Tampere University

Sami Sinisalmi

Acoustic scene classification using spatial spectrum estimation

Signal Processing
Bachelor of science Thesis
December 2019

Abstract

Sami Sinisalmi: Acoustic scene classification using spatial spectrum estimation
Bachelor of science Thesis
Tampere University
Degree Programme in Computing and Electrical Engineering, BSc
December 2019

Analysis of audio from our surroundings gives us important cues about the acoustic scene, with automatic analysis usually done by sound event detection or analysing the audio scene as a whole. On the other hand, inspecting the auditory space characteristics, or openness of the space, is a much less studied aspect. This thesis aims to study the classification of audio scenes based on the aforementioned auditory space characteristics with the use of different audio features. In this work, log-mel band energies and spatial spectrums for the audio recordings are calculated and used in the classification. The results revealed that best performance is obtained when using the combination of mel features and spatial spectrum, instead of either one of them. It was also observed how any differences inside a class can affect the results.

Keywords: audio scenes, auditory space characteristics, feature comparison, log-mel band energy, spatial spectrum.

Preface

This thesis was written as a part of Bachelor of Science program in the fall 2019. Audio processing has been an interesting subject for me and working on this project has only increased my interest on the subject. I would like to thank my instructors Annamaria Mesaros and Toni Heittola for all the help and advise they have given me throughout the fall.

Tampere, 4th December 2019

Sami Sinisalmi

Contents

List of symbols and abbreviations

ANN	Artificial neural network
DCASE	Detection and Classification of Acoustic Scenes and Events
DFT	Discrete Fourier Transform
DOA	Direction Of Arrival
LDA	Linear Discriminant analysis
ReLU	Rectified Linear Unit
STFT	Short-Term Fourier Transform
SVM	Support Vector Machine
TOA	Time Of Arrival

1 Introduction

Classifying acoustic scenes is an application of audio signal processing with many uses. Analysis of audio from our surrounding environments gives us information about the acoustic scene and for instance, we are able to recognize the scene we currently are in and individual sound sources in the scene. Ability to extract this information from audio could be used for example to recognize a car model by its engine sound, bird species from recordings or in surveillance and security systems[1]. So far the analysis has been mostly done by using classification or recognition of individual sounds in the scene or the scene by whole.

The data set of DCASE 2019 challenge task 1, acoustic scene classification provides us the largest amount of audio data from our everyday public environments, the data being recorded audio from public spaces in multiple different cities across Europe. This data set provides us material to study the acoustic scene and compare different methods in classification of the acoustic scenes.

This thesis aims to study the classification of audio scenes based on their auditory space characteristics or more simply put, by the openness of the space. This is done by classifying the examples using spectral and spatial features calculated from the audio. The study presents a comparison between the two feature representations, and the performance of a system that uses both.

Chapter 2 introduces the problem of the thesis and some background information about used techniques. Chapter 3 presents the used data set, classification setup and the results. Finally, Chapter 4 gives the final conclusions from the work.

2 Background

Classification of acoustic scenes is the classification of environments based on the general acoustic characteristics of the scene. This is usually done by training a classifier with recorded audio from various scenes, such as street, park, office, etc. Using different types of classifiers and by extracting various kinds of information from audio samples, impacts the results of classification.

2.1 Problem

This thesis focuses on classifying different audio samples to three different classes based on their auditory space characteristics, with a focus on investigating how different features from audio samples affect the results. The considered acoustic scenes are *outdoor*, *indoor* and *vehicle*, the expectation is that acoustic scenes which are indoors (*airport*, *shopping mall*, *metro station*) have some acoustic characteristics in common that differentiate them from scenes which are outdoor or inside vehicles. In order to characterise this aspect, the study uses spatial features and compares them with the commonly used mel energies. How does these different features and their combinations impact the results in a case of classifying audio by auditory space characteristics?

2.2 Features

There are many different types of information that can be extracted from a single audio signal, which can help us to identify the correct category for each sample. Different types of features extracted from audio often have an effect on the classification performance, which can be used as an indicator of the usefulness of the features in the given problem. From the results we should be able to understand better which kind of features are the most beneficial in terms of recognising acoustic scenes.

2.2.1 Log mel-band energies

Mel-frequency representation is a frequency representation of sound that uses a scale of mels instead of hertz. This Mel-scale represents perceived pitch and a commonly used formula of converting hertz into mel is:[2]

$$m = 2595\, log_{10}\,(1 + \frac{f}{700}) \tag{2.1}$$

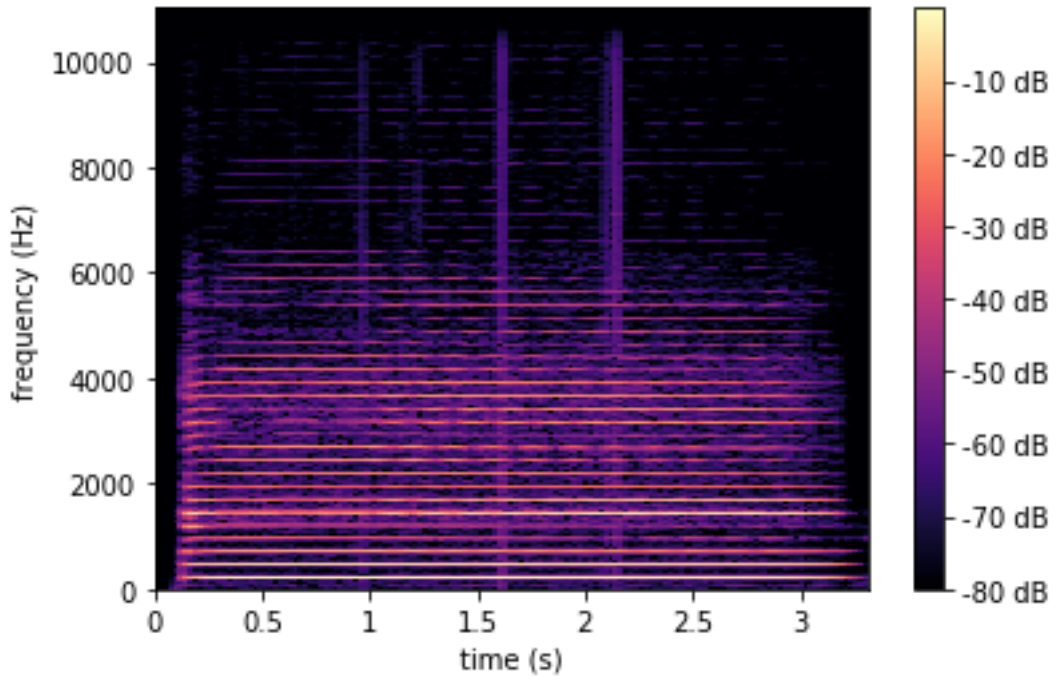

Figure 2.1 *Linear-frequency power spectrogram taken from an oboe playing*

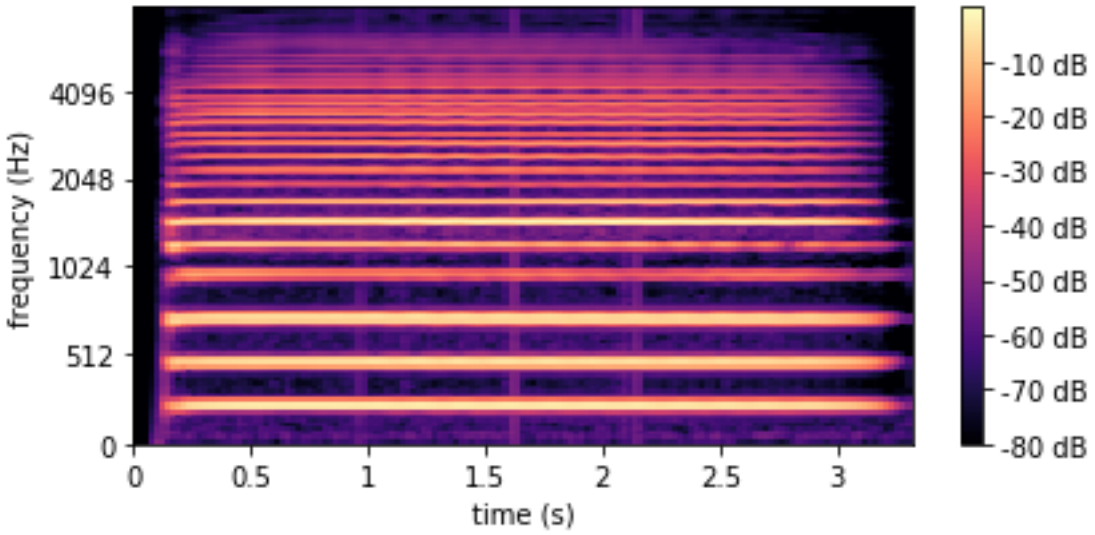

Figure 2.2 *Mel-frequency spectrogram taken from an oboe playing*

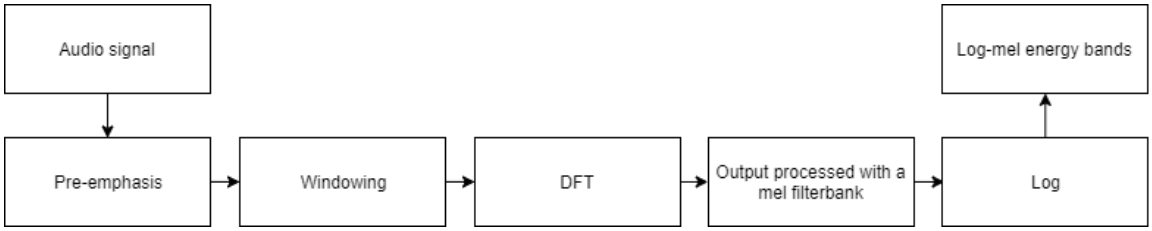

Figure 2.3 *Block diagram of extracting log-mel energies from an audio signal*

where frequencies m stands for mel and f stands for hertz. Figures 2.1 and 2.2 present a linear frequency and mel frequency spectrogram, respectively, for comparison of the two feature representations.

Extracting log-mel energies from an audio signal is illustrated in Figure 2.3. First an pre-emphasis filter is applied to the audio signal to amplify high frequencies. Then the audio signal is processed in short windows $x(n)$, with a certain window size and a hop-size that results in the windows overlapping each other. Windowing is applied using a window function $w(n)$, after which a DFT operation is done to these windows resulting in a spectrum of the window $X(f)$.

To obtain the mel energies, the spectrum of the signal is processed using a mel filterbank as one in Figure 2.4. The mel filterbank consists of triangular shaped band-pass filters W_k such that the center frequencies of these filters are equally spaced on the mel scale. The k stands for the filter number, while total amount of filters is usually $K = 40$. Mel-band energies are calculated by,

$$E(k) = \sum_{k=1}^{K} W_k(f)|X(f)|^2. \tag{2.2}$$

Taking logarithm of each $E(k)$ results on log mel-band energies of the audio signal.

These log mel-band energies provide information from audio signals that is correlated to human perception of sound. The bands mimic the approximately logarithmic perception of pitch of human hearing while log energy mimics the perception of loudness by humans. The motivation of these log mel-band energies is to model the human perception of audio.

2.2.2 Spatial spectrum

Direction of arrival or DOA has an important role in sound event detection[3] as well as in acoustic scene classification, as it gives us more information about direction of occurring sounds. Taking a DOA estimation of binaural audio results in a spatial spectrum, giving us information about direction of sounds occurring in the scene.

Extracting a spatial spectrum from binaural audio signal is illustrated in Figure 2.5. Binaural audio recording results in an audio signal for each of the two audio

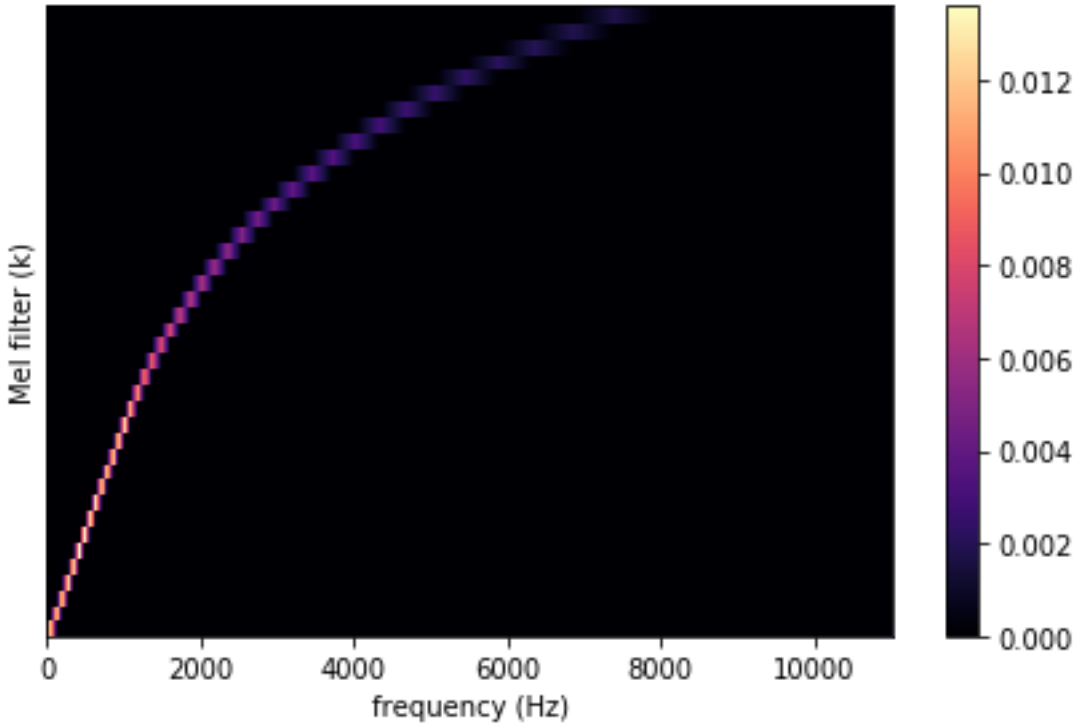

Figure 2.4 *A mel filter bank for $K = 40$ mel bands in range of 0-8kHz*

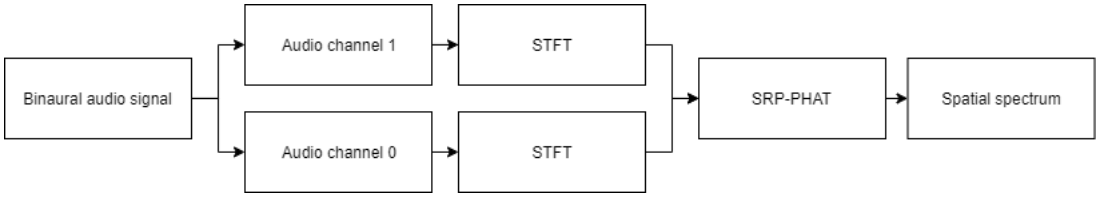

Figure 2.5 *Block diagram of extracting a spatial spectrum from binaural audio signal*

channels, STFT is then applied to both of these signals. A predetermined window size and a hop-size is used that results in the windows overlapping each other, just like in computing mel-band energies. DOA can be then estimated with steered response power with phase transform (SRP-PHAT)[4][5] for each of these windowed signals produced by the STFT.

SRP-PHAT is a very robust sound source localisation algorithm, but it has its drawbacks as it is computationally expensive[6]. In frequency domain the spatial spectrum's P locations q are calculated for a frequency band by[7],

$$P_q = \int_0^{2\pi} |\sum_{m=1}^{M} \frac{X_m(\omega)}{|X_m(\omega)|} e^{j\omega\tau[q,m]}|^2 d\omega, \tag{2.3}$$

where M is the number of microphones, $X_m(\omega)$ is the Fourier transform of the mth microphone signal for frequency ω. $\tau[q, m]$ is the time of arrival or TOA of a sound

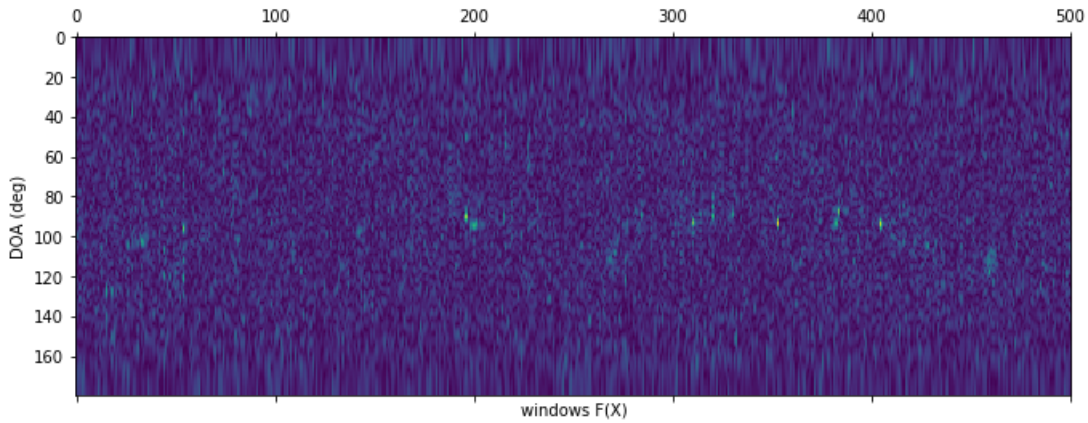

Figure 2.6 *A spatial spectrum of an 10 second audio recording from an airport in Barcelona*

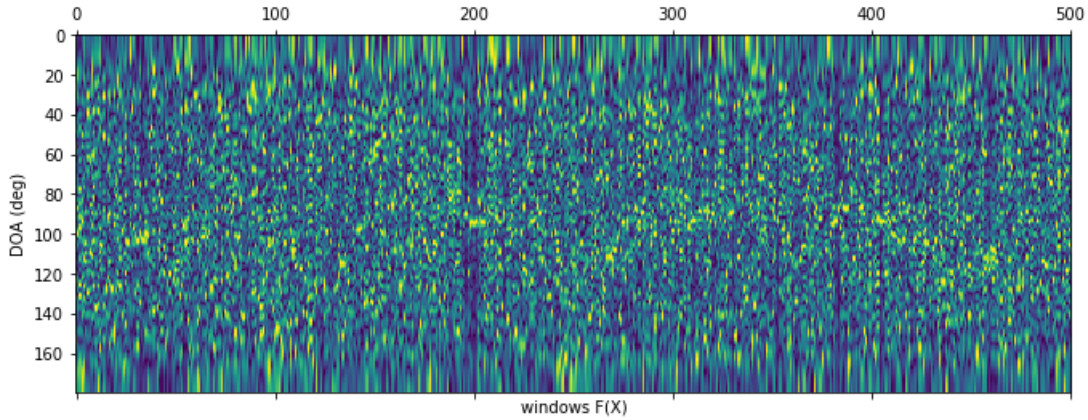

Figure 2.7 *Normalised version of the Figure 2.6 spatial spectrum*

signal from the location q to the mth microphone. DOA can then be extracted from

$$\hat{q} = arg\,max\,p_q. \tag{2.4}$$

From these results a spatial spectrum can be formed, such as the one illustrated in Figure 2.6 (calculated using window length of 40ms and a hop length of 20ms), which contains information about the sound energy per direction with an accuracy of one degree. Using binaural audio recording leads to the calculated spatial spectrum to contain observed sounds from all 360°. However, the 360° spectrum is obtained by mirroring the first 180° as it is not possible to disambiguate the sounds coming from front or behind of the recording system.

Spatial spectrum can be normalised by using min-max feature scaling

$$P_n(\theta) = \frac{P(\theta) - P(\theta)_{min}}{P(\theta)_{max} - P(\theta)_{min}}, \tag{2.5}$$

where $P(\theta)$ is the spatial spectrum, $P(\theta)_{max}$ and $P(\theta)_{min}$ being the spatial spectrum's maximum and minimum value respectively and $P_n(\theta)$ being the resulting normalised spatial spectrum. This can help when values in data have a really wide variation and it can affect negatively in the classification process.

This spatial spectrum and normalised spatial spectrum provides information from audio signals that is correlated to where occurring sounds happen in the scene, compared to the position of the listener.

2.3 Classifiers

The term classifier refers to an algorithm used in learning and classification process of a classification task. Used classifiers are similar in audio processing and in image processing. Classical classifiers like SVM and LDA are able to complete classification tasks with a shorter training time. These classifiers are linear classifiers, which have good performance if the classes are linearly separable. In a case, where the classes can not be separated linearly, the results of using such classifiers result in a low classification accuracy.

Complicated classification problems need classifiers that can learn complicated nonlinear relationships. Artificial neural network or ANN, is an network that consists of neurons which are connected to each others, connecting a layer of input neurons to a layer of output neurons. Information (like audio features) is fed to the network while giving the network an desired output (label of the audio example) from the information, while the network uses this operation to detect patterns in the information that leads into the desired output. This is done with different pairs of features and desired outputs until the network has learned to classify the information correctly.

A more advanced neural network called convolutional neural network or CNN was created in 1998[8], and has been gaining popularity in machine learning ever since. CNN is an artificial neural network, which has been popular in image processing as well as in other classification problems. An ANN propagates information from a layer to another, whereas a CNN works in the same way, but includes different kinds of layers, combining activation maps and pooling layers[9]. These layers then convolve with each other.

A single neuron output is defined by the input the neuron receives and by the neuron's activation function. If the neuron uses an step function as its activation function, it can only produce zeros or ones as its output, or more simply put, the neuron is activated. Layers consisting of neurons that use these activation functions or activation maps are used as filters to determinate which parts to activate from the previous layer. Pooling layers on the other hand are layers that divide the previous layer evenly to smaller areas. Then, values based on those areas are taken from

them and given to the next layer.

In CNN, the most commonly used activation function is ReLu, which gives an output of zero if its input is negative, otherwise its output is the same as its input. The most commonly used pooling option is max pooling, which takes the highest value of an area, where these areas are evenly divided in the previous layer. Usually the output layer of the classification uses a sigmoid function.

These layers then try to detect patterns, while having certain weights, giving the gained information to the next layer until the data reaches the last layer. In this output layer, each of the neurons' output is the probability of that neuron's represented class. The highest probability compared between the neurons is the predicted class.

State-of-the-art in audio classification includes CNN classifiers similar to the ones used in image processing, modified to take features from audio as their inputs. Most of the advancements have been made by improving the feature extraction process with audio data and feeding these neural networks better information for improvements in classification accuracy.

3 Methods

3.1 Data set

This section introduces the used data set and its different properties. The used data set in this work is the one provided in the DCASE 2019 challenge task 1, acoustic scene classification. The data set from DCASE 2019 challenge acoustic scene classification includes a total of 40 hours of audio from 12 cities in Europe where the audio has been collected in 10 different scenes: airport, metro station, shopping mall, park, pedestrian street, public square, street with traffic, bus, metro and tram [10].

Each audio file is recorded simultaneously with four audio recording devices [11], main recording device being Soundman OKM II Klassik/studio A3, electret binaural in-ear microphone and a Zoom F8 audio recorder with 48 kHz sampling rate and a 24 bit resolution. To simulate human auditory system, the microphones were worn in ears. All the data used in this work is recorded by the mentioned recording system.

In this work, the data set has been divided differently compared to the original purpose. To be able to study the auditory space characteristics more efficiently, the classes have been regrouped according to auditory space characteristics. Classes *airport*, *metro station* and *shopping mall* were combined into *indoor* class. *Park*, *pedestrian street*, *public square* and *street with traffic* into *outdoor* class and finally *metro*, *bus* and *tram* are combined into vehicles class. Now the number of classes is reduced to just three, where the classes have more variation inside them, but still having the related element of the auditory space characteristic which then enables us to study this aspect more precisely.

In total, the data set consists of 14,400 audio samples, each having a length of 10 seconds. So where each of the 10 classes had exactly 1,440 audio samples, the regrouped 3 classes have 4,320 audio samples each. Class *outdoor* is an exception though, as it had been made combining four classes instead of three, making its size 5,760 audio samples.

3.2 Setup

For this work, a baseline system from DCASE2019 challenge's acoustic scene classification task[12] was used (Figure 3.1). This baseline system provided a simple entry level state-of-the-art system that already gave some reasonable results for the DCASE2019 challenge's task 1, enabling easy testing of different features. The system was modified to perform the classification with three classes, for obtaining a baseline performance using mel energies.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_7 (Conv2D)            (None, 40, 500, 32)       1600
_________________________________________________________________
batch_normalization_7 (Batch (None, 40, 500, 32)       160
_________________________________________________________________
activation_7 (Activation)    (None, 40, 500, 32)       0
_________________________________________________________________
max_pooling2d_7 (MaxPooling2 (None, 8, 100, 32)        0
_________________________________________________________________
dropout_10 (Dropout)         (None, 8, 100, 32)        0
_________________________________________________________________
conv2d_8 (Conv2D)            (None, 8, 100, 64)        100416
_________________________________________________________________
batch_normalization_8 (Batch (None, 8, 100, 64)        32
_________________________________________________________________
activation_8 (Activation)    (None, 8, 100, 64)        0
_________________________________________________________________
max_pooling2d_8 (MaxPooling2 (None, 2, 1, 64)          0
_________________________________________________________________
dropout_11 (Dropout)         (None, 2, 1, 64)          0
_________________________________________________________________
flatten_4 (Flatten)          (None, 128)               0
_________________________________________________________________
dense_7 (Dense)              (None, 100)               12900
_________________________________________________________________
dropout_12 (Dropout)         (None, 100)               0
_________________________________________________________________
dense_8 (Dense)              (None, 10)                1010
=================================================================
Total params: 116,118
Trainable params: 116,022
Non-trainable params: 96
_________________________________________________________________

Input shape                      : (None, 40, 500, 1)
Output shape                     : (None, 10)
```

Figure 3.1 *Network summary of the used classifying system[12] on its default values*

The plan was to gain knowledge of how different features affect the classification accuracy. The features used in this study are log-mel energies, spatial spectrum and a normalised spatial spectrum. The training was done with all of these features separately as well as with combining mel features with spatial spectrum and mel features with the normalised spatial spectrum. The combination of features was done by concatenating the mel features with spatial spectrum as well as with the normalised spatial spectrum.

The baseline system included a log-mel band energy extraction of the audio files with 40 bands and could predict classes through training with these features. To get the spatial features, a spatial feature calculation was added into the system's feature extraction part.

Calculation of the spatial spectrum turned out to be a very computationally heavy process, even when using multiple cores. To make the calculation of spatial spectrum faster, a longer window length of 240ms and a hop length of 120ms was used instead of the originally planned 40ms/20ms. To match the window sizing with the mel energies, log-mel energy bands were also calculated using this window length and hop length, using 40 mel bands. Spatial spectrum calculation resulted in a spatial spectrum that represented directions of sound from 360°, but since it was just a mirrored 180°, only 180° part of the spectrum was used.

This modified the sizes of resulting feature matrices, which required modifying the network structure from the original baseline architecture. Sequencing processors sequence length and hop length was changed from 500 to 84 while the pooling layer's pool size was changed from 4x100 to 4x16.

Feature extraction was then done to each of the 14,400 audio samples; the training, testing and evaluation were done with the original training/testing/evaluation split that was provided in the DCASE2019 acoustic scene classification task.

3.3 Results

Training of the classifier was done separately with different features. Other aspects were left untouched, meaning that in each case the system would be identical except for the used features. This way, the impact for the classification result can be attributed to the different features used. The results are presented in Table 3.1.

It can be observed that the log-mel band energies concatenated with spatial spectrum (Mel and SP) and normalised spatial spectrum (Mel and NSP), gave the best results. Just using log-mel energy bands (Mel) proved to be good, unlike to just using the spatial spectrum (SP) or the normalised spatial spectrum (NSP) which performed more poorly.

As for the difference in accuracy between the tested classes, recordings from *Outdoor* class and *Vehicle* class were the easiest to classify. On the other hand,

Table 3.1 *Classification accuracy with different features*

Features	Indoor	Outdoor	Vehicle	Average
Mel	79.9	**90.7**	89.6	86.7
Mel and SP	**84.3**	87.0	**93.7**	**88.4**
Mel and NSP	81.7	90.3	93.0	**88.4**
SP	60.3	81.9	82.0	74.7
NSP	61.4	78.1	86.8	75.4

audio recordings from *Indoor* class were the hardest ones to predict correctly, this can be explained with how the *Indoor* class' acoustic scenes are rather different. For this reason, a separate investigation was performed in an additional experiment.

The data was split such that one of the acoustic scenes from each class was included to the test set, while the others were used to train the system. This resulted in three different test cases, where each of the case had one acoustic scene removed from each of the classes after which testing and evaluating was done to the removed class to study the generalisation of the system. These results are presented in Tables 3.2, 3.3 and 3.4

Table 3.2 *Classification accuracy for airport, park and bus acoustic scenes, while they were not in the training set*

Features	Airport	Park	Bus	Average
Mel	**69.4**	**90.2**	93.5	84.3
Mel and SP	56.1	89.6	**97.6**	81.1
Mel and NSP	**69.4**	89.6	96.9	**85.3**
SP	45.4	56.2	84.1	61.9
NSP	57.0	52.1	82.4	63.8

Table 3.3 *Classification accuracy for mall, public square and tram acoustic scenes, while they were not in the training set*

Features	Mall	Square	Tram	Average
Mel	78.9	76.7	89.9	81.9
Mel and SP	**87.5**	82.4	93.6	**87.8**
Mel and NSP	81.2	**85.0**	**95.9**	87.4
SP	66.0	83.7	82.3	77.3
NSP	51.2	80.4	90.1	73.9

It can be observed that even now the best performance was achieved using mel energies with a normalised spatial spectrum or a spatial spectrum as the features. It can be noted though, that usage of spatial spectrum compared to a normalised spatial spectrum is situational as there are situations where one outperforms the other.

Table 3.4 *Classification accuracy for metro station, street and metro acoustic scenes, while they were not in the training set*

Features	**Metro station**	**Street**	**Metro**	**Average**
Mel	48.7	66.1	**87.5**	67.4
Mel and SP	**51.3**	64.5	86.6	67.5
Mel and NSP	48.5	**74.8**	86.8	**70.1**
SP	47.4	53.9	86.6	62.6
NSP	50.3	50.1	82.4	61.0

There is even some variation where the spatial spectrums have better performance compared to just using mel energies.

But as what comes to classification accuracy in the classes, we can see that the *indoor* class had the worst prediction accuracy. Inspecting the 3.4 reveals us that the classification accuracy for *metro station* is significantly worse compared to others. This shows that the metro station is the most different acoustic scene in the *indoor* class, and cannot be reliably learned using the *shopping mall* and *airport* acoustic scenes, affecting the class' prediction accuracy as a whole.

4 Conclusion

In this thesis, we compared audio features in classification of audio scenes by their auditory space characteristics. The used features were log-mel energy bands and spatial spectrums, where comparison between use of these features and a combination of these features was made.

Good results were obtained by using just the log-mel energy bands, but an increase in the classification accuracy was gained by using these mel energies with a spatial spectrum. Using only the spatial spectrum resulted in a lower performance than using just the mel energies. From this, we can make an assumption that sound's DOA plays a role in acoustic scene classification and should be considered when dealing with acoustic scenes. There is still room for more improvements though, using shorter window lengths and hop lengths, we should be able to get higher accuracy. The use of GPU in calculation of SRP-PHAT[6], should significantly decrease the time spent when estimating the spatial spectrum. This would give us more time to do additional tests and try different window- and hop lengths.

It was also noticed that one of the acoustic scenes didn't fit that well with the class it was assigned to, resulting in a class that was harder to predict correctly. The acoustic scene didn't represent the auditory space characteristic of that class, compared to other scenes in the class. The classes consist of spaces of different dimensions, and very likely this acoustic scene was the one with a different room size. The class would benefit from having subclasses with similar room sizes resulting in similar auditory space characteristics.

In addition, when testing the generalisation with the classes, it was observed that the normalised version of the spatial spectrum was in some case a better choice than the non normalised and vice versa. This is something that could use more study, where especially does the normalised version perform better, when does it perform worse and why does either of these happen? Definitive conclusions cannot' be said from these tests, so more study could be done from this topic.

Bibliography

[1] Aki Härmä, Martin McKinney, and Janto Skowronek. *Automatic surveillance of the acoustic activity in our living environment*. July 2005.

[2] Douglas O'Shaughnessy. *Speech communication: human and machine*. 1987, p. 150.

[3] Sharath Advanne et al. *Sound event detection in multichannel audio using spatial and harmonic features*. Available: `http://dcase.community/documents/workshop2016/proceedings/Adavanne-DCASE2016workshop.pdf`. Sept. 2016.

[4] Joseph Hector DiBiase. *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. May 2000.

[5] *SRP-PHAT from pyroomacoustics python library*. Available: `https://pyroomacoustics.readthedocs.io/en/pypi-release/index.html`.

[6] Taewoo Lee, Sukmoon Chang, and Dongsuk Yook. *Parallel SRP-PHAT for GPUs*. Jan. 2016.

[7] Vicente Peruffo Minotto et al. *GPU-based approaches for real-time sound source localization using the SRP-PHAT algorithm*. Aug. 2012.

[8] Yann LeCun et al. *Gradient-based learning applied to document recognition*. Available: `http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf`. Nov. 1998.

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016, p. 326.

[10] *DCASE2019 Challenge subtask A*. Available: `http://dcase.community/challenge2019/task-acoustic-scene-classification#subtask-a`. DCASE Community. 2019.

[11] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. *A multi-device dataset for urban acoustic scene classification*. Available: `http://dcase.community/documents/workshop2018/proceedings/DCASE2018Workshop_Mesaros_8.pdf`. Nov. 2018.

[12] Toni Heittola. *DCASE2019 challenge task 1 baseline system*. Available: `https://github.com/toni-heittola/dcase2019_task1_baseline`.