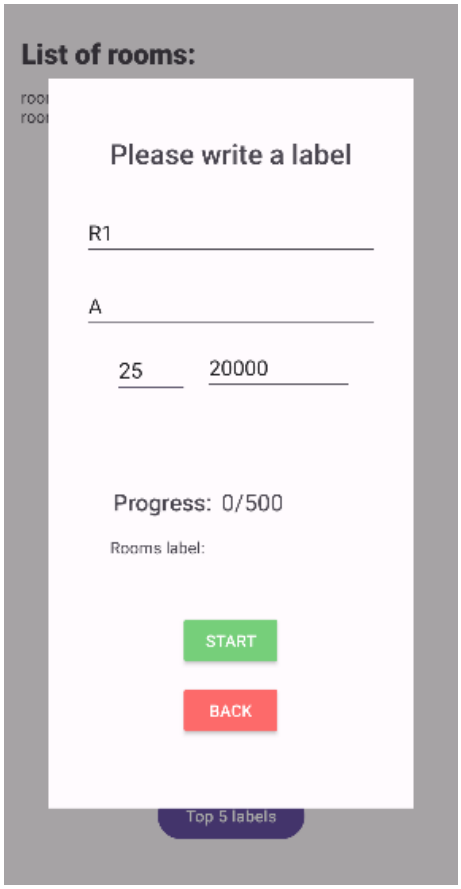


Figure 3: Screenshot showing a part of the Android Application with the options to set the location (room and location), duration and frequency of the recording.

The chirps consists of a sine wave with a frequency of 20k Hz. To limit clicking from the speakers, a window is applied to smooth out the chirps as described by B. Zou et. al. [13]. A frequency of 20k is used since this frequency is inaudible to the human ear and still within the frequency range that a smartphone can emit and record. The server is implemented using the Python Flask library. It receives the audio data from the client and applies different preprocessing techniques depending on the iteration of the experiment. The data is split up to gain individual fragments containing the response period (around 95ms) of each chirp. This is done by calculating the enveloped function and retrieving the largest peaks from that same enveloped function. These peaks are then used to chop up the recording as shown in Figure 4

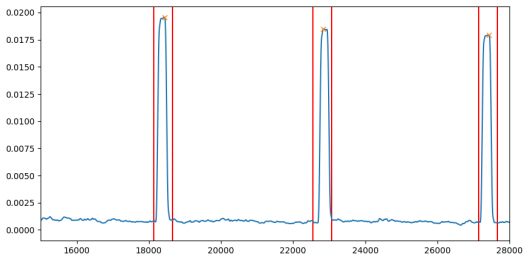


Figure 4: This image shows how the sound fragment is divided. The orange x shows the identified peak while the red lines show how the individual fragments are created.

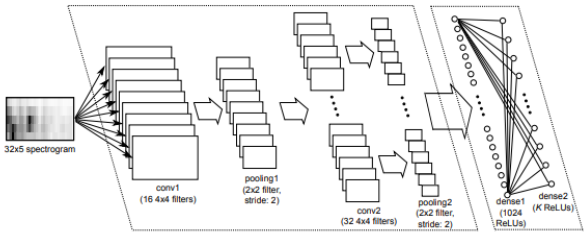


Figure 5: Image by Q. Song et. al. showing the CNN

Using the Librosa library, each window is then processed using the following methods:

- *librosa.stft*.This produces a standard spectrogram as used in experiments like RoomRecognize [11]. Two versions are created using this technique: one raw version and one version where each value is converted to a linear scale.
- *librosa.mfcc*.This creates a spectrogram using Mel-Scaled Frequency Cepstral Coefficients.
- *librosa.chroma_stft*. This creates a chromagram. For this technique two versions were made as well. One version where the input data has a high pass filter applied to it to cut out all lower frequencies and one version without filter.
- *librosa.melspectrogram*. This creates a Mel-Scaled Spectrogram.

To keep the spectrograms relevant. All spectrograms are capped between 19.5k Hertz and 20.5k Hertz except for one instance of the chromagrams since the chromagrams display pitch value and not frequency values.

Each resulting spectrogram or chromagram is then converted to a 32x5 black white image as described by Song et. al. [11]. For the deep model, a Convolutional Neural Network is implemented using the Tensorflow and Sklearn libraries. This model is implemented using the following parameters:

- A 2D convolution layer with a 16 4x4 filters
- A pooling layer with a 2x2 filter and 2 strides
- A second convolution layer with 32 4x4 layers
- A second pooling layer with a 2x2 filter and 2 strides
- A Dense layer with 1024 ReLu nodes
- And finally another dense layer containing k nodes corresponding to k target locations

Tensorflow is used to implement the model (and therefore also train and evaluate it) and the Sklearn library is used to prepare the training data. Each location contains around 1000 images. These images are collected and then split into 3 data sets as desribed in section 4.
Since the result of every feature extraction method is converted to a 32x5 image, every method can be fed to the same model.

6 Experimental Setup and Results

6.1 Setup

The objective of the study was to determine the most effective processing techniques and data for recognizing different locations and rooms. For this purpose, a local home was chosen as the test environment. The experiment involved re-running the same experiment using various techniques as described earlier. By comparing the results of these experiments, more can be learned about what techniques affect the performance of the deep models.

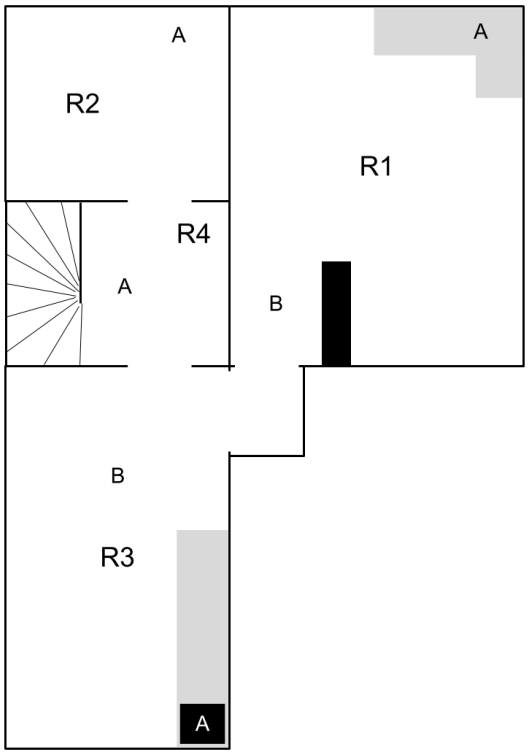


Figure 6: The floorplan of containing the locations that were used in the experiments

Figure 6 shows most of the locations that were tested. Here, R1 is a bedroom with location A being the desk and location B being behind the door between a closet and the wall. This givens location A a more open space while location B is more enclosed. Room R2 is the bathroom with only one location A. The bathroom has much more echo than the other rooms which should result in a stronger echo response for every chirp. Room R3 is a longer room where location A is in the corner and location B is in a more open space similar to location A in room R1. Room R4 is the only room without any windows and thus completely enclosed between walls except from the stairs that lead to another floor above and beneath room R4. R5 is located one floor above the floor described in figure 6. This location was chosen since the roof has a slope which differentiates it from the other locations. These location were chosen to explore different environmental aspects such as reverb and room size while also testing the deep model's ability to distinguish locations that are more similar to each other such as locations R3 B and R1 A.

6.2 Results

Appendix B shows the validation accuracy per epoch and a confusion matrix showing the performance after training.

When looking at the results, an order of performance can be created among each FET:

1. The Mel-Scaled spectrograms
2. The standard spectrograms
3. The standard spectrograms where the values were converted to a linear scale
4. The chromagrams where the input data was filtered using a high pass filter
5. The chromagrams where the input data was not filtered using a high pass filter
6. The Mel-Scaled Frequency Coefficients

FET	Approximate accuracy
Mel-Scaled spectrograms	85%
Standard spectrogram	80%
Linear standard spectrogram	75%
Chromagram with filter	70%
Chromagram	55%
Mel-Scaled Frequency Coefficients	45%

Table 1: The results from the deep models

Table 1 shows the results from each FET. Figures 15a and 14a show the chromagrams without filter and mel-scaled frequency spectrograms with a performance that is significantly worse than the other spectrograms with accuracy's around 50%. Figure 16a shows the results for the chromagrams where a high pass filter was applied before the conversion to the chromagrams. Here, the performance is much better which does follow a logical pattern as the lower frequencies no longer influence the chromagrams. The best performing FET's are shown in figures 12a, 13a and 17a. The two standard spectrogram version are quite similar in their performance which also makes sense since the only difference between these two FET's is the scale of the data. Out of all 6 FET's tested in this research, the mel-scaled spectrograms perform the best with an accuracy of around 85%. Given the images shown in section 3 this is not that suprising since the mel-scaled spetrograms were the clearest among all other images.

7 Responsible Research

The main concern that might come up when talking about recordings to identify a location is privacy. Since the recording are made on a regular basis if a user want real time information about their indoor location, any privacy sensitive information might be recorded without permission. For most of the FET's used in this paper, the only data that is saved are small audio samples with a frequency between 19.5k Hertz and 20.5k Hertz. This mostly eliminates the privacy concerns stated earlier. For one of the chromagrams however, the frequency range was not limited. This increases the privacy risk when deployed. However, since the accuracy of these chromagrams was among the lowest of all the FET's that were tested, the chance is less likely that the unfiltered chromagrams will be used for any future acoustic location sensing.
Another issue lies within the future of this research. Nowadays, a lot of applications that use location data, sell this data to third party companies [3]. When indoor location sensing becomes more mainstream, the same issue might occur as

companies can now also get access to more indoor location data which can be even a larger breach of privacy.

8 Discussion

Investigating what data set should be collected and how it should be processed is a broad subject that contains a huge amount of different options to research. In the rather short time span of this research, not all possible FET's could be given equal attention. This means that many other FET's have not been tested that might achieve a better accuracy than the FET's tested in this paper. The FET's that were used might also have an incorrect processing. Since these methods of feature extraction are quite different from each other, converting them to a 32x5 image might not work in the best interest of some FET's. Secondly, cutting of the lower frequencies in some of the resulting spectrograms might have causes valuable data loss..
A second limitation could have been the amount of locations that were used in the experiment. Using more locations will make results more reliable than the results shown in this paper. Another possible limitation is the hardware of the smartphone that was used in the experiments. Using chirps of 20k Hertz is something that the smartphone used in the experiments had difficulties with so the raw recorded data might already have been slightly unreliable.
In the end, the scale of this experiment was of a rather small size which might not have yielded the best results. For a better understanding of the performance that different FET's can deliver, an experiment of a larger scale using more rooms with more distinct features might be needed.

9 Conclusions and Future Work

9.1 Conclusion

In this paper, different methods were tested to see which of these methods performed better on recognizing different locations. 7 different locations where used to investigate which feature extraction technique performed with the highest accuracy. The experiments included the use of standard spectograms, Mel-Scaled spectrograms, Mel-Scaled Frequency Cepstral Coefficients and chromatograms. The results show that the standard spectrograms, Mel-Scaled spectrograms and filtered chromagrams deliver a higher accuracy than the mfcc's and chromagrams without filtering. The resulting accuracies ranged between 45% for the worst performing FET's and 85% for the FET's that had the best performance. Out of all the 6 techniques that were tested, the Mel-Scaled spectrograms turn out to be the best fit for indoor location sensing.

9.2 Future work

This paper shed some light on the way audio data should be captured and processed for effective indoor location sensing. While there are a lot of researches that have implemented quite successful classifiers, there is still more work needed before this technology can be used on a wider scale:

1. Data size: the size of the data set grows rather quickly. For this experiment, the total size of the data set grew to 4.55MB with 30 MB size taken on disk for only seven locations. This means that for a database containing multiple buildings with a large amount of rooms, a huge data set is needed to train a deep model.
2. Robustness. The training environment used in this paper contained almost no interference and the phone was not moved or rotated during the training. This means that most applications where a user does frequently move or rotate might not get the same accuracy as shown in this paper.
3. World wide integration. Should this technology one day be integrated into applications such as Google Maps to enhance the navigation to indoor locations, more research is needed on the scalability of this technology on a large scale.

A Examples

Standard spectrograms

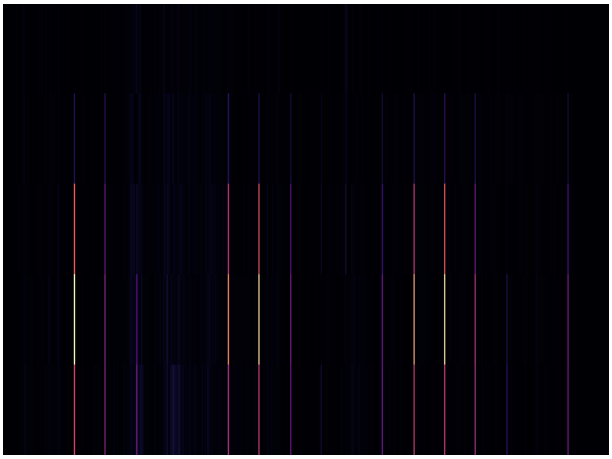


Figure 7: Example of a standard spectrogram.

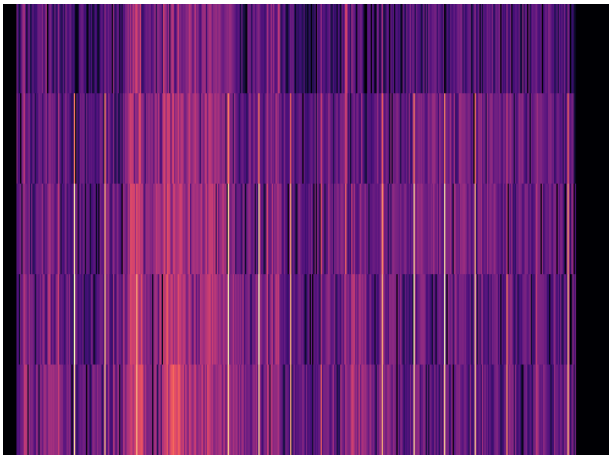


Figure 8: Example of a standard spectrogram where the values are converted to a linear scale.

Chromatograms

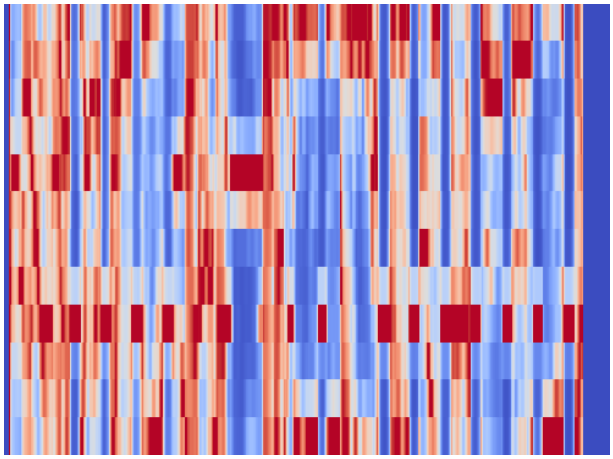


Figure 9: Example of a chromatogram.

Mfccs ceptrograms

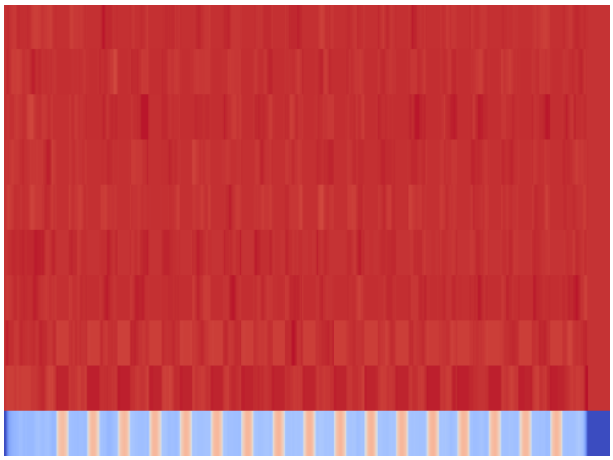


Figure 10: Example of a Mel-frequency cepstral spectrogram.

Mel-scaled spectrograms

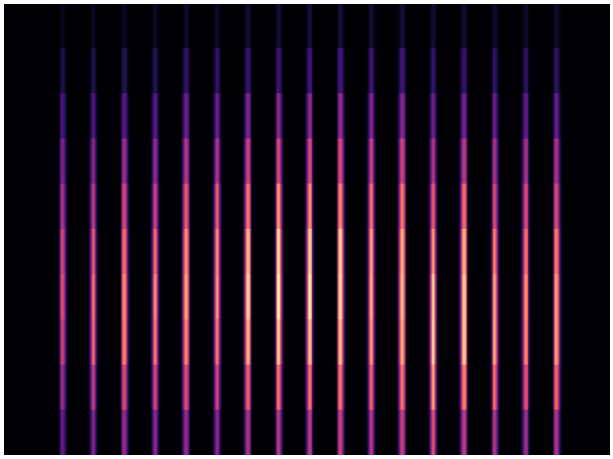
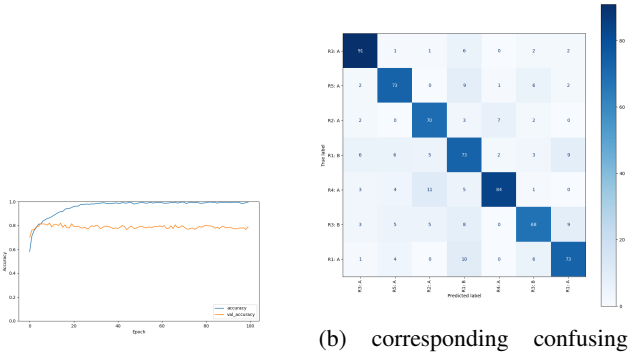


Figure 11: Example of a mel-scaled spectrogram.

B Accuracy results

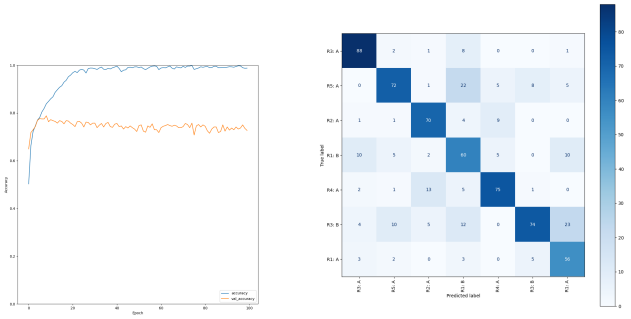
Standard method



(a) Performance per epoch (b) corresponding confusing matrix

Figure 12: Training results using the standard spectrograms

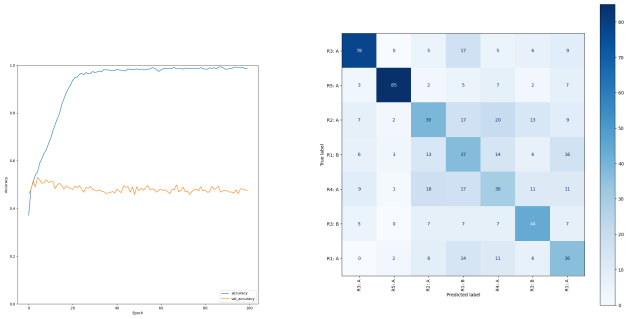
Standard method (converted to linear scale)



(a) Performance per epoch (b) corresponding confusing matrix

Figure 13: Training results using the linear scaled standard spectrogram

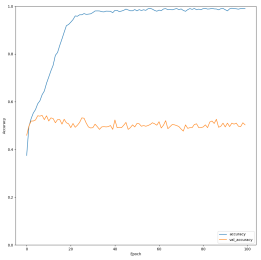
Mfccs



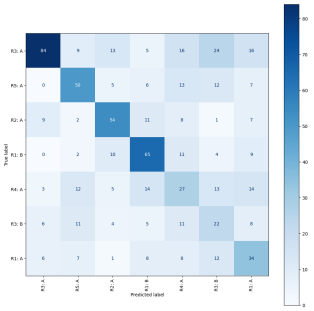
(a) Performance per epoch (b) corresponding confusing matrix

Figure 14: Training results using the mfccs spectrograms

Chromatogram



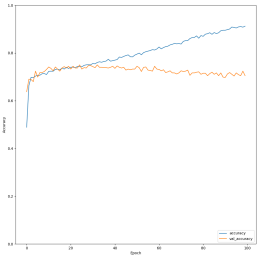
(a) Performance per epoch



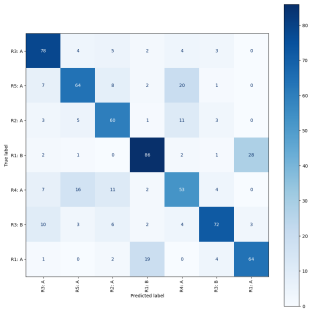
(b) corresponding confusing matrix

Figure 15: Training results using the chromatograms

Chromatogram with high pass filter



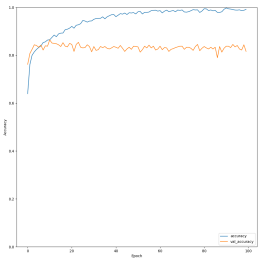
(a) Performance per epoch



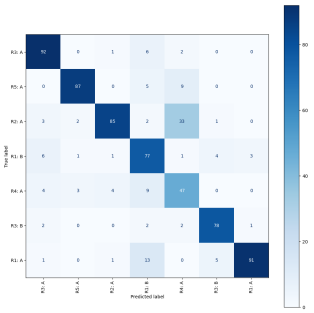
(b) corresponding confusing matrix

Figure 16: Training results using the chromatograms

Mel spectrograms



(a) Performance per epoch



(b) corresponding confusing matrix

Figure 17: Training results using the mel-scaled spectrograms

References

[1] J. Lian; J. Lou; L. Chen; and X. Yuan. Echospot: Spotting your locations via acoustic sensing. page 21, 2021.

[2] I. Diaconita; A. Reinhardt; F. Englert; D. Christin and R. Steinmetz. Do you hear what i hear? using acoustic probing to detect smartphone locations. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, pages 1–9, 2014.

[3] S. Brennan; S. Coulthart; and B. Nussbaum. The brave new w e new world of thir orld of third party location data ty location data. In *Journal of Strategic Security*, volume 16, pages 81–95, 2023.

[4] K. Liu; X. Liu; L. Xie; X. Li. Towards accurate acoustic localization on a smartphone. page 5. University of Florida, 2013.

[5] W. Luo; Q. Song; Z. Yan; R. Tan; Guosheng Lin. Indoor smartphone slam with learned echoic location features. page 15. Nanyang Technological University, The Chinese University of Hong Kong, 2022.

[6] S. P. Tarzia; P. A. Dinda; R. P. Dick; Gokhan Memik. Indoor localization without infrastructure using the acoustic background spectrum. page 14. Northwestern University, University of Michigan, 2011.

[7] P. Seetharaman and Z. Rafii. Cover song identification with 2d fourier transform sequences. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 616–620, 2017.

[8] H. Murakami; M. Nakamura; S. Yamasaki; H. Hashizume; M. Sugimoto. Smartphone localization using active–passive acoustic sensing. page 8. Stony Brook University, Beijing Jiaotong University, 2018.

[9] C. Ittichaichareon; S. Suksri; and T. Yingthawornsuk. Speech recognition using mfcc. pages 135–138, 2012.

[10] D. Guo; W. Luo; C. Gu; Y. Wu; Q. Song; Z. Yan; R. Tan. Demo abstract: Infrastructure-free smartphone indoor localization using room acoustic responses. page 2. Nanyang Technological University, The Chinese University of Hong Kong, 2021.

[11] Q. Song; C. Gu; R. Tan. Deep room recognition using inaudible echos. page 28. Nanyang Technological University, 2018.

[12] W. Huang; Y. Xiong X. Li; H. Lin; X. Mao; P. Yang and Y. Liu. Shake and walk: Acoustic direction finding and fine-grained indoor localization using smartphones. In *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, pages 370–378, 2014.

[13] B. Zhou; M. Elbadry; R. Gao; F. Ye. Batmapper: Acoustic sensing based indoor floor plan construction using smartphones. page 14. Stony Brook University, Beijing Jiaotong University, 2017.