Environmental Sound Monitoring Using Machine Listening and Spatial Audio

Marc C. Green

PhD

AudioLab

Electronic Engineering

University of York

UK

February 4, 2021

Abstract

This thesis investigates how the technologies of machine listening and spatial audio can be utilised and combined to develop new methods of environmental sound monitoring for the soundscape approach. The majority of prior work on the soundscape approach has necessitated time-consuming, costly, and non-repeatable subjective listening tests, and one of the aims of this work was to produce robust systems reducing this need.

The EigenScape database of Ambisonic acoustic scene recordings, containing eight classes encompassing a variety of urban and natural locations, is presented and used as a basis for this research. Using this data it was found that it is possible to classify acoustic scenes with a high level of accuracy based solely on features describing the spatial distribution of sounds within them. Further improvements were made when combining spatial and spectral features for a more complete characterisation of each scene class.

A system is also presented using spherical harmonic beamforming and unsupervised clustering to estimate the onsets, offsets, and direction-of-arrival of sounds in synthesised scenes with up to three overlapping sources. It is shown that performance is enhanced using higher-order Ambisonics, but whilst there is a large increase in performance between first and second-order, increases at subsequent orders are more modest.

Finally, a mobile application developed using the EigenScape data is presented, and is shown to produce plausible estimates for the relative prevalence of natural and mechanical sound in the various locations at which it was tested.

Contents

Declaration of Authorship

I declare that this thesis presents entirely original work and that I am its sole author. The work has not been previously presented for any award at the University of York or other university. The bibliography contains proper acknowledgement of all sources. Parts of this research have been previously presented as conference and journal publications, as follows:

- M. C. Green and D. T. Murphy, "Acoustic Scene Classification Using Spatial Features,", in *Detection and Classification of Acoustic Scenes and Events Workshop*, Munich, Germany, 2017
- M. C. Green and D. T. Murphy, "EigenScape: A Database of Spatial Acoustic Scene Recordings", *Applied Sciences* vol. 7, no. 11:1024, November 2017, doi: 10.3390/app7111204
- M. C. Green, D. T. Murphy, S. Adavanne, T. Virtanen, "Acoustic Scene Classification Using Higher-Order Ambisonic Features", in *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2019
- M. C. Green and D. T. Murphy, "Sound Source Localisation in Ambisonic Audio Using Peak Clustering" in *Detection and Classification of Acoustic Scenes and Events Workshop*, New York City, USA, 2019
- M. C. Green and D. T. Murphy, "Environmental Sound Monitoring Using Machine Learning on Mobile Devices", *Applied Acoustics*, vol. 159, pp. 107041, February 2020, issn: 0003-682X, doi: 10.1016/j.apacoust.2019.107041

Acknowledgements

This thesis represents the culmination of 4 years of doctoral study, but also of the 6 years since I first moved to York, with interludes in Tampere Finland, Redmond Washington, and finally back home in Manchester. A great number of people have been part of my life during these years without whom this thesis would not have been written.

I would firstly like to thank Dr. Jude Brereton, who re-opened the door to the scientific world that I had closed by some silly decisions as a teenager. Her support, especially in my early years at York, literally changed the course of my life.

Thanks to my supervisor, Prof. Damian Murphy, who was instrumental in shaping this project when all I had were a few vague ideas. Damian has remained a most enthusiastic advocate of my many research endeavours, even (especially) at those times when my own levels of enthusiasm became rather low.

Very many thanks to the wider AudioLab community for welcoming me into the fold. I will never forget the lunchtime rituals. Thanks especially to Frank, whose generous friendship and frank (ahem) advice has helped me get through tough times even when I was thousands of miles away.

Thanks to Prof. Tuomas Virtanen, Prof. Annamaria Mesaros, Toni Heittola and Sharath Adavanne for their genuine warmth in a very cold country!

To my family, thanks for putting up with me and keeping me going during the final months of writing up as I hastily returned to base during a global pandemic!

Finally, thanks to Woolly Bully and Mr. Bubbly, who can always be counted on to be silly when I get too serious.

Keynote

"There was always the distant bustle of the city, a deep and throbbing space-filling rumble of ironclad wagon wheels on cobbled streets and the grind of streetcars. It was almost like the sound of the ocean or the wind in the forest, yet deep with the brutality that only a city can offer in fact and spirit, no matter how glamorous the environment or euphoric the social veneer. This was a resonance we cannot experience today; rubber tires on smooth paved streets have muted the old, rough sounds of iron on stone and the clopping of thousands of horses' hooves, timing the slow progression of ponderous wagons and more sprightly buggies. It was a sound not to be forgotten: a pulse of life in vigorous physical contact with earth."

– Ansel Adams

1 | Introduction

It is one of the defining characteristics of humanity to alter the environments in which we live [1]. Over the past two-hundred years, industrialisation has impacted the geological makeup of the planet to the point that many geologists now term our current epoch the *Anthropocene* - the 'age of humans' [1, 2]. The impact of industrialisation on the sounds of our environments has been profound, with the most apparent manifestation being the increase of the background noise level in which people in developed countries live their lives. The internal combustion engine is, after all, much more cacophonous than windmills or water wheels. It is difficult to accurately track the changes in sound level that have occurred over time, as sound measurement equipment was not invented until the industrial revolution was well underway. One way is to measure the output of sound-making devices that have been preserved. An American police siren from 1912, for instance, measures at 88 dB(A) at a distance of 3.5 metres. A siren from 1974, on the other hand, measures at 114 dB(A) at this same distance [3], giving some indication of the increase in background sound level in cities between these two dates. There are well-documented health risks, including lack of sleep and increase in irritability, leading to more severe health effects, that have been related to background noise levels [4]. Thus the vast majority of laws regarding environmental sound focus on 'noise abatement'. There has been comparatively little consideration of how industrialisation has affected not only the level of background sound, but also its character. This, in fact, is key to the perception of sound.

A clear demonstration is made by Raimboult and Dubois' example of music in a

nightclub compared with aircraft sound from an airport [5]. The first is considered entertainment and will largely please listeners (though perhaps not neighbours), whereas the other is generally regarded as an annoying pollutant. This is despite the fact that both have very similar noise levels, and in fact the consequences of voluntary prolonged exposure to loud music can be much more severe, in terms of hearing loss, than more occasional exposure to loud aircraft. Whilst a fairly extreme case, this shows the key roles that content and context play in the perception of sound.

Recently, there has been some movement towards viewing environmental sound as 'a resource rather than a waste' [6], with the focus shifted towards the perception of environmental sound rather than absolute noise levels. This is has come to be known as the 'soundscape approach' [4]. There is, as yet, no standardised model for the subjective environmental sound assessment necessary for the soundscape approach, and adoption of its tenets in legislation has been limited. Such assessment methods as have been used so far have the key disadvantage of being typically more time-consuming, hence more expensive, than noise level measurement. It is perhaps no surprise, then, that these methods have not gained much traction beyond the academic research community, though there are projects underway aiming to change this [7].

Meanwhile, the discipline of machine learning for audio, or 'machine listening', including Computational Auditory Scene Analysis (CASA), has arisen as an area of interest amongst researchers looking to replicate mechanically the human aptitude for decoding acoustic scenes [8]. As implied in its name, CASA research has focused on modelling aspects of how the human hearing system processes sound, particularly for Automatic Speech Recognition (ASR), which aims to generate text from spoken language (as opposed to speech processing as a whole, which includes the reverse of this - synthesis of audible speech from text), and to some extent for Music Information Retrieval (MIR). There has been comparatively little investigation into systems that do not necessarily explicitly aim to model human hearing,

or into more generalised analysis of everyday acoustic scenes beyond speech and music. This newer approach has been termed Computational Analysis of Sound Scenes and Events (CASSE) [9]. Interest in CASSE has increased in recent years, with a notable milestone being the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge established in 2013 [10], and now running yearly.

1.1 Motivation and Thesis Outline

This thesis describes efforts to utilise machine listening technologies as a method for the study of the properties of acoustic environments in and of themselves. Spatial audio, that is, multi-microphone recordings capturing the spatial distributions of sound sources in an environment, has not been significantly explored in machine listening research until recently. A key area of investigation in this thesis is therefore the investigation of the utility of spatial audio and the importance, or otherwise, of the spatial properties of sound scenes to their effective classification. A sufficiently robust machine listening system could ultimately be used to generate parametric representations of acoustic environments, including information about the spectral and spatial properties of the kinds of sounds constituting each scene. This has the potential to vastly reduce the costs of sophisticated soundscape assessment and could go some way towards bringing this approach into the mainstream, where it could start having influence on the laws surrounding urban planning and noise management. The work in this thesis seeks to explore several avenues in this regard.

A large proportion of this thesis is dedicated to the development of approaches for environmental sound monitoring using spatial audio recordings. This ranges from the high-level classification of entire acoustic environments to the partial development and testing of a method to track individual sound sources. Practical considerations, as well as reflection on the required level of detail for relevant soundscape information, led to the additional development of a mobile application as a tool for on-site sound monitoring. All of these experiments and sub-projects were fa-

cilitated by the recording and organisation of EigenScape, a database of 64 acoustic scene recordings in the high resolution fourth-order Ambisonic spatial audio format. This research is presented in detail in the following chapters as follows:

Chapter 2 introduces such fundamentals of acoustics and spatial audio as are required for understanding the work presented later in the thesis. An intuitive picture of a sound field is built up from first principles, from the nature of a sound wave, to the properties of sound propagation in space, and the representation of a sound field as a superposition of several travelling waves. There is discussion of the acoustic phenomena most relevant to the types of acoustic environments considered in this work, and those features of the human hearing system relevant to the human-centric features extracted from recordings later in the work. The chapter concludes with a thorough review of the theory of spherical harmonics underpinning the Ambisonic format, central to much of the subsequent work in the thesis.

Chapter 3 provides background regarding environmental sound monitoring related to the main motivation of this work. A thorough review of the soundscape approach is presented and contrasted with standard approaches taken to tackling environmental noise. There is discussion on the terminology used in the field, as distinctions between the underlying physics of acoustic environments and their perception as soundscapes is a subtle but important factor in understanding the divergence between the environmental noise and soundscape approaches. There is also an overview of various systems of categorisation that have been proposed for environmental sounds, along with how these relate to these two approaches and may inform attitudes therein. The chapter concludes with a summary of the prevalent methods for subjective soundscape assessment including on-location soundwalks and laboratory-based testing using various means of acoustic environment reproduction.

A detailed background of the machine learning frameworks, algorithms, and techniques that are utilised in this work is presented in Chapter 4. This includes a high-level overview of the specifics of acoustic scene classification and sound event detection, along with detail on the long-standardised mel-frequency cepstral coeffi-

cient (MFCC) features and Gaussian mixture models (GMM). These were used as a baseline in the original DCASE challenge and are used in a similar capacity in later chapters. The other algorithms covered are support vector machines (SVM), density-based spatial clustering of applications with noise (DBSCAN), and a short summary of the ideas behind convolutional neural networks (CNNs). Some examples of early acoustic scene classification work are presented, illustrating potential difficulties to be avoided, followed finally by more recent systems which utilise spatial audio in their operation.

The EigenScape database is introduced in Chapter 5, beginning by outlining several previously-available datasets. This highlights reasons why none of these were appropriate for the work conducted in this thesis, but also emphasises the organisational elements and content of these datasets that inspired the specification for EigenScape. The equipment used to record the database, in particular the mh Acoustics Eigenmike, is then detailed, along with technical description of the specific recording settings and Ambisonic file format. The chapter continues with a description of the eight location classes contained in the database and maps are included showing specific recording locations. A description of the cross-validation technique used to increase the utility of small databases, and how this can be applied to EigenScape, concludes the chapter.

Chapter 6 covers the first investigation using the EigenScape data, which investigated the usefulness of spatial audio features for acoustic scene classification. The work had the motivation of determining whether acoustic environments could be classified based solely on their spatial features, and, through this, validating the EigenScape data. The processes by which spatial features are extracted from both first-order and higher-order Ambisonics are explained, and results from classifiers trained using both sets of features are compared to those from classifiers trained using standard MFCCs. The chapter concludes with a summary of work conducted in collaboration with colleagues at Tampere University, Finland, using CNNs to classify environments using these same features, which led to additional insights.

The logical next step for the work was the development of methods to identify individual sound sources based on spatial features. This is the subject of Chapter 7, which begins by describing how features extracted using similar techniques to the previous chapter were analysed to estimate sound source directions and trajectories of movement. The performance of this system was, for practical reasons, assessed using synthesised data, rather than EigenScape.

The final piece of original work in this thesis, presented in Chapter 8, is the development of a mobile application featuring a machine learning model trained using EigenScape data. This model produces metrics describing sound environments in terms inspired by the soundscape approach. The app also makes use of the Sennheiser AMBEO smart headset and Apple ARKit, enabling the user to place virtual sound objects to test the effects of potential modifications to actual acoustic scenes. This represents an attempt to implement some of the findings and techniques explored previously in an application for practical use.

1.2 Statement of Hypothesis

In this thesis, a portfolio of work is presented, integrating a variety of aspects of machine listening, spatial audio, and the soundscape approach. These works are all enabled by the EigenScape data, which was originally recorded to provide a basis on which to investigate the following hypothesis:

> *The monitoring of acoustic environments, with a view to deriving information useful to the soundscape approach, can be assisted using spatial audio analysis.*

Details of the work that has been undertaken in this regard are covered in the rest of the thesis as outlined in this introductory chapter. The results of these investigations will support or refute this hypothesis, whilst likely providing new insights and uncovering new questions to investigate. In conducting this research, novel contributions to the field have been noted as follows:

- The EigenScape database of fourth-order Ambisonic acoustic environment recordings.
- Proof that acoustic environments can be characterised by their spatial properties alone.
- The combination of spherical harmonics beamforming and DBSCAN clustering for source tracking.
- Investigation of trade-off between Ambisonic order and source tracking performance.
- The estimation of soundecology metrics using scene classification techniques.

These contributions will be explored in depth in the subsequent chapters as outlined above. Chapter 9 will conclude the thesis with a summary and short reflection on the findings of the work, and will return to the hypothesis in light of these. Finally, recommendations will be proposed for various ways in which this work might be continued in the future.

2 | Fundamentals of Sound in Space

2.1 Introduction

In order to constructively discuss how a sound field might be measured and characterised by a human listener or a machine learning system, it is first necessary to set out the nature of sound waves and how they combine in a space to produce a sound field that can be perceived and recorded spatially. In this chapter, we will explore the basic concepts of sound, from the characteristics of sound waves and room acoustics, to the measurement of sound using microphones, and how spherical harmonic signals obtained from microphone arrays are able to record the variation of sound in space as well as time. The chapter will conclude with information on how the human hearing system perceives spatial sound.

2.2 Basic Properties of Sound

2.2.1 Sound Waves

In essence, sound is the transmission of vibration from a source object through a medium, such as air, as a series of compressions and rarefactions of said medium. A useful representation often used is the 'ball-and-spring' model, in which the balls are analogous to the masses of the molecules in the transmission medium and the springs are the inter-molecular forces between them [11]. Sound waves are, therefore, longitudinal; the direction of propagation is parallel to the direction of disturbance.

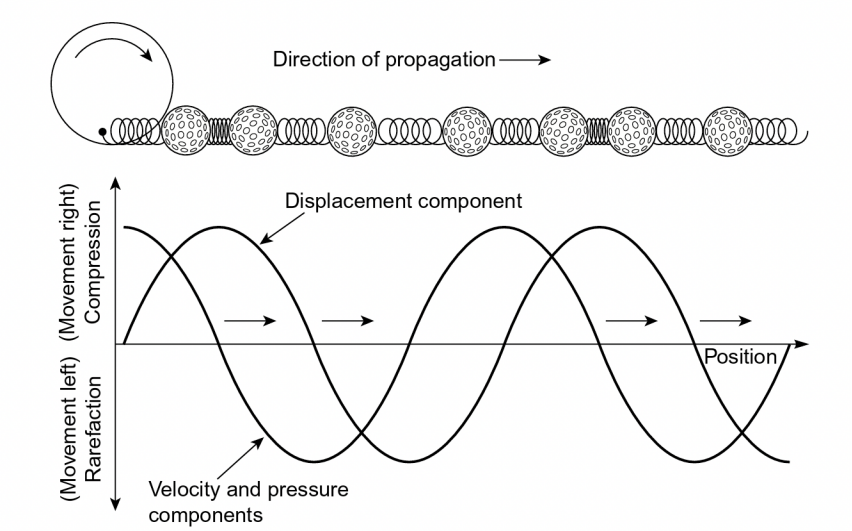


Figure 2.1: An illustration of the ball-and-spring model of sound transmission, together with its transverse visualisation (from [11]).

This is in contrast to transverse waves, in which the disturbance is orthogonal to the direction of propagation. Figure 2.1 shows a ball-and-spring visualisation of a sine wave, together with transverse visualisations of the displacement and velocity/pressure components of the wave.

The pressure component is the force needed to accelerate the air molecules, and is the scalar quantity that is measured when sound is recorded at a point in space. This will be explored in more detail in Section 2.3.1. Velocity is in phase with pressure, but is a vector which defines the direction of travel for the sound wave. The ratio of the pressure component amplitude p and velocity component amplitude U is the acoustic impedance Z_0:

$$Z_0 = \frac{p}{U} \tag{2.1}$$

This value is analogous to resistance in an electrical circuit and is a constant for any given medium. Z_0 is determined by the mean density ρ of the medium together with the velocity of sound in the medium c:

$$Z_0 = \rho c \tag{2.2}$$

The velocity of sound in a gas c_{gas} can be calculated by [11]:

$$c_{\text{gas}} = \sqrt{\frac{\gamma RT}{M}} \tag{2.3}$$

where γ is the heat capacity ratio of the gas, M is the molecular mass, R is the gas constant (8.31 J K^{-1} mol^{-1}) and T is the temperature in Kelvin. Estimating M for air at 2.89×10^{-2} kg mol^{-1} [12], the speed of sound at the nominal temperature of 20°C (293.15K) is 343.5 ms^{-1}. Estimating ρ_0 of air at 1.205 kg m^{-3} [12], Z_0 for air is 413 kg m^{-2} s^{-1}. Since R is a constant and γ and M will not change excepting very small local differences in the ratios of the constituent gases making up air, temperature is the most important variable influencing the speed of sound.

Apart from very short pulses, the majority of sound waves are periodic in nature, that is, they consist of regularly repeating patterns. The simplest periodic vibration is the sine wave, of which the wave depicted in Figure 2.1 is an example. Sine waves are defined based on three parameters:

- Amplitude A, which in terms of a sound wave can be thought of as the difference in pressure between the peak of a compression or rarefaction and the middle-point of the cycle. The amplitude of a wave is experienced as its loudness.

- Frequency f, which is the number of complete cycles the wave completes in one second. This is the most salient property of the wave relating to the perception of pitch.

- Phase ϕ, which is the 'starting position' of the wave. This has very little perceptual bearing in isolation, but does affect how waves interfere.

Sine waves can therefore be defined as:

$$y(t) = A\sin(2\pi ft + \phi) \tag{2.4}$$

The frequency of a sine wave is related to its wavelength λ, which is the distance between two points in the wave that are at the same phase. This is perhaps easiest to visualise as the distance between two peaks or troughs. Wavelength is related to frequency by [11]:

$$c = f\lambda \tag{2.5}$$

2.2.2 The Frequency Domain

Most sounds are, of course, much more complex than sine waves. It has long been known, however, that any given periodic waveform can be modelled as a combination of sines of various amplitudes, frequencies and phases. This is Fourier's theorem, mathematically [11]:

$$f(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos(n\omega_0 t) + b_n \sin(n\omega_0 t) \tag{2.6}$$

where ω is the angular frequency ($2\pi f$), a_0 is the offset of the entire signal from 0, known as the bias or d.c. component, and a_n and b_n are the contributions of the nth sine and cosine components, respectively. It's important to note that according to the trigonometric addition formulae, the linear sum of a sine and a cosine is equivalent to a single weighted and phase-shifted sine [13]. Euler's formula [14] also states that any combination of sine and cosine can be expressed as a complex exponential:

$$e^{j\theta} = \cos\theta + j\sin\theta \tag{2.7}$$

where $j = \sqrt{-1}$, the imaginary unit. This allows the re-expression of the Fourier series as:

$$f(t) = \sum_{n=-\infty}^{\infty} C_n e^{jn\omega_0 t} \tag{2.8}$$

where C_n are complex coefficients describing the contributions of each sinusoid. The absolute value (modulus) of C_n represents the magnitude of contributing sinusoid n, whilst the angle (argument) represents the phase. The set of sinusoids that make up a periodic signal are known as a Fourier series. The frequency ω_0 is the fundamental frequency of the signal, and it can be seen that the frequencies of the other sinusoids are integer multiples of this fundamental. As is stated in Equation 2.8, an infinite number of sinusoids may be needed to properly construct the signal. Figure 2.2 shows the first four sine waves (or *partials*) required to synthesise a square wave, together with the signal that results from adding only these first four partials. The signal that results from adding 50 partials is also shown, followed by the idealised wave that would result from the addition of an infinite number of partials. Square waves are characterised by the inclusion only of odd partials at the relative amplitudes of $1/n$. Note that the synthesised waves exceed the bounds of the ideal square wave. This phenomenon is known as the Gibbs overshoot, and is apparent when synthesising any waveform with discontinuities, including the square wave. As the number of partials used to synthesise the waveform increases, this overshoot decreases, as is apparent when comparing the $N = 4$ and $N = 50$ waveforms. Theoretically, when an infinite number of partials are used, this overshoot becomes infinitely small and disappears.

The frequency content of any given signal $F(\omega)$ can be obtained using the Fourier transform. Although in reality not all signals are periodic, aperiodic signals can be treated as periodic signals with an infinite period [11]:

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt \tag{2.9}$$

The above equation is the continuous form of the Fourier transform, however since digital audio is comprised of samples of a continuous sound signal, more often used in practise is the discrete form:

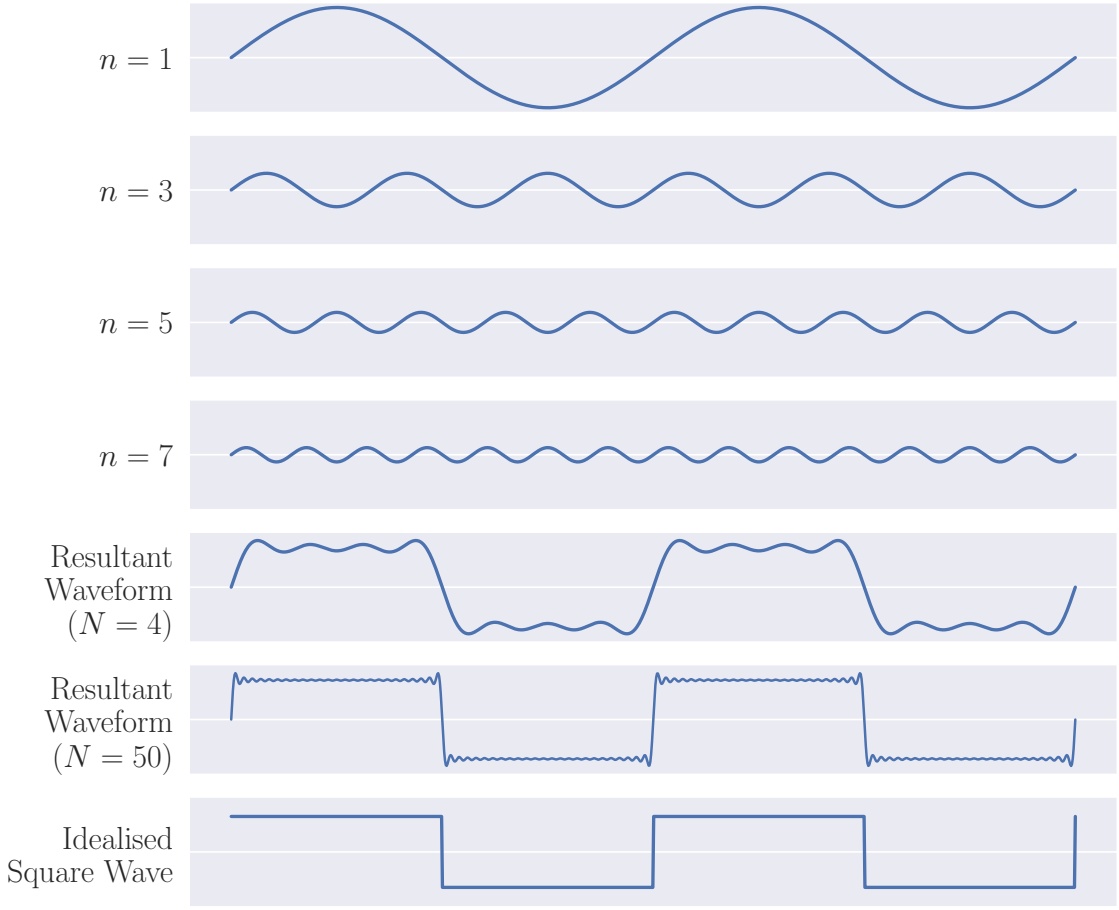


Figure 2.2: Fourier synthesis of a square wave, showing the first four partials, the waveform resultant from combining these, the waveform obtained from 50 partials, and the idealised wave.

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N} \tag{2.10}$$

The continuous signal $f(t)$ is replaced by a sampled signal $x[n]$, with the infinite integral replaced by a finite summation over the available samples. The discrete Fourier transform (DFT) returns N equally-spaced frequency samples $X[k]$ with bandwidths and centre frequencies determined by the sampling frequency of the signal. $X[k]$ is known as the frequency domain representation of the signal. The original time domain representation of a signal is completely recoverable from the

frequency domain representation using the inverse Fourier transform:

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] e^{j2\pi nk/N} \tag{2.11}$$

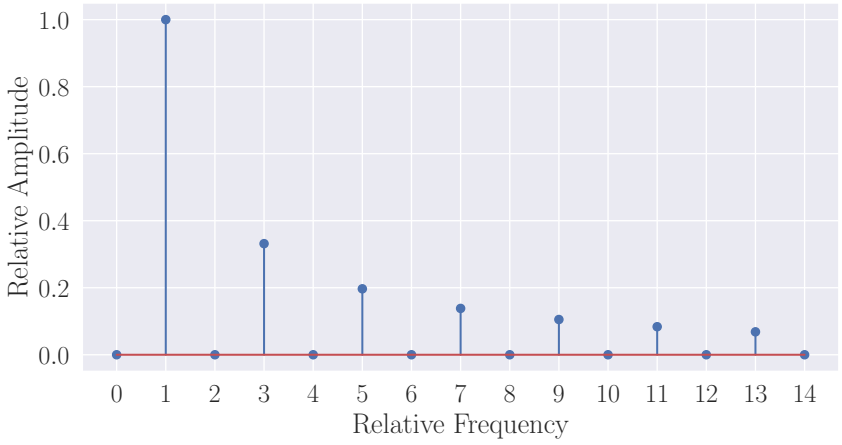
Figure 2.3 shows the frequency domain representation of a square wave obtained using the DFT, truncated at the 14th frequency bin. The DFT is often used on short frames of audio, so changes in frequency content over time can be monitored, producing a time-frequency representation. Whilst the frequency spectrum of harmonic signals such as the square wave will always consist of evenly-spaced harmonics, applying the DFT to noisy signals will give a frequency spectrum showing wider bands of contributing frequencies. The DFT of white noise theoretically shows equal power at all frequencies. Naturally-occurring sounds generally consist of combinations of harmonic and noisy components.

2.3 Sound Propagation in Space

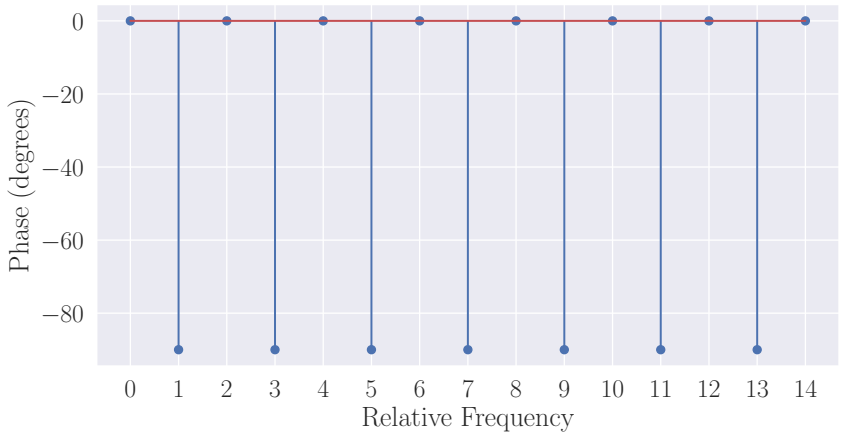
2.3.1 Sound Level

Now we have established the basic makeup of sound signals, we can begin to consider how these signals propagate through space and combine to produce a sound field. The first consideration is how the amplitude of a sound is quantified. There are many different measurable quantities that can be used, but since the human ear is sensitive to pressure, the predominant measure is sound pressure level (SPL), defined as the root mean square (RMS) pressure at a given point [11]. Real-world SPLs range over several orders of magnitude, from less than 2×10^{-5} Pa (the threshold of hearing) to more than 20 Pa (the threshold of pain). Given this enormous range, SPL is typically reported on the logarithmic decibel (dB) scale, defined as:

$$\text{dB SPL} = 20 \log_{10} \left(\frac{\text{RMS Sound Pressure}}{2 \times 10^{-5} \text{Pa}} \right) \tag{2.12}$$



(a) Amplitude component.



(b) Phase component.

Figure 2.3: The frequency domain representation of a square wave.

Sound Sources at Distance	**dB SPL**	**RMS Pressure**
Jet Aircraft at 50 m	140 dB	200 Pa
Threshold of pain	120 dB	20 Pa
Disco at 1 m from loudspeaker	100 dB	2 Pa
Busy road at 5 m	80 dB	0.2 Pa
Conversational speech at 1 m	60 dB	2×10^{-2} Pa
Quiet library	40 dB	2×10^{-3} Pa
Background of recording studio	20 dB	2×10^{-4} Pa
Threshold of hearing	0 dB	2×10^{-5} Pa

Table 2.1: Some examples of typical environmental sound levels (after [15]).

Table 2.1 shows typical SPL levels for some example sound sources. The distance from the sound source is an important factor on the sound level at the measurement position. Whilst Figure 2.1 shows a model for sound propagating in one dimension, in reality sound propagates through air in three dimensions. In order to consider the effect of this on sound level, we must consider an alternative measurement which takes into account the area affected by the sound waves, namely sound intensity level (SIL), which is measured in watts per square meter (W m^{-2}). Since free-field radiation from a point source in 3D space will necessarily be spherical, the sound power at the source P will be spread over a spherical surface of ever-increasing radius r, thus the sound intensity at a distance r can be calculated as [11]:

$$I = \frac{P}{4\pi r^2} \tag{2.13}$$

A measurement of SPL at a point at distance r will show a reduction in level corresponding to that of the SIL, as a rule of thumb a 6 dB reduction with each doubling of r. This formulation comes with a few caveats, as it assumes an infinitesimally small source radiating unimpeded omnidirectionally. Real sources are always of finite size and rarely perfectly omnidirectional, and real environments are never actually 'free

field' but have several features that affect propagation, as will be discussed. Nevertheless this equation generally provides a good estimate and is a useful intuition for basic sound propagation.

2.3.2 Room Acoustics

In this section, factors affecting the real-world propagation of sound, namely interaction with surfaces or boundaries, will be introduced. It should be noted that a large part of the work in this thesis is concerned with outdoor sound, in which such boundaries are much more sparse than in a room. Apart from the ever-present ground reflection, large open spaces are about the closest to free-field propagation one can find [16]. Although several of the concepts from room acoustics do not strictly apply outdoors, in built-up urban environments there can be much interaction with the surrounding architectural infrastructure, leading to unique sonic imprints at given locations. It is therefore beneficial to briefly review the basics of room acoustics.

In general, a recording at a point in a room of a sound signal emitted in another part of the room will consist of three parts [11]:

- **Direct sound** - The sound as it has travelled along the shortest possible path between source and receiver. Thus far, this can be considered a free-field propagation apart from a small amount of absorption by the air, and as such the level of the sound will be attenuated according to the inverse square law as defined in Equation 2.13.

- **Early reflections** - The first few reflections to arrive at the receiver, having been reflected from one or more surfaces in the room. These reflections are specular, that is they appear to come from a definite point, and are generally lower in level than the direct sound, having travelled a greater distance through air and been subject to some absorption by the reflecting surfaces. Early reflections give a listener a sense of the size and character of the space. If early reflections arrive within about 30 milliseconds, they will be perceptually

fused with the direct sound, altering its timbral 'colour' [17]. Often in open outdoor spaces, reflecting surfaces other than the ground are much farther away than in a typical room, so many early reflections will take a relatively long time to arrive and be perceived as distinct 'echoes'.

- **Reverberation** - Also known as late reflections, this is the field that builds up as the sound continues reflecting from multiple surfaces. Whilst the reflections are still physically distinct, over time the set of reflections becomes dense both temporally and spatially. This means individual reflections cannot be perceived and the reverberant field is diffuse, appearing to envelop the listener. Over time, the reverberation is attenuated to the point where it is no longer audible. Open outdoor spaces do not build up late reverberant fields.

The reverberation characteristic of a room can be recorded as its room impulse response (RIR). The RIR contains complete information about the reflections in a space for the source and receiver positions from which it was recorded. The excitation *impulse* is the Dirac δ-function:

$$\delta(t) = \begin{cases} \infty & t = 0 \\ 0 & t \neq 0 \end{cases} \tag{2.14}$$

The Fourier transform of the δ-function is flat across the frequency spectrum. This ensures that the RIR records the response of the room to all frequencies. It is physically impossible to produce a perfect impulse, so often rooms are excited using a swept sine wave signal which covers all frequencies of interest [18, 19]. The RIR is recovered from this by deconvolution of the recorded sweep, effectively collapsing the sine sweep to a single instant in time. Figure 2.4 shows a simplified RIR, and how the three components of a typical room acoustic are represented therein.

There are several parameters that can be used to describe the characteristic of a room acoustic. Predominant among these is the reverberation time RT_{60}, defined as the amount of time taken for a 60 dB decrease in sound energy upon the cessation

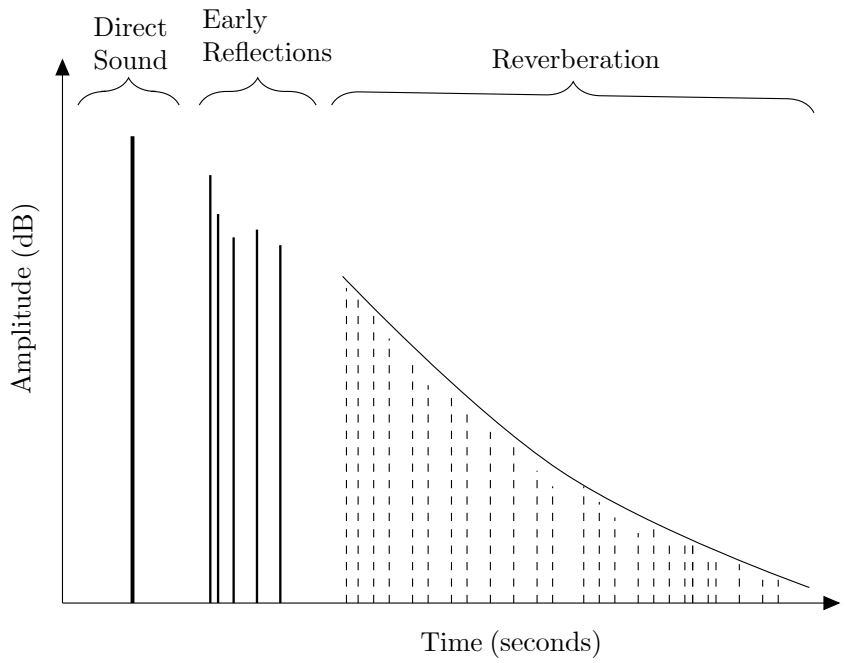


Figure 2.4: A simplified representation of a typical RIR structure (from [20]).

of the source sound. The technique to estimate RT_{60} from the RIR is to calculate the energy decay curve (EDC) by taking the reverse integral of the squared RIR [19, 21]:

$$\mathrm{EDC}(t) \triangleq \int_t^\infty h^2(\tau) d\tau \tag{2.15}$$

where τ is the time constant and h is the impulse response. Generally the noise floor of the RIR is high enough that the EDC does not linearly reach -60 dB, so typically the RT_{60} is estimated from two points on the linear decay portion of the EDC. Often -5 dB and -35 dB are used as these points, yielding the T_{30} estimate for RT_{60}. Figure 2.5 shows this method graphically. The plot depicts a log squared RIR, with the EDC shown in orange. The two red crosses show the points on the decay curve at -5 and -35 dB, with dotted lines showing the linear extrapolation to the energy at -60 dB. In this case, the RT_{60} is estimated to be around 1.65 seconds.

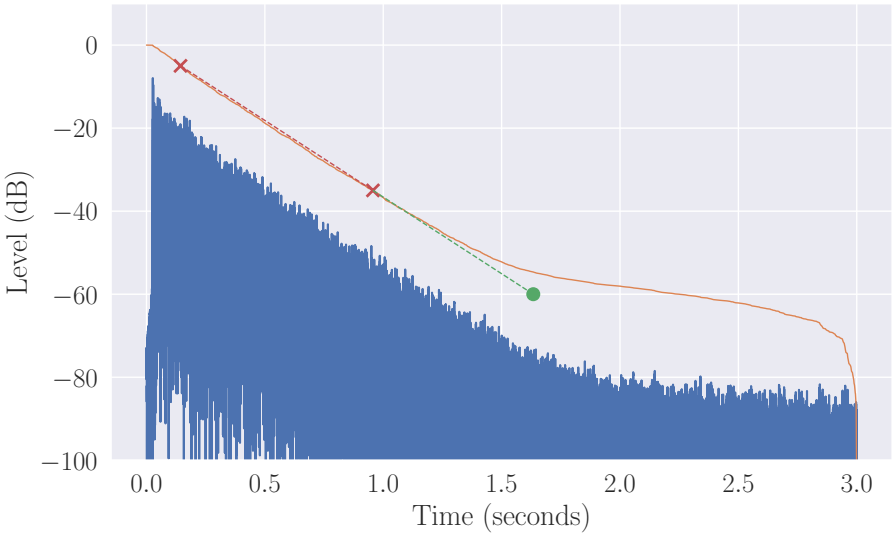


Figure 2.5: Calculating the T_{30} estimate of the RT_{60} from a recorded RIR.

2.3.3 Diffraction

Another important property of sound propagation in real spaces is diffraction. This is the bending of sound waves as they encounter openings (apertures) or obstacles where the sound is free to propagate on one or both sides. Diffraction is an especially important consideration for built outdoor environments and sound barriers [22]. The angle of diffraction bending θ_{spread} depends on the relationship between the size of the aperture or obstacle (L) and the wavelength of the sound [11]. If the wavelength is much smaller than the object ($\lambda \ll L$), only a small amount of diffraction occurs, whereas if the wavelength is comparable to the object size, or larger ($\lambda \geq L$), the amount of diffraction is large. This can be estimated by [23]:

$$\theta_{\text{spread}} \approx \frac{\lambda}{L} \tag{2.16}$$

As such, the farther off-axis a listening position to an opening or edge of a barrier, the greater the attenuation of high-frequency sound.

2.4 Recording Sound Fields

A basic model of a sound field is that of a set of travelling wavefronts in three dimensions. Although waves emitted from point sources are spherically curved, it simplifies the mathematics involved considerably to consider the wavefronts as planar. This assumption does not incur too much error if the measurement of the sound field is made at a reasonable distance from any sources of sound. A point in spherical co-ordinates $\mathbf{r}$ includes radius r, elevation $\theta \in [0, \pi]$ and azimuth $\phi \in [0, 2\pi]$, related to cartesian co-ordinates as [24]:

$$\begin{aligned} x &= r \sin\theta \cos\phi \\ y &= r \sin\theta \sin\phi \\ z &= r \cos\theta \end{aligned} \tag{2.17}$$

Sound pressure p from a single travelling plane wave as measured at a point $\mathbf{r}$ can be defined as [24]:

$$p(\mathbf{k}, \mathbf{r}) = e^{-j\mathbf{k}\cdot\mathbf{r}} \tag{2.18}$$

$\mathbf{k} = [\, k \;\; \theta_k \;\; \phi_k \,]$ is the wave vector, denoting the wave's direction of travel where k is the angular wavenumber (spatial frequency), in radians per meter ($k = {}^{\omega}/_{c} = {}^{2\pi f}/_{c}$). The exponent is negative as the measured direction of arrival (DOA) of the wave is opposite to its direction of travel. A more complex sound field featuring a continuum of plane waves can be described as:

$$p(\mathbf{k}, \mathbf{r}) = \int_0^{2\pi} \int_0^{\pi} a(\mathbf{k}) e^{-j\mathbf{k}\cdot\mathbf{r}} \sin\theta_k d\theta_k d\phi_k \tag{2.19}$$

where $a(\mathbf{k})$ is the directional amplitude density, which quantifies the amplitude for wave vector $\mathbf{k}$, thus capturing the variations in amplitude for both frequency and DOA. Figure 2.6 shows a simplified sound field with two active sound sources.

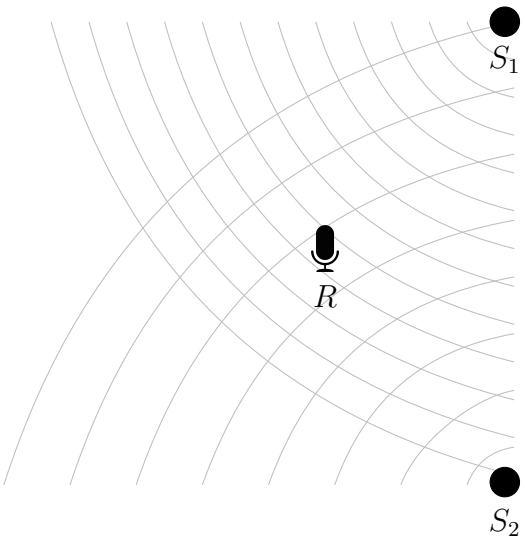
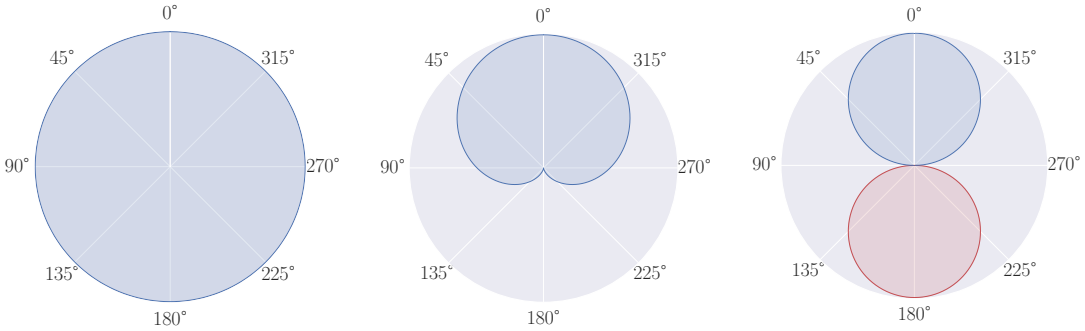


Figure 2.6: Diagram of a sound field based on basic model, featuring two point sources.

A recording of a sound field made with a single microphone capsule represents the integration of sound incident on the microphone capsule into a single signal with the contributions of sound from a given direction proportional to the microphone's polar pattern. A polar pattern defines the relative sensitivity of a microphone to sound incoming from any given direction. Figure 2.7 shows some common polar patterns found in a wide variety of microphones. Note that these are two-dimensional cross-sections of three-dimensional patterns. The red shading in the bidirectional pattern indicates a region of reversed polarity - the contribution of sounds incoming from this direction will be phase-inverted.

Consider an omnidirectional microphone employed to record the sound field shown in Figure 2.6 at position R. Since sources S_1 and S_2 are roughly equidistant from the microphone, assuming equal power for each source, each would be equally prominent in the recording regardless of microphone orientation. If, on the other hand, a cardioid microphone was used and oriented towards the top of the diagram, then source S_1 would sound more prominent in the recording than S_2. This leads to the intuitions that:



(a) Omnidirectional (b) Cardioid (c) Bidirectional

Figure 2.7: Two-dimensional cross-sections of three common microphone polar patterns (after [25]). Red portions indicate negative polarity.

1. A single-microphone recording encodes no spatial information about the scene.

2. If one were able to rotate a directional microphone to sample multiple directions simultaneously, a map of sound power varying over angle of incidence could be recorded.

2.4.1 Spherical Microphone Arrays

Whilst it is clear that the procedure of simultaneous directional sampling outlined above cannot be achieved in practical reality, it is possible to record the impact of a sound field surrounding a point in space using a spherical microphone array (SMA). SMAs sample the sound field with Q sensors arranged in a sphere. Though open spheres can be used, allowing free sound propagation between the microphones, more typical is the rigid sphere, with sensors mounted on a solid surface. The set of microphones samples the sound field at intervals, with the resolution of the spatial information determined by the number and arrangement of the sensors, together with the radius of the sphere. Typically, the raw microphone signals are encoded to give spherical harmonic (SH) signals. This encoding of a soundfield using spherical harmonics is known as Ambisonic format [26, 27, 28].

Spherical harmonics can be used to represent functions that are defined on the

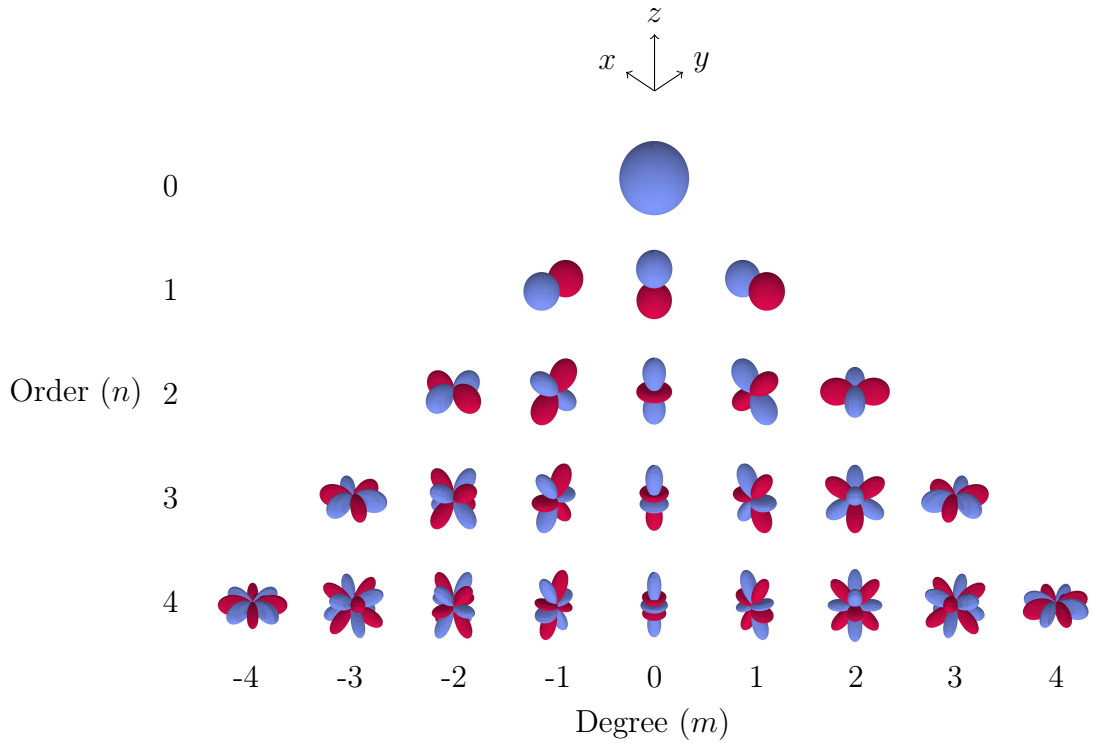


Figure 2.8: The spherical harmonics up to fourth-order (after [24]).

surface of a sphere. Figure 2.8 shows the spherical harmonic functions up to fourth-order, with reversed polarity lobes shaded red. They are the spherical basis functions, equivalent to the sinusoids used as basis for one-dimensional functions, as covered in Section 2.2.2. The spherical Fourier series is [24]:

$$f(\theta,\phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} f_n^m Y_n^m(\theta,\phi) \tag{2.20}$$

where θ and ϕ are spherical co-ordinates as defined in Section 2.4 and f_n^m are the weights for the spherical harmonic functions Y_n^m, defined as:

$$Y_n^m(\theta,\phi) \equiv \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{jm\phi} \tag{2.21}$$

Variables n and m are the SH *order* and *degree*, respectively, and P_n^m is the associated Legendre function [29]. Compare Equation 2.20 to the standard Fourier series in Equation 2.8. Just as higher-frequency sinusoids are required for higher-resolution

representations of 1-D functions (i.e. the square wave example in Figure 2.2), higher-order spherical harmonics are required for higher spatial resolution. Continuous spherical space-frequency domain functions can be mapped to the spherical harmonic domain using the spherical harmonic transform (SHT):

$$f_n^m(k) = \int_0^{2\pi} \int_0^{\pi} f(k, \theta, \phi) \left[Y_n^m(\theta, \phi) \right]^\star \sin\theta d\theta d\phi \tag{2.22}$$

where $\star$ represents the complex conjugate. Since microphone arrays provide samples of the continuous spherical function, the discrete form must be used [30, 31]:

$$f_n^m(k) = \sum_{q=1}^{Q} \alpha_q f(k, \theta_q, \phi_q) \frac{\left[Y_n^m(\theta_q, \phi_q) \right]^\star}{b_n(kr)} \tag{2.23}$$

where α_q are the quadrature weights governing the relative contribution of microphone q, and $b_n(kr)$ is the mode strength, describing the scattering of the sound field by the measurement sphere itself, which is dependant on the interaction between wavenumber k and the sphere radius r. For a given SH order N, it is required that $Q > (N+1)^2$.

The microphone array used in this work is the mh-Acoustics Eigenmike array [32], which employs a nearly-uniform arrangement of 32 microphone capsules flush-mounted on a rigid sphere, and can output SH signals up to fourth-order. Given uniform or nearly-uniform sampling, the quadrature weights become constant:

$$\alpha_q = \frac{4\pi}{Q} \tag{2.24}$$

and for a rigid sphere:

$$b_n(kr) = 4\pi j^n \left[j_n(kr) - \frac{j_n'(kr)}{h_n^{(2)'}(kr)} h_n^{(2)}(kr) \right] \tag{2.25}$$

with j_n being the spherical Bessel function, $h_n^{(2)}$ the spherical Hankel function of the second kind, and $'$ denoting their derivatives [24, 29]. Figure 2.9 shows the magnitude of the mode strength for spherical harmonic patterns, or 'beams', up to sixth-order. The upper x-axis labels show frequencies derived from kr for the

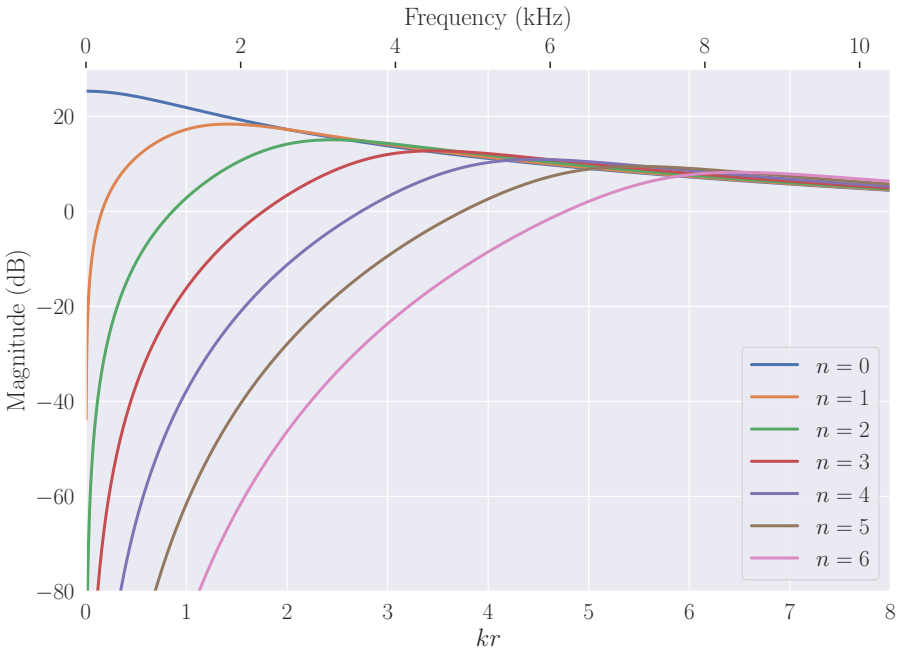


Figure 2.9: The magnitude of $b_n(kr)$ for orders up to $n = 6$. The upper frequency scale shows equivalent frequencies for an array radius of 4.2 cm (e.g. the Eigenmike).

Eigenmike radius $r = 4.2$ cm. It is apparent that the magnitudes of all orders $n > 0$ become very small as kr becomes significantly smaller than n. This has implications for the operational frequency range of the microphone array. Dividing out $b_n(kr)$ in Equation 2.23 to compensate for the array scattering causes large magnifications of the low frequencies in higher-order channels, which will also magnify any sensor noise present. The higher the order, the larger the increase in noise [24, 30]. This causes the lower frequencies in high order beams to degrade performance, and so these must be filtered out. The exact frequencies at which performance degradation becomes too great vary between arrays. Table 2.2 shows the lower cutoff frequencies for each order for the Eigenmike array [33].

The upper frequency limit for the reconstruction of the spherical function is restricted by spatial aliasing. Spatial aliasing occurs when a given SH order n cannot adequately describe high-frequency changes over space. Looking again at Figure 2.9 we can see that for $kr > n$ the contribution of the next highest SH

SH order n	**Lower cutoff frequency**
0, 1	30 Hz
2	400 Hz
3	1000 Hz
4	1800 Hz

Table 2.2: Lowest operating frequencies for Eigenmike per order [33].

order $n + 1$ is around 10 dB lower than that of n, with higher orders than this lower still. We can therefore consider the error incurred by the omission of these higher-order channels to be negligible at lower frequencies [24, 30]. For $kr > n$, on the other hand, the error becomes significant and the information contained in the higher-order channels is required to properly reconstruct the spherical function. The upper cutoff frequency should therefore satisfy $kr \leq N$ in order to avoid significant aliasing. For the Eigenmike, with $N = 4$, this makes the upper cutoff frequency 5.2 kHz, though mh Acoustics employ a proprietary 'high-frequency extension' stage in an attempt to maintain directivity above this limit [34].

2.4.2 Spherical Harmonic Beamforming

Recording the sound field as a series of spherical harmonics gives all the information required to make the hypothetical instantaneously-rotating directional microphone referred to in Section 2.4 a reality. It is one of the strengths of the spherical harmonic format that by simple weighting of the available SH channels, one can synthesise a polar pattern in any direction of interest and with a given directivity, within the limits of the maximum available SH order. This is known as beamforming.

The method by which a beam is synthesised is:

$$Z(\theta, \phi) = \sum_{k} \sum_{n=0}^{N} \sum_{m=-n}^{n} w_n f_n^m(k) Y_n^m(\theta, \phi) \tag{2.26}$$

Note that this equation is essentially identical to the description of the spherical

Fourier series in Equation 2.20, with added wavenumber dependence for f_n^m (as SH beamforming is typically done in the SH-wavenumber domain) and added weights w_n. In a sense, the SH-domain signals f_n^m are already weights for the spherical harmonics, and w_n are additional weights applied to modify the resultant beampattern. As such, the simplest approach is to set these weights at a constant value, so all spherical harmonics for a given *look direction* contribute equally. This has the effect of synthesising an axis-symmetric beam that approximates a Dirac δ function pointed in the look direction. Since plane waves incident upon the measurement array are in spherical harmonic representation equivalent to δ-functions at angles indicating their DOA, this technique can be thought of as decomposition of the sound field into a set of plane waves (as described in Equation 2.19) and therefore is typically termed plane wave decomposition (PWD) [35]. The constant weights yielding unity-gain PWD beams are derived in [24] as:

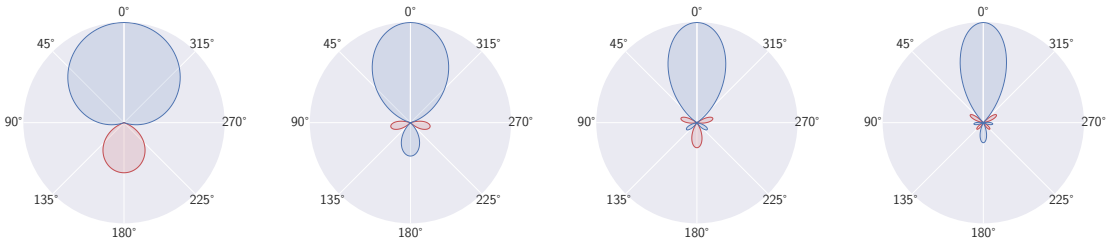
$$w^{\mathrm{PWD}} = \frac{4\pi}{(N+1)^2} \tag{2.27}$$

PWD yields the maximally-directive beam for any given order [36], at the expense of the presence of sidelobes arising from order-limited beampattern synthesis, analogous to the Gibbs overshoot phenomenon shown in Figure 2.2. The higher the spherical harmonic order used, the narrower the beam and the smaller the sidelobes. A theoretical beam synthesised using infinite orders would be infinitely narrow with no sidelobes. Figure 2.10 shows the beampatterns synthesised using PWD from 1st to 4th-order.

An alternative to PWD that is often used when decoding a SH signal to a loudspeaker array for playback (Ambisonic decoding) is the max-r_E weighting [37]. This weighting is designed to maximise energy in the look direction, producing minimal sidelobes at the expense of a wider main lobe [38]. The weights are calculated by:

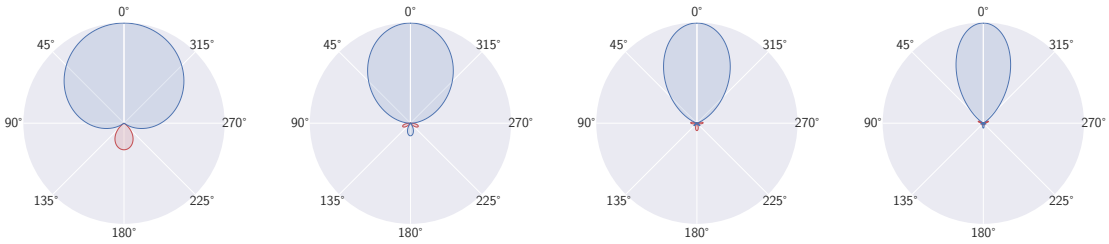
$$w_n^{\text{max-rE}} = P_n(E) \tag{2.28}$$

where P_n is the Legendre polynomial for order n and E is the largest root of P_{N+1},



(a) N=1 (b) N=2 (c) N=3 (d) N=4

Figure 2.10: Cross-sections of PWD beampatterns from 1st to 4th-order.



(a) N=1 (b) N=2 (c) N=3 (d) N=4

Figure 2.11: Cross-sections of max-r_E beampatterns from 1st to 4th-order.

approximately [39]:

$$E \approx \cos\left(\frac{137.9^\circ}{N+1.51}\right) \tag{2.29}$$

Figure 2.11 shows beampatterns synthesised using max-r_E weights.

2.4.3 Cross-Pattern Coherence

A recently-developed technique enabling much narrower beams than would be achievable using either the PWD or max-r_E weightings at any given order $N \geq 2$ is cross-pattern coherence (CroPaC) [31, 40]. At the expense of increased computational complexity, this technique utilises coherence between beams synthesised from multiple SH orders, combined with axis-symmetric rotations, to achieve complete sidelobe suppression. To generate CroPaC beampatterns, two beams are synthesised for orders N and $N-1$ with initial weights w_n set to select individual SH channels where

$m = n$, e.g.:

$$\begin{aligned} w_1 &= \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \dots \end{bmatrix} \\ w_2 &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \dots \end{bmatrix} \end{aligned} \tag{2.30}$$

Since these weights select individual spherical harmonic patterns, which are not symmetric in all axes, Wigner-D rotation (described in Section 2.4.4) is used to steer these beams to the look direction. Outputs Z_N and Z_{N-1} are calculated according to Equation 2.26 with the summation over k omitted, retaining the wavenumber dependence at this stage. Next, the cross-spectrum of these two beampatterns is calculated:

$$Z_{\text{cross-spectrum}}(\theta, \phi, k) = \mathbb{R}\left[Z_N(\theta, \phi, k)^\star Z_{N-1}(\theta, \phi, k)\right] \tag{2.31}$$

As is shown in Figure 2.12, this results in a dipolar beampattern with a negative pole on the opposite side to the desired look direction. A half-wave rectification is used to suppress sounds incoming from this negative side:

$$Z_{\text{CroPaC}} = \max\left[0,\ Z_{\text{cross-spectrum}}\right] \tag{2.32}$$

Figure 2.13 shows the results of this stage. Although the main-lobe of each pattern is noticeably narrower than the PWD or max-r_E equivalents shown in Figures 2.10 and 2.11, there are very prominent sidelobes which would cause unwanted contribution from off-axis sounds. Therefore, in order to suppress these sidelobes, an additional series of rotations are performed about an axis defined by the look direction (θ, ϕ). Taking the product of these rotations results in a final beampattern free of sidelobes, as only the parts of the beampattern common to all rotated versions (i.e. the mainlobe in the look direction) are retained:

$$Z_{\text{Suppressed}}(\theta, \phi, k) = \prod_{n-1}^{N} \Delta_n \left(Z_{\text{CroPaC}}(\theta, \phi, k)\right) \tag{2.33}$$

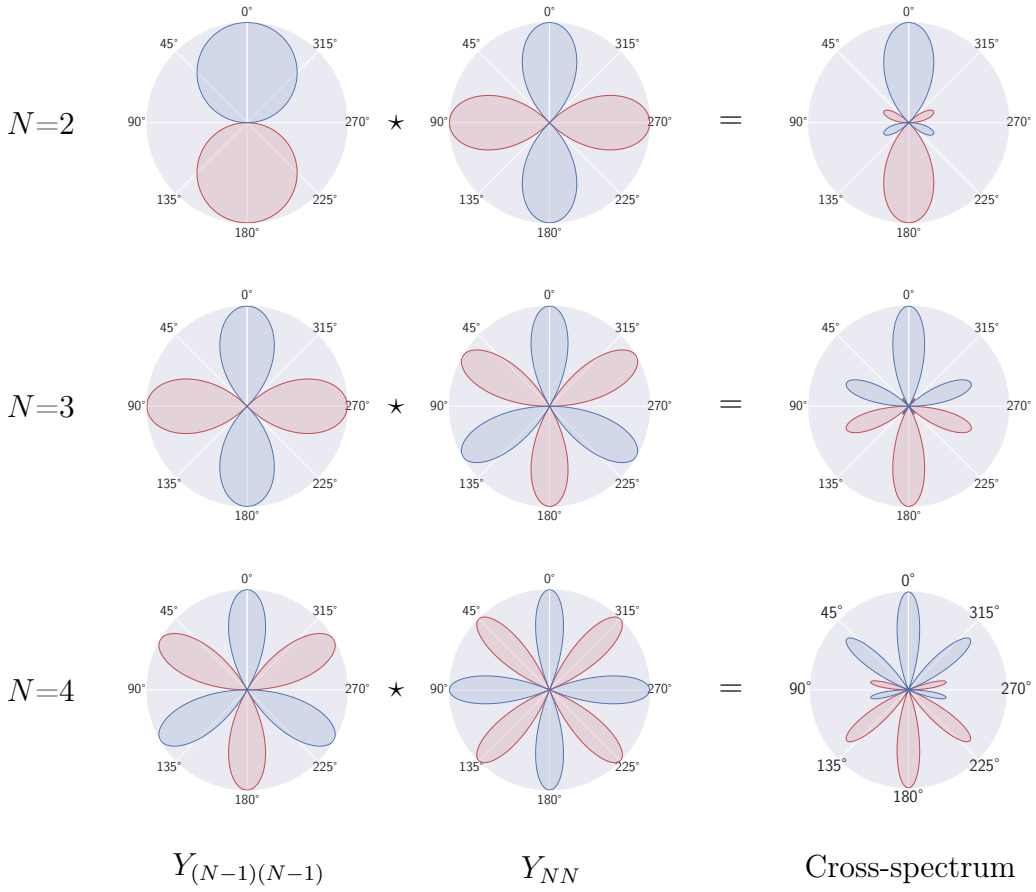
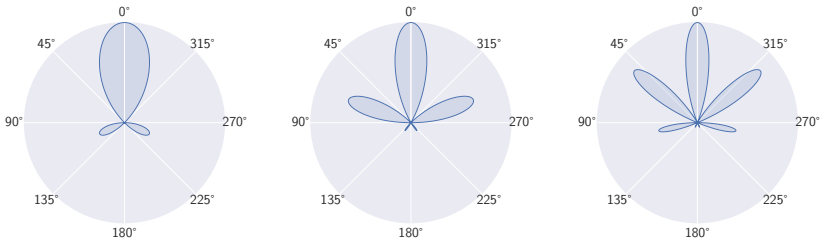


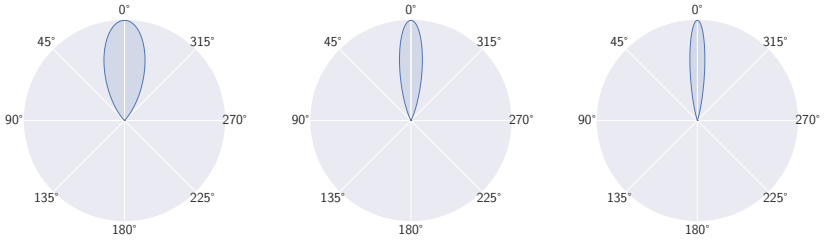
Figure 2.12: Calculation of cross-pattern coherence patterns for $N = \{2, 3, 4\}$. Note that the set of patterns used to calculate the cross-spectrum shown here are 2D cross-sections of the same 3D SH functions shown in Figure 2.8.

where Δ_n represents a rotation of $\frac{n\pi}{N}$ radians. In practise, these rotations are included as part of the process of steering the beampattern weights. Equations 2.31 and 2.32 must effectively be used N times to calculate the rotated versions of the beampattern for sidelobe suppression using Equation 2.33. The final CroPaC beampatterns are shown in Figure 2.14.



(a) N=2 (b) N=3 (c) N=4

Figure 2.13: Half-wave rectified cross-pattern coherence patterns for $N = \{2, 3, 4\}$.



(a) N=2 (b) N=3 (c) N=4

Figure 2.14: CroPaC beampatterns for $N = \{2, 3, 4\}$.

2.4.4 Rotation of Spherical Harmonic Functions

For axis-symmetric beampatterns such as PWD and max-r_E, all that is required to steer the beams is the summation of the weighted SHs for a given look direction. CroPaC, on the other hand, utilises non axis-symmetric beamforming, and as such it is necessary to describe here the process by which any arbitrary spherical harmonic function can be rotated.

A rotation Δ of a complex spherical harmonic function for a given n and m can be defined as a weighted sum of complex spherical harmonics of the same order n and that order's full range of degrees, m', e.g.:

$$\Delta(\alpha, \beta, \gamma) Y_n^m(\theta, \phi) = \sum_{m'=-n}^{n} D_n^{m'm}(\alpha, \beta, \gamma) Y_n^{m'}(\theta, \phi) \tag{2.34}$$

where α, β and γ are the Euler rotation angles. The rotation weights are given by the Wigner-D function [24]:

$$D_n^{m'm}(\alpha,\beta,\gamma) = e^{-jm'\alpha} d_n^{m'm}(\beta) e^{-jm\gamma} \tag{2.35}$$

where $d_n^{m'm}(\beta)$ is the Wigner-d function:

$$\begin{aligned} d_n^{m'm}(\beta) &= \zeta^{m'm}\sqrt{\frac{s!(s+\mu+\nu)!}{(s+\mu)!(s+\nu)!}}\sin(\beta/2)^{\mu}\cos(\beta/2)^{\nu}P_s^{(\mu,\nu)}(\cos\beta), \\ \mu &= |m'-m|, \quad \nu = |m'+m|, \quad s = n - {}^{\mu+\nu}/_2, \\ \zeta^{m'm} &= \begin{cases} (-1)^{m-m'}, & \text{if } m < m'. \\ 1, & \text{otherwise.} \end{cases} \end{aligned} \tag{2.36}$$

and P is the Jacobi polynomial [41]. Since functions on a sphere can be represented by different weightings of spherical harmonics (Equation 2.20) it follows that rotations of functions represented in this way effectively require a redistribution of these weights to describe the same pattern in a different orientation [24]. This means a function represented by spherical harmonics from a single order n will not require SHs of any other order to represent its rotation. Usually, however, a function will require rotation across multiple orders:

$$\hat{f}_n^m = \sum_{n=0}^{N}\sum_{m'=-n}^{n}\sum_{m=-n}^{n} f_n^m D_n^{m'm} \tag{2.37}$$

The directional dependency of $D_n^{m'm}$ has been omitted here for clarity. This can be expressed as a matrix operation:

$$\hat{\mathbf{f}} = \mathbf{D}\mathbf{f} \tag{2.38}$$

$\mathbf{f}$ is a vector containing the spherical harmonic weights describing the function:

$$\begin{bmatrix} f_{00} & f_{1(-1)} & f_{10} & f_{11} & \cdots & f_{NN} \end{bmatrix}^T \tag{2.39}$$

where T is the array transpose operation. $\mathbf{D}$ is a block diagonal matrix:

$$\begin{bmatrix} \mathbf{D}_0 & 0 & 0 & \dots & 0 \\ 0 & \mathbf{D}_1 & 0 & \dots & 0 \\ 0 & 0 & \mathbf{D}_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{D}_N \end{bmatrix} \tag{2.40}$$

The matrices $\mathbf{D}_n$ comprise of entries covering $D_n^{m'm}$, e.g., for $N = 1$:

$$\begin{bmatrix} D_1^{(-1)(-1)} & D_1^{(-1)0} & D_1^{(-1)1} \\ D_1^{0(-1)} & D_1^{00} & D_1^{01} \\ D_1^{1(-1)} & D_1^{10} & D_1^{11} \end{bmatrix} \tag{2.41}$$

The above equations describe rotation matrices for complex spherical harmonic functions. For real-valued spherical harmonics, a method is required to transform these functions to real values. The real-valued rotation matrices $\mathbf{D}^{\mathbb{R}}$ can be calculated as follows [42]:

$$\mathbf{D}^{\mathbb{R}} = \mathbf{C}^{\star}\mathbf{D}\mathbf{C}^{T} \tag{2.42}$$

where $\mathbf{C}$ is a block diagonal matrix (similar to that shown in Equation 2.40) consisting of elements $\mathbf{C}_n$. These are matrices constructed similarly to that shown in Equation 2.41, comprising entries covering $C_n^{m'm}$, which is defined as:

$$C_n^{m'm} = \frac{1}{\sqrt{2}} \begin{cases} 0 & |m'| \neq |m| \\ \sqrt{2} & m' = m = 0. \\ (-1)^m & m' = m > 0. \\ 1 & m' > 0 > m. \\ -j(-1)^m & m' < 0 < m. \\ j & m' = m < 0. \end{cases} \tag{2.43}$$

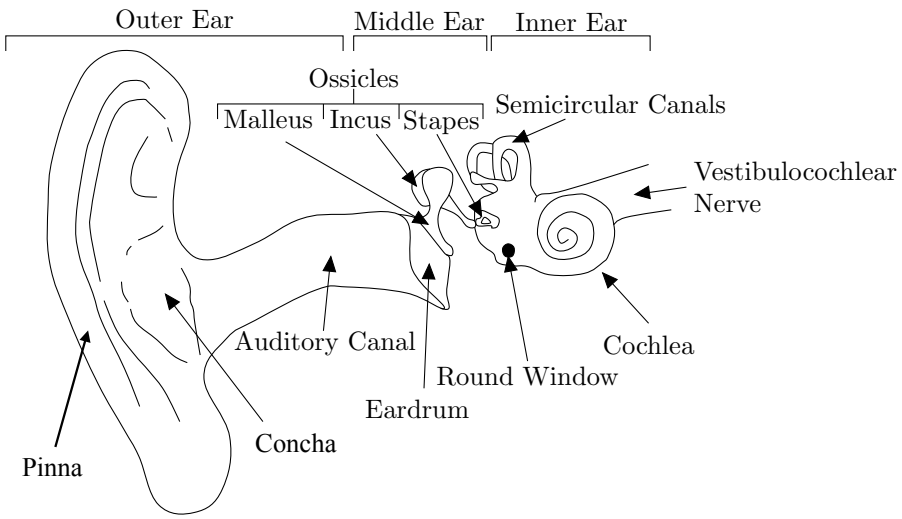


Figure 2.15: The anatomy of the human ear (from [20]).

2.5 Human Hearing

So far, the underlying physics of sound waves and sound fields have been explored, yet equally important to understand in any study involving sound perception is the way in which a sound field interacts with the human hearing system, making its various facets perceptible to a listener.

2.5.1 The Ear

Central to the human hearing system is, of course, the ear. Figure 2.15 is a cross-section of a human ear, indicating its three main sections; outer, middle and inner [11]. The outer ear consists of the pinna, concha, auditory canal, and tympanic membrane, or eardrum. The pinna is the large fold of tissue visible on the head, and serves to direct sound down to the concha, which is the entrance to the auditory canal. The pinna and concha also have the effect of enhancing particular frequencies depending on the DOA of incoming sound. This is a key component of the human ability to localise sounds, explored further in Section 2.5.2. Sound travels along the auditory canal and induces vibrations in the tympanic membrane, which is

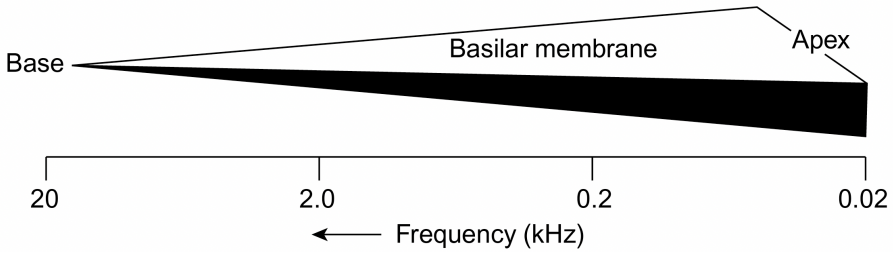


Figure 2.16: Simplified diagram of the uncoiled basilar membrane, indicating its resonances across the range of human hearing (from [11]).

the threshold of the middle ear. The middle ear consists of three small bones, the ossicles, that transmit the movements of the tympanic membrane to the oval window, entrance to the cochlea and inner ear [11]. The cochlea is a coiled tube, with oval and round windows at its base. Contained within the cochlea is the basilar membrane, a simplified diagram of which is shown in Figure 2.16. This membrane resonates at frequencies ranging from 20 Hz at its apex to 20 kHz at its base, stimulating tiny hair cells contained in the organ of Corti, distributed along the length of the membrane. These hair cells trigger the vestibulocochlear nerve to transmit information to the brain [11, 20].

This system by which the *place* of vibration on the cochlea is key to differentiation of frequencies gives rise to the phenomenon of critical bands, effectively determining the frequency resolution of human hearing. If the frequency components of an incoming sound are relatively well-spaced, the vibrations on the basilar membrane will be distinct. If, however, frequencies are closely spaced, vibrations on the membrane will not be definitely separated, giving rise to a perceptual 'roughness' [11]. The required frequency difference for perceptual separation is known as the 'critical bandwidth'. This phenomenon is key the design of many perceptually-motivated machine listening features, some of which will be covered in Chapter 4.

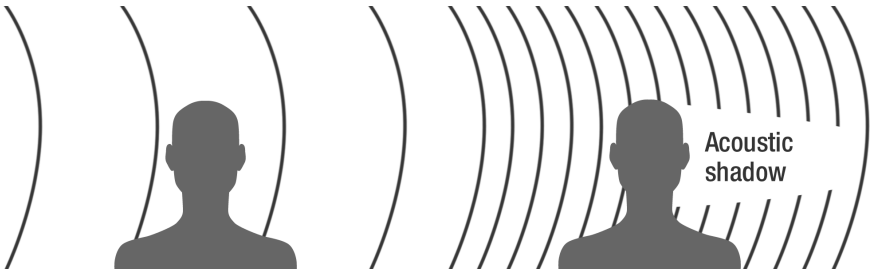


Figure 2.17: Acoustic shadowing by the head causing ILD at higher frequencies (from [43]).

2.5.2 Spatial Hearing

As noted in Section 2.4, a single receiver encodes no spatial information on sounds in an environment. Thus, two ears are required for the perception of spatial sound. Now that the basic functioning of a single ear has been established, the manner by which the signals from both ears are used in tandem to perceive these spatial properties can be explored.

The two main cues that can be derived from binaural hearing are the interaural level difference (ILD) and interaural time difference (ITD) [11]. ILDs arise when the head creates an acoustic shadow for incoming sound waves, which causes the sound arriving at the ear in shadow to be attenuated. Following the principles of diffraction described in Section 2.3.3, this effect is greater at frequencies where the wavelength of the sound is smaller than the head, as shown in Figure 2.17, and virtually nonexistent for low-frequency sounds where the wavelength is larger than the head. At low frequencies, therefore, ITD is a vital cue for localisation [11]. Figure 2.18 shows a simplified model of ITD. From this, estimates can be derived as [11]:

$$\text{ITD} = \frac{r\left(\theta + \sin(\theta)\right)}{c} \tag{2.44}$$

This can be used to derive a maximum possible ITD value (for sounds arriving at $\theta = 90°$) of 6.6×10^{-4} s. This will result in ITD ambiguities for frequencies

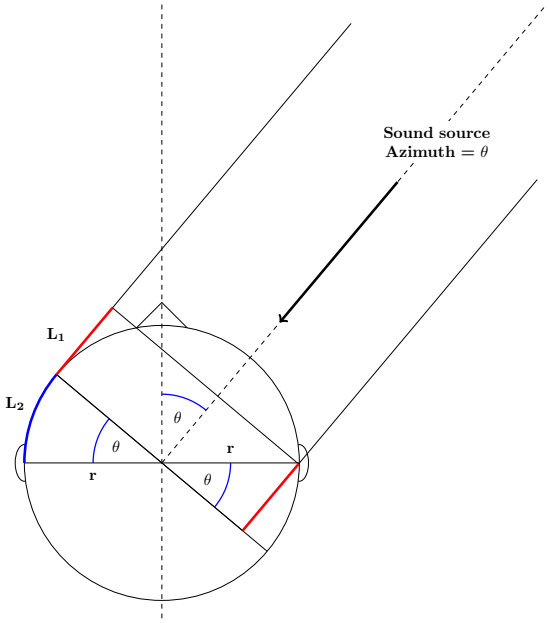


Figure 2.18: Simplified model of ITD (from [20]).

completing more than one vibration in this period of time; approximately 1.5 kHz and above. In practise, then, ITD cues are more important for low frequencies, and ILD for high frequencies.

Whilst ILD and ITD contain a great deal of spatial information, they can only account for perception of direction on the horizontal plane, and even then do not account for the human ability to tell whether sounds are coming from in front or behind. The direction-dependent filtering of the pinna mentioned in Section 2.5.1 is key in both these regards. Subtle head movements are also used to disambiguate front from rear [11].

All of the spatial cues described in this section can be encoded as a head-related transfer function (HRTF) [44]. These are recorded along the lines of the method used for RIRs described in Section 2.3.2, using a pair of small microphones placed at the entrance to each ear canal, typically in an anechoic chamber. Separate HRTFs must be recorded for every DOA of interest. A high spatial resolution is required for realistic binaural synthesis, so this can be a very time-consuming process.

2.6 Summary

This chapter has provided the foundation for the technical aspects of the work presented in this thesis, starting with the basic makeup of sound waves and describing the nature of their propagation through space, and how this builds up a sound field. It was shown that just as a one-dimensional signal can be represented and analysed using a Fourier series based on sinusoids, a spherical signal can be represented using spherical harmonics, and as such these make an ideal basis with which to record a sound field spatially. The chapter concluded with a short summary of the hearing system, the means by which spatial sound is perceived. The next chapter will explore the more human-centric concerns of soundscape and environmental noise, which provide the motivation for the work presented in this thesis.

3 | Environmental Noise and the Soundscape Approach

3.1 Introduction

It is a fact so obvious that it is not often stated that we are constantly immersed in sound. Whilst vision can be shut off simply by closing one's eyes, sound is not so easily shut off. Even modern noise-cancelling headphones, whilst very effective, do not ever completely block all sound. Given such constant stimulus, it should be clear that sound can have a profound effect on both physical and mental wellbeing, though in fact it is this very ubiquity that often leaves sound taken for granted and regarded as a secondary concern.

In the previous chapter, the scientific grounding of sound fields was explored. It was shown that in basic terms sounds vary only in amplitude and frequency, with spatial interactions and variations thereof. It was also shown that it is relatively straightforward given modern technology to measure sound across these dimensions. To quantify human response to these apparently simple variations is much more difficult. There are physiological dimensions, such as noise-induced hearing loss, that are easy to define using standard SPL measurement [11]. Psychological effects (which can in the long term lead to physiological effects) are much harder to evaluate. How can the fact that the sounds of a bustling city with cars, honking horns, chatter and music can energise and enthuse one person, yet enrage the next, be quantified?

Can a simple number be put to the calming effect of a babbling brook, that even the distant intrusion of a single passing vehicle [45] could shatter?

This chapter will explore the concept of the soundscape approach, which arose partially as an attempt to answer these kinds of questions. The prevailing environmental noise approach will also be summarised, with the two compared and contrasted by way of an exploration of the terminology and language used in each. Various taxonomies for sound categorisations will be investigated and assessed in relation to their applicability to machine listening and human perception, and the chapter will conclude with some methods by which the soundscape approach has been implemented in practise.

3.2 Terminology

To consider the concept of *soundscape*, especially as related to the field of machine listening, one must first return to the source of many of the ideas that pervade the literature, and tackle the surprisingly challenging issue of the terminology used in this field of study. The notion of 'The Soundscape' was crystallised by R. Murray Schafer in his seminal book of the same name [3]. Though the exploration of the subject in this work in exhaustive, a single formal definition of the term is elusive. Schafer presents a number of potential definitions in quick succession:

> "The soundscape is any acoustic field of study. We may speak of a musical composition as a soundscape, or a radio program as a soundscape or an acoustic environment as a soundscape" [3].

Whilst it makes sense to be inclusive in terms of the kinds of sounds encompassed by the term, the main focus of soundscape studies has been on the sounds present in urban environments [5, 46, 47], and in wilderness environments, where often the terms 'acoustic ecology' or 'soundecology' are used alongside 'soundscape' [48, 49, 50]. Schafer's text converges on the metaphor of the soundscape as the

auditory equivalent to the visual concept of the landscape, with an emphasis on the soundscape as a product of human perception [3]. The formal definition of *soundscape*, published as an international standard in ISO 12913-1 [51], extrapolates from the European Landscape Convention Agreement's definition of *landscape*, thus arriving at:

> "the acoustic environment of a place, as perceived by people, whose character is the result of the action and interaction of natural and/or human factors" [6].

Although this definition was not universally agreed upon [52], it is prevalent in the literature and since it is now enshrined as an international standard, this thesis will use the term based on this definition.

Since a soundscape depends on perception, it is unique to each individual experiencing a particular place. The raw input to a machine listening system will not have been filtered by any such perceptual mechanism, and will instead be an audio stream dependent on an environment and altered only by the properties and location of the microphones used to capture the sound. It can therefore not be considered a soundscape according to the above definition, so alternative terminology is needed. One option is to use the term 'auditory scene', borrowing from computational auditory scene analysis (CASA). However, by using the word 'auditory', that is "Relating to the sense of hearing" [53], 'auditory scene' still implies a certain focus on the *perception* of sound.

Many alternative terms have been used in the literature, most prevalently 'acoustic scene' [54], and sometimes 'sound scene' [9, 55] in relation to recent machine listening work. In some respects, the term 'sound scene' might be considered less ambiguous than 'acoustic scene', as the latter might perhaps be taken as a reference to room acoustics. On the other hand, the concept of 'sound' could be taken as perceptual. Together with the semantic similarity, this could make the term 'sound scene' more easily confusable with 'soundscape', so it will not be used in this thesis.

Noise Approach	**Soundscape Approach**
Sound managed as a waste	Sound perceived as a resource
Focus is on sounds of discomfort	Focus is on sounds of preference
Integrates all sounds at a receptor	Differentiates between sound sources
Managed by reducing levels	Works towards wanted sounds not being masked by unwanted sounds

Table 3.1: Environmental Noise Approach vs Soundscape Approach (after [4]).

Instead, this thesis will generally use the term 'acoustic environment', defined by Brown as "the sound from all sources that could be heard by someone in that place" [6]. An acoustic environment is characterised by the kinds of sound sources present, together with the reshaping and colouration that are given to those sources by reflections, absorption, diffractive bending of sound waves, or any other modifications happening within the vicinity. Hence, 'acoustic scene' will be used when referring to ASC, but 'acoustic environment' is generally preferred due to its explicit referencing within the ISO definition of 'soundscape'.

3.3 Environmental Noise

Whilst the choice of terminology to describe this subject might seem a somewhat trivial starting point for this discussion, the particulars of the language used can have an effect on the wider framing of the issue. This is exemplified in the writing of Brown, who categorises the prevailing approach to managing the acoustic environment as the 'environmental noise' approach, calling this the "traditional, objective energy-based model of the acoustic environment", as opposed to the more nuanced 'soundscape' approach, a "subjective listener-centred model" [4]. Table 3.1 presents some key differentiators between these two approaches.

From the point of view of an audio researcher, noise is a specific type of signal,

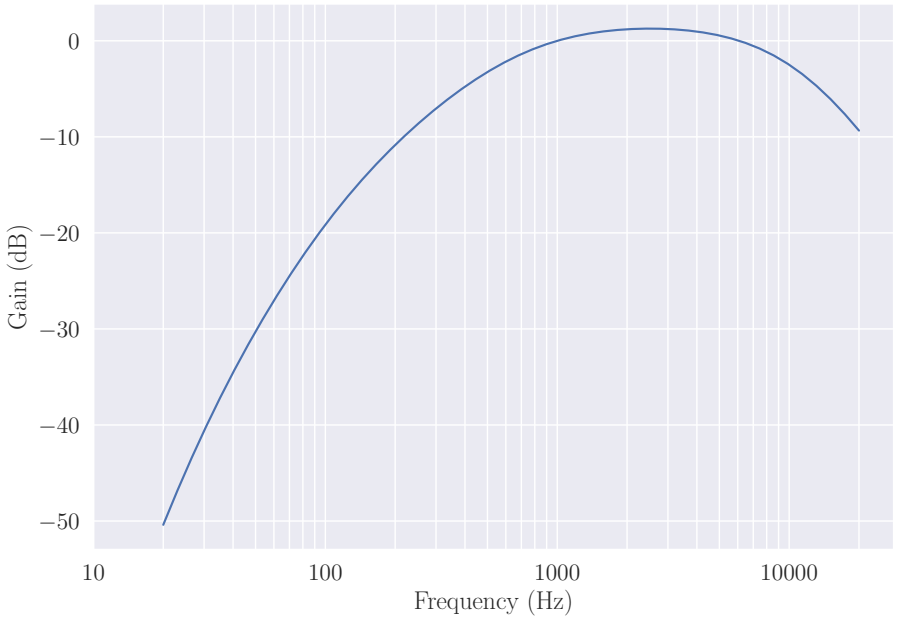


Figure 3.1: The 'A' weighting for SPL measurement.

but to members of the general public the terms *noise* and *sound* may be much more synonymous. 'Environmental Noise' - a term that is used in much of the legislation on this matter [56, 57] - feels very reductive in both senses, conflating all environmental sound into an 'unwanted' mental space. This is borne out by the standard metric for environmental noise, the L_{Aeq} measurement, which integrates all sounds present in a scene into a single one-dimensional reading.

3.3.1 L_{Aeq}

$L_{\text{Aeq, T}}$ denotes the equivalent continuous SPL 'dose' required to give the same total amount of sound pressure received over a designated period of time T at the measurement point, essentially the average SPL [11, 43]. The 'A' indicates A-weighting, which is an international standard scale designed to compensate for the fact that human perception of loudness is not flat across all frequencies, especially at lower amplitudes [58]. A-weighting is defined to normalise the perceived loudness of tones relative to the perception of a 1 kHz tone at a given SPL, as shown in Figure 3.1.

Most SPL meters take A-weighting into account by default, and A-weighted readings are denoted by the unit dB(A).

L_{Aeq} is an important measurement for protection of hearing, as the amount of hearing damage incurred is a function of both SPL and exposure time. Exposure to loud sounds upwards of 100 dB(A) can cause permanent hearing damage in a matter of minutes, or even seconds as the SPL increases. Lower level sounds on the order of 85 dB(A) can be tolerated for several hours before hearing is permanently affected [11]. This is reflected in UK legislation covering workplace noise exposure, which states that, if the daily workplace noise dose ($L_{\text{Aeq, 8h}}$) exceeds 85 dB(A), employees must be provided with hearing protection, and that if the dose exceeds 80 dB(A) they are entitled to request it [11, 59].

Ambient noise in urban environments is often measured over a 24-hour period ($L_{\text{Aeq, 24h}}$). The level considered by legislation to cause annoyance is considerably lower, at 50-55 dB(A), than that which can cause increased risk of physiological health effects, which for a 24-hour period is 70 dB(A) [60]. The 24-hour SPL measurement is sometimes weighted to reflect temporal variations in annoyance and disturbance levels based on typically lower activity levels in evening and night periods. This measurement is denoted L_{den} (day-evening-night) and is calculated as [61]:

$$L_{\text{den}} = 10 \log \frac{1}{24} \left(12 \cdot 10^{\frac{L_{\text{day}}}{10}} + 4 \cdot 10^{\frac{L_{\text{evening}}+5}{10}} + 8 \cdot 10^{\frac{L_{\text{night}}+10}{10}} \right) \tag{3.1}$$

This applies a 5 dB penalty to noise occurring in the evening hours (7 pm to 11 pm) and a 10 dB penalty to the night hours (11 pm to 7 am), in relation to noise occurring during the day.

There is no doubt in the usefulness of measuring L_{Aeq}, especially in terms of the prevention of hearing loss and as a broad indicator of annoyance. The desire to measure sound in an objective manner such as this is understandable when one considers the motivation of devising universally-applicable laws that are relatively simple to implement. The problem with the approach is that, as shall be explored

in Section 3.4.3, and as outlined in the introduction to this thesis, the emotional and physiological responses to sound stimuli cannot be completely quantified by the sound level alone. L_{den} represents an acknowledgement that perception of sound levels is context-sensitive, but this is still a level-based metric that treats all sound as inherently of equal potential disturbance. For a more nuanced understanding, L_{Aeq} could be considered one component of a broader multi-variate contextual analysis.

Brown's definition of the soundscape approach frames the technique as inherently subjective, and indeed perceptions of a given soundscape can vary greatly from person to person. There are, however, several broad patterns in soundscape perception whereby advanced measurement and machine learning technology could be brought to bear in more discerning measures than L_{Aeq}. Of particular interest is the aspect of the soundscape approach that "requires differentiation between sound sources" [4]. This is an area where a machine listening system could be very useful.

3.4 Soundscape Taxonomies

3.4.1 Schafer's Features of the Soundscape

Whilst soundscapes tend to be considered as a single entity, they are always built up from distinct sonic events. There have been numerous systems proposed for the categorisation of the kinds of sounds that contribute to a soundscape, beginning with Schafer's original work on the subject, which proposed a four-category system:

- **Keynotes** - The ever-present sounds that define the character of the soundscape to which they belong.
- **Signals** - Sounds that are actively listened to and convey meaning or messages.
- **Soundmarks** - The auditory equivalent to landmarks - sounds specific to the location in which they are heard.

- **Archetypal sounds** - "mysterious ancient sounds, often possessing felicitous symbolism" [3].

Keynote is a term borrowed from music, identifying the tone that gives the key signature of a composition. One might also call to mind the keynote talk at a conference, intended to 'set the tone' for the rest of the proceedings. According to Schafer, keynote sounds are not typically listened to consciously, but serve as the background [3, 62], underpinning, colouring, and contextualising all other sounds in a scene. These are the kinds of sounds one might not notice until they are gone. Schafer identifies keynote sounds as typically "created by geography and climate", though the sounds of machinery, in particular the motor car, now fulfil this role in many modern urban soundscapes [63, 64].

In describing *signals* Schafer invokes Gestalt visual perception theories, positioning signals as the 'figure' to the keynote's 'ground' [3]. Schafer posits that any sound source can be a signal if a listener consciously chooses to direct their attention towards it, though some sounds typically draw attention more than others. In a modern urban soundscape, police sirens or the alert sounds made by pedestrian crossings are examples of signals.

Keynotes and signals could be thought of as the main categories, with soundmarks and archetypal sounds being special cases of either. In the city of York this is typified by the famous Minster bells. These are a soundmark, a sound characteristic to the city which cannot be found anywhere else, whilst also exemplifying an archetypal sound of antiquity, a church having been on the site of the Minster for more than a thousand years. In a city with such depth of history, this is a sure qualification of "felicitous symbolism". Further, Schafer specifies bells as a type of signal as they convey warning or information (perhaps the time, or a call to prayer), whereas their ubiquity in York could even see them considered part of the city's keynote sound. Schafer notes that keynote sounds "may have imprinted themselves so deeply on the people hearing them that life without them would be sensed as a distinct impoverishment". This was exemplified by the widespread outrage at the

2016 dismissal of the Minster bell ringers, causing the bells to fall silent [65].

3.4.2 Acoustic Environment Classification

Whilst there is clear benefit to Schafer's categorisations when considering soundscapes, they are inherently subjective and therefore unsuitable as labels that could be objectively applied to sounds by a machine listening system. There is no guarantee that two people will perceive the same things when hearing an acoustic environment. One person's signal could be another person's keynote, and in fact these perceptions could change from moment to moment depending on which sound is given special attention. It is clear that the objective analysis of acoustic environments requires a less subjective taxonomy.

One categorisation system, originally proposed for research into biodiversity [66], consists of three categories into which all sounds can be fitted:

- **Anthrophony** - Sound created by human activity.
  E.g. Vehicle noises, machinery, music...
- **Biophony** - Sounds produced by biological activity (non-human).
  E.g. Birdsong, animal calls...
- **Geophony** - Sounds originating in non-biological natural processes.
  E.g. Running water, wind, ocean waves...

These groupings reflect their origins in soundecology research. In wilderness environments it is usually a fair assumption that the majority of sound sources will be non-human in origin. Since much research in soundecology is concerned with quantifying the degree of human encroachment upon these environments, it is entirely reasonable in this context to take a low-resolution view of human sound as a single agglomeration. In soundscape research, however, the focus is on human perception of sound, which, as shall be explored in Section 3.4.3, tends not to be uniform across all the sounds included under the umbrella of anthrophony. Furthermore,

since the sound sources in urban environments are overwhelmingly anthrophonic, a finer resolution is needed for these, with perhaps less detail required for biophonic and geophonic sounds.

Additional shortcomings of directly mapping soundecology techniques onto urban soundscape analysis can be seen in the recent work of Devos [50], in which soundecology indicators (summarised in [67]) are applied to urban soundscape recordings. Of particular interest is the Normalised Difference Soundscape Index (NDSI), calculated using

$$\text{NDSI} = (\beta - \alpha)/(\beta + \alpha) \tag{3.2}$$

where β and α represent estimations of the biophonic and anthrophonic sound power respectively, averaged over a complete recorded signal. The output is a number ranging between -1 to 1, where negative numbers indicate more prevalent anthrophony and positive numbers indicate more prevalent biophony. To gain estimates for β and α, a system is used whereby the frequency spectrum is split into two bands. In [67] these are 200 Hz - 2 kHz, assumed to contain mainly anthrophonic sound, and 2 kHz - 8 kHz, assumed to consist mainly of biophonic sound. It is clear that this method of separating sound sources is extremely crude, with a great deal of anthrophonic sound being much more wide-band and many animals producing sound well below 2 kHz. This drawback is noted in much of the literature that uses the metric, but NDSI has nevertheless been shown to give usable results in analysis of wilderness environments [49, 67].

Devos' work shows that this technique fails when attempting to measure NDSI in urban environments [50]. Several isolated anthrophonic sounds, including hedge trimmers and mopeds, give incorrect output from the NDSI metric. The paper states that the technique fails in urban environments due to the large number of sound sources present. Despite the failure of the soundecology indicator in this case, the research does at least represent an attempt to glean more meaningful objective measures from an acoustic environment recording than L_{Aeq}. This raises

the question of whether a machine listening system could be used to improve the accuracy of the NDSI or similar metric. Devos' paper does in fact mention using a "computational model of auditory attention" as a potential area for future work.

Arguing that there is "insufficient resolution" [6] for urban acoustic environments using the soundecology categories, Brown *et al.* developed an acoustic environment schema, shown in Figure 3.2, that divides anthrophony into various subcategories [6, 68]. Brown's schema could make a good hierarchical categorisation system for an automated acoustic environment analysis system. The *air traffic* class, for instance, would be a subclass of 'Motorised Transport', itself a subclass of *Anthrophonic* sound. Such a system could give figures for the contribution of individual sound sources or categories [69]. This would facilitate the calculation of metrics such as a more reliable NDSI ratio, and hierarchical classification could also work to the advantage of an imperfect system. Certain sounds that might be more difficult to tell apart and more prone to misclassification - perhaps different types of motor vehicles - would likely still be classed within 'motorised transport', allowing metrics involving this superclass to be calculated without issue. This approach is proposed in [70] for the Instrument for Soundscape Recognition, Identification and Evaluation (ISRIE) system, though in that work a slightly expanded version of the basic soundecology categories is used, rather than Brown's schema, which was published later.

3.4.3 Perceptual Dimensions

Having investigated Schafer's soundscape features, which are perceptually relevant but difficult to objectively quantify, the question arises of the perceptual relevance of the quantifiable soundecology categories. In some studies, after experiencing an acoustic environment by some means (see Section 3.5), participants are interviewed on their perception of these as soundscapes [71]. Whilst the depth of insight into individual experiences gained this way can be valuable, it is difficult to map this kind of data to numerical values that can be analysed statistically. For this reason, most recent studies use semantic differentials, characterised by numerical scales positioned

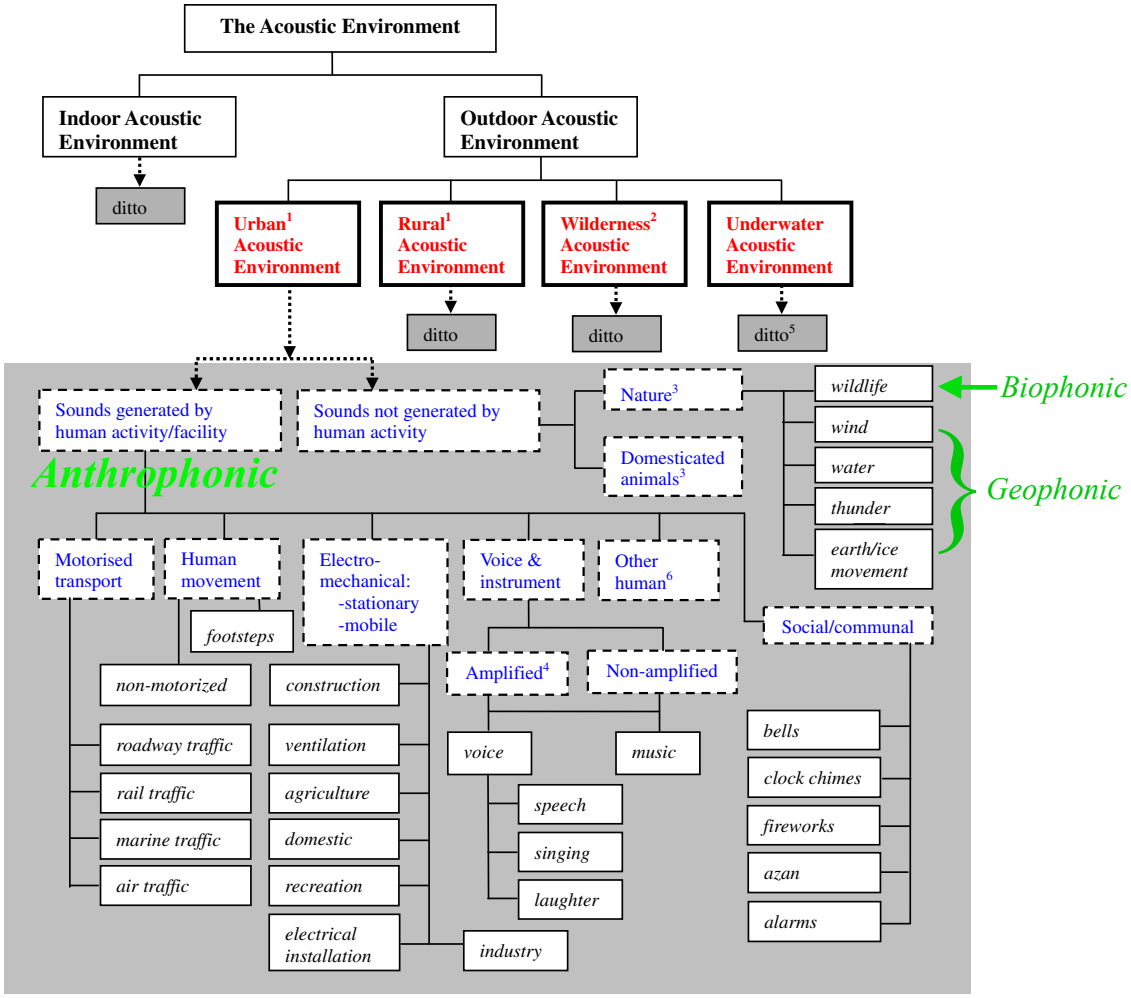


Figure 3.2: Brown's acoustic environment schema. Dashed boxes indicate sound source categories, whereas sound sources themselves are denoted by italic text and non-dashed boxes (Modified from [68] after [6]).

between pairs of opposing adjectives [71, 72, 73, 74]. Whilst there is no standardised set of scales, some commonly-used adjective pairs are Comfort/Discomfort, Smooth/Rough and Quiet/Loud [72, 75].

In [76], Axelsson *et al.* present work in which five binaural acoustic environment recordings were rated by subjects using 116 different semantic differential pairs. Principle Component Analysis revealed two components that explained 66% of the variation. Figure 3.3 shows the loading plot of these two components. Since the first component is "best explained by the five adjectives Uncomfortable, Comfortable, Appealing, Disagreeable, and Inviting", the authors labelled it *pleasantness*, with the second component labelled *eventfulness* as it is "best explained by Eventful, Lively, Uneventful, Full of life, and Mobile" [76]. A third component, dubbed *familiarity* in this initial study, was shown to account for a further 8% of the variation in responses. This *familiarity* component has since been superseded by the notion of *appropriateness* [71, 74], perhaps missed in this initial study as only recordings from large city environments were used.

Though there has been some concern over the use of linguistic labels given their potential to be interpreted differently by people from different cultural backgrounds [43], a great deal of research is converging on this 2D *pleasantness/eventfulness* perceptual space, sometimes augmented by the additional *appropriateness* dimension [71, 74, 77, 78]. These scales are of particular interest to this research as they provide key insights into soundscape perception. *Pleasantness*, for instance, has been shown to positively correlate with geophonic and biophonic sound, such as running water and birdsong, and negatively correlate with the sounds of machinery, including traffic and construction sound [76]. *Eventfulness*, on the other hand, is shown to be positively correlated to human sounds, yet largely unaffected by technological sounds [73, 74].

This is a key point of differentiation between human soundscape perception and the standard soundecology categories. Axelsson *et al.*'s research indicates that human perception is better reflected by splitting anthrophony into sound directly made

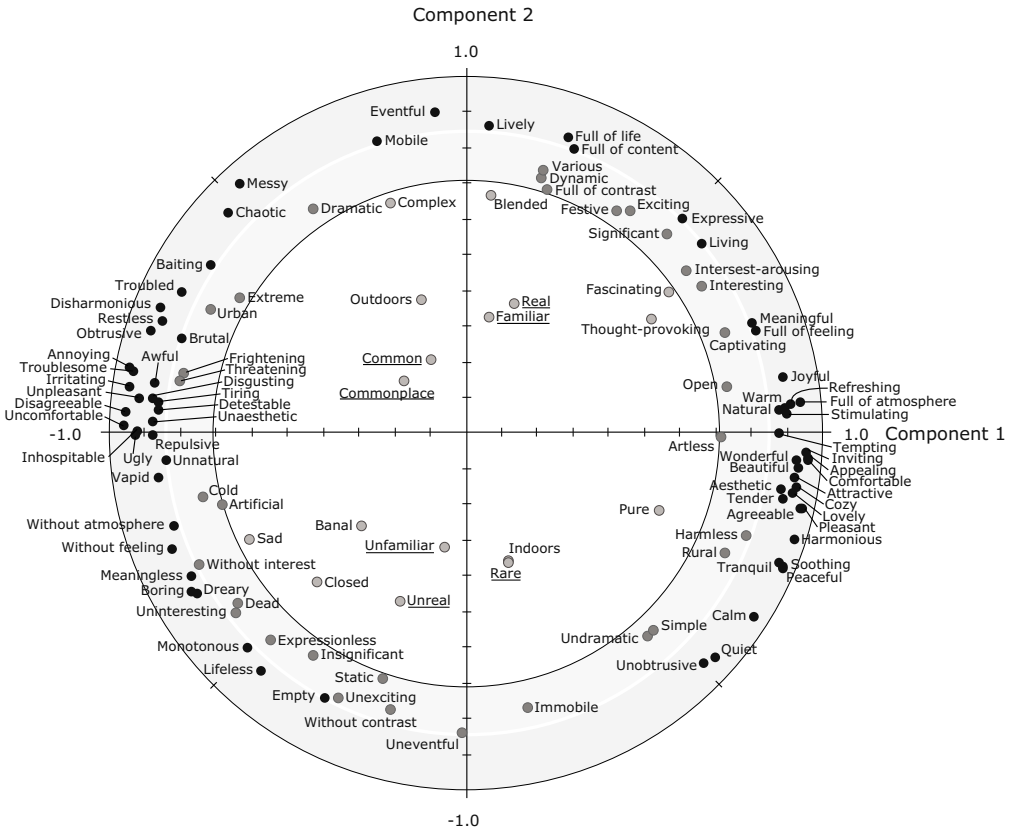


Figure 3.3: Component loading plot of Axelsson *et al.*'s 116 tested semantic differential scales. Component 1 has been labelled *pleasantness*, with Component 2 labelled *eventfulness* (from [76]).

by humans (speech, music, footsteps) and sound made by machines. Likewise, whereas in soundecology it is relevant to distinguish between geophony and biophony, these can, perceptually, be grouped together under the umbrella of natural sounds. This leads to the following perceptually-motivated sound categories:

- **Human**: Non-mechanical sounds indicative of the presence of humans. This primarily consists of speech, but also footsteps, music and laughter.
- **Natural**: The sounds of all manner of fauna except humans, together with sound created by weather and geological forces including rainfall, wind and flowing water.
- **Mechanical**: Sounds from machinery, including transport and construction.

These categories are reflected in the research of Watts *et al.* [79], who found them emerging from questionnaire answers completed by visitors to various green urban locations. They are also used by Stevens *et al.* in their research into soundscape perception and categorisation based on auralisation and visualisation of recorded acoustic environments [45, 80]. Recent work by Kroos et al. [81] used hierarchical cluster analysis to generate a taxonomy based upon participants' sorting of sounds corresponding to the top 60 search terms on freesound.org. This resulted in five top-level labels, including 'human', 'nature' and 'urban', closely corresponding to the three perceptually-motivated categries detailed above, and adding 'music' and 'effects' categories. Since these categories emerged from sorting of mostly isolated sound events, it is unsurprising that musical sounds have been split from human sounds. As discussed previously, context plays a large part in perception, and it is not unreasonable to suggest that musical sounds in an urban acoustic environment setting would be best described as, and elicit similar emotional response to, human sounds. Likewise the 'effects' category consists of isolated events (perhaps indeed recorded as sound effects) with labels such as 'thud' and 'swoosh'. Again, in the context of an acoustic environment, sounds such as these are likely to be perceptually

grouped as belonging to any of the three categories, based on context. A CNN classifier was able to group sounds into these five proposed categories with 80.8% accuracy, a result suggesting the viability of this method for deriving soundscape affect metrics.

Looking back at Figure 3.3 with these categories in mind, it can be seen that *pleasantness* equates to the balance between mechanical sounds and natural sounds, whereas *eventfulness* represents the degree of human sound present. As Lundén *et al.* put it: "Pleasantness is positively associated with the sound of nature and negatively associated with the sound of technology, and ... Eventfulness is positively associated with the sound of people" [78]. This leads to the idea of using these two dimensions as perceptually-motivated metrics alternative to the ecologically-motivated NDSI. A system capable of accurately quantifying the contributions of human, natural and mechanical sounds to an acoustic environment could provide the data required to calculate accurate estimates of the perceptual attributes of a soundscape.

3.4.4 Considering 'Value Judgements'

Despite the neat mapping of objective sound classes to a subjective perceptual space, there remains the question of how the third perceptual dimension - *appropriateness* - interacts with and colours the first two. In Chapter 1 the examples of aeroplane sound and music in a nightclub were used to illustrate how sound level alone does not capture human responses to sound stimuli. In that example, the aeroplane was used as an example of unpleasant sound, whereas the nightclub music was used as a pleasant sound. But what about those neighbours to the nightclub? It's the appropriateness dimension that takes the music from pleasant for the people in the club, to unpleasant for the neighbours. Compare and contrast the adjectives in the upper-right quadrant of Figure 3.3 ('dynamic', 'exciting', 'expressive'), to those in the upper-left ('chaotic', 'disharmonious').

Likewise, mechanical sounds are not universally unpleasant. To elaborate on the aeroplane example, consider enthusiasts attending an air show. In this case, the loud

aircraft sound could be a large part of the attraction, especially in cases such as the famous 'howl' of the Avro Vulcan bomber [82]. The upper-right quadrant adjectives seem as apt to describe the experience here as to nightclub revellers. It is clear that whilst *appropriateness* might be a lesser component, its ability to influence the main two dimensions is considerable.

It is therefore important to avoid the trap of much of the literature in describing certain sources as 'intrusive' and others as 'pleasant' seemingly by default. Bunting *et al* [70], for instance, make mention of "intrusive sources of noise, e.g. mechanical, or non-intrusive, e.g. birdsong..." without any qualification. In contrast, Brown's schema (Figure 3.2) was expressly developed in such a way as to "avoid inputting value judgements" of sound sources in any given context. Harriet posits that such value judgements made on natural and urban soundscapes could be due to the effects of enculturation. The idea is presented that a "collective shift in attitudes" towards soundscapes might be something to be encouraged, citing the human tendency to change tastes over time in regards to art and fashion [43].

There is an interesting dichotomy here in that taking perception into account when producing acoustic environment descriptions could result in a reduction to the black-and-white 'noise' approach, especially if all mechanical sound is considered 'bad' by default. Alternatively, the total removal of the pleasantness/eventfulness subjective descriptors runs the risk of losing sight of the perceptual aspects of this field, when these are the main motivation for development of the soundscape approach. There is clearly a balance to be found between considering the *appropriateness* of an acoustic environment, which is dependent on culture and context, and much less easily quantifiable as it is not directly equatable with purely acoustic content of the sound sources, and the *pleasantness* and *eventfulness* of that environment, which are much simpler to predict in terms of acoustic content for most people most of the time.

3.5 Subjective Soundscape Assessment

The biggest challenges surrounding the subjective assessment of soundscapes have been how to collect useful data and how to quantify that data in a standardised fashion. Broadly speaking, most of the work on subjective soundscape assessment is conducted using one of two techniques:

- **Soundwalks** - Subjects experience an acoustic environment on location.
- **Laboratory Reproduction** - The acoustic environment is reproduced in a controlled listening environment.

3.5.1 Soundwalks

In a typical soundwalk, a group of participants are led by a researcher along a prescribed route through an area of interest. Subjects are usually encouraged to concentrate on their perception of the acoustic environment. At designated locations along the walk, participants are asked to stop and answer questions, either in the form of a questionnaire or an interview [71, 72]. Soundwalks can last as long as ninety minutes. Figure 3.4 shows a typical soundwalk route through a "green space" park area, with the numbered points indicating the locations where participants were asked questions.

Soundwalks have the advantage of presenting subjects with the most realistic stimulus possible, direct from the environment itself. This means that results gained from this technique will be as representative as can be achieved of their reactions to the real-world acoustic environment, a factor known as 'ecological validity' [75]. A key disadvantage of this technique, however, is the lack of experimental control. Whilst the researcher can set the route of the walk, the actual sounds presented cannot be controlled and so the test is not reproducible. This has led to criticism that soundwalks "only represent the case in question, and will not contribute to general knowledge" [71].

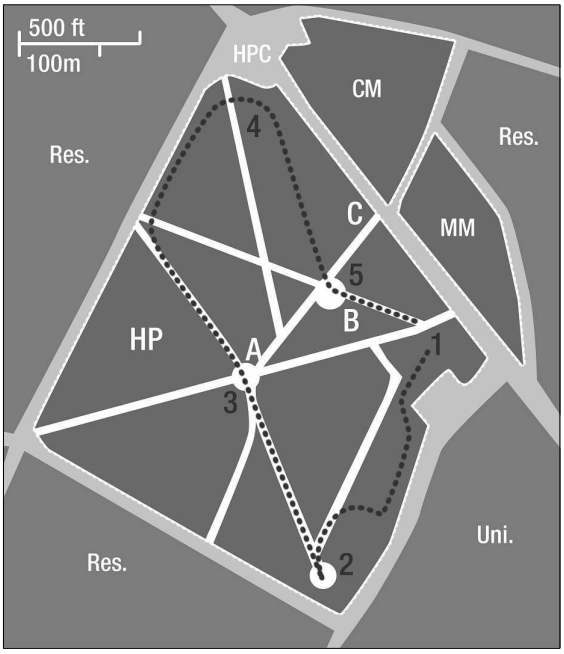


Figure 3.4: Example soundwalk route conducted in Woodhouse Moor, Leeds by Harriet & Murphy. Stopping locations are numbered 1 to 5 (From [72]).

3.5.2 Laboratory Reproduction

In laboratory-based soundscape assessments, subjects are presented with recordings of acoustic environments, which might also be presented binaurally [76], or using Ambisonics [72, 74, 78]. The recordings presented are usually of much shorter duration than a typical soundwalk, making it more convenient to collect extensive results.

Lab experiments have the advantage of allowing more experimental control. The researcher can present subjects with consistent sound recordings, so experiments are reproducible. It is also possible to test the potential perceptual impact of soundscape interventions before implementing any physical changes to the infrastructure on-location. The work presented in [72], for instance, tests the impact of the introduction of a noise barrier via simulation and laboratory playback. The clear disadvantage of this approach is the potential for reduced ecological validity, which leads to the criticism that lab results "ought to be validated in situ" [71]. Such a fun-

damental limitation to the technique would seem to defeat the object of lab-based testing, so there has been some work recently on establishing the ecological validity of these tests.

Boren *et al.* note that, "Providing quick A/B comparisons between auditory scenes make differences in loudness, texture, or clarity stand out to listeners more distinctly than if they visited each site in person" [83]. It is interesting to consider this assertion from the point of view of ecological validity. Whilst representing an advantage in some contexts, it could be inherently ecologically invalid to be able to switch rapidly between acoustic environments in a manner that could not happen in the natural world. Similarly, lab-based reproduction allows the listener to focus on sounds, absent from the reality of having to deal with a chaotic city environment, or exposure to the sights, smells and other sensations that make up a real-world experience. Schafer contended that soundscapes cannot and should not be considered separately from their environments [3]. Perhaps, however, the question in terms of practical assessment is how realistic lab-based reproduction needs to be to produce results representative of those that would be obtainable in a real environment, rather than the more philosophical concern of what is considered to be a real experience.

Davies *et al.* have presented work that gives some evidence that Ambisonic reproduction may give ecologically valid results [75]. In their test, a set of thirty-second long clips were recorded of urban city locations and presented to subjects over an Ambisonic loudspeaker array. Results showed some correlation with in-situ soundwalk experiments conducted earlier by Kang [84]. These recordings were made seven years after Kang's initial study and in a different city, however, so this can hardly be considered a fair comparison.

Harriet and Murphy give some more robust evidence for the ecological validity of Ambisonic rendering [72]. In this research, results from the soundwalk outlined in Figure 3.4 are compared with results from a 'Virtual Sound Walk' using Ambisonic recordings made in locations 3 and 5, with a third recording made next to the

road. Although these recordings were not made during the real sound walk, they were made under "similar" conditions and represent a more valid comparison than the work of Davies *et al.* Results from this test show strong correlation between responses to the Ambisonic reproduction and location 2 on the real soundwalk. This provides evidence that laboratory testing can be ecologically valid, though it is unclear why there is such a strong correlation to location 2, given that the recordings were actually made in other locations.

One potential explanation might be found in the preparation of the Virtual Sound Walk test clip. A concern is raised in [72] regarding how to "compress the soundscape in time, yet without compromising its authenticity". This is a key issue with regards to ecological validity. In the context of Harriet and Murphy's paper, a one-hour soundwalk experience had to be condensed into a seven-minute clip. Harriet and Murphy essentially condense the soundscape 'by ear', carefully annotating recordings with sound events and reconstructing a new acoustic environment as a "composition" that "tries to give a balanced impression" by including alternating sections of busy and calm [72]. By contrast, in Davies *et al.*'s paper there is little consideration of this factor, with no detail given on how the thirty-second clips were chosen [75].

Whilst Harriet and Murphy's method is probably more robust than presentation of a random short clip, theirs is not a foolproof process. The recomposition of the acoustic environment by the researcher introduces a source of bias that could be reflected in the results. Results from tests described in [75, 85], in which subjects are asked to design an Ambisonic reproduction of an acoustic environment using prerecorded sound samples, show that participants tend to design the soundscape based more on their expectation of what it *should* sound like than any objective reality. This could have been the case in Harriet and Murphy's soundscape composition.

The ability to effectively sidestep this issue is one of the key application areas of an automatic acoustic environment classification system. The information gathered by such a system could be used to assist with the synthesis of shorter acoustic environment clips that remained statistically representative of the longer experiences

of real acoustic environments, thus helping to increase the ecological validity of lab-based soundscape work.

3.5.3 Mixed Reality

Modern virtual and augmented reality (VR/AR) technologies, capable of creating immersive visual and auditory worlds or adding virtual objects to real environments through head-mounted displays and smartphones, are increasingly being used as tools in soundscape research.

Motivated by research indicating the influence of visual factors on soundscape perception [86, 87], Ruotolo *et al.* [88] used VR to assess the response of participants to a rural location in southern Italy. A virtual model of a real area where motorway construction was planned was created and paired with binaural recordings made at locations near existing motorways. Participants were tested on versions of the model with and without the proposed motorway and asked to complete a series of cognitive tasks whilst immersed in the simulation. It was found that performance in a task involving verbal memory decreased with increasing proximity to the motorway. VR was used in this case in order to "reproduce the perceptual richness of the environment whilst keeping experimental control", a claim that clearly links the use of VR with ecological validity, termed "behavioural realism" in this study [88].

The same research team later explicitly tested the ecological validity of VR reproductions in tests at Via Partenope in Naples and a virtual reconstruction of the location [89, 90]. A six-part questionnaire was used to compare the experience of passersby at the real city location with participants experiencing the virtual version, with results showing "congruence" of experience between the two conditions. This gives some evidence for the ecological validity of the method, though it should be noted that the questionnaire was far less extensive than many used in soundwalk studies.

One of the few examples of AR use in soundscape research is a system created by Kınayoğlu to test the perceptions of subjects to altered acoustic environments at

four locations around the University of California Berkely campus [91]. Participants carried a bag containing a laptop computer with a program which rendered spatial binaural acoustic environments composed by researchers for each area based on geolocation and head rotation. This had the effect of completely replacing the natural acoustic environment. The study is especially interesting as a targeted enquiry into how the appropriateness of the acoustic scene to its location affects perception, as the alternative acoustic environments presented at each location varied in their congruence with the real physical locations. Participants completed a survey including ratings for categories including *pleasantness*, *eventfulness*, and *appropriateness* analogues. Replacing the sounds at a public square with sound from a similar location in Morocco did not result in a reduction in *appropriateness* and actually resulted in increased *pleasantness* and *eventfulness* ratings. This suggests that subtle cultural differences in sound sources (e.g. the most prevalent spoken language) may not result in a reduction of the feeling of *appropriateness* as long as the holistic effect of the soundscape is broadly consistent with the location. When busy urban sounds were presented at a park location, *appropriateness* ratings did decrease. Some listeners reported the sense that things must be happening "beyond the trees" [91], an effect that the authors compare to off-screen diegetic film sound. This suggests that the brain will attempt to construct a plausible model of the environment even when visual and sonic information is mismatched.

Since Kınayoğlu's study, Apple's ARKit [92] has emerged as a cutting edge AR technology. Running on an iPhone, it is capable of tracking features in the device's surroundings to enable a smooth AR experience, and so far has been applied to interior design and measurement [93, 94] as well as gaming. More recently, devices specifically developed for true AR audio applications have been introduced to the market. The Sennheiser AMBEO Smart Headset [95] is an iOS accessory that combines high-quality binaural microphones and earbuds with a smartphone interface to enable mobile augmented audio in conjunction with ARKit. The 'transparent hearing' mode relays incoming audio to the in-ear speakers with very low latency.

This allows virtual audio sources to be convincingly superimposed onto real acoustic environments and experienced on-location in real time. The Apple AirPods Pro [96] have brought this capability to a more mainstream consumer product, though use of this technology in academic soundscape research remains limited so far and is one of the areas of novelty in this thesis, discussed further in Chapter 8.

It is possible, if AR technology becomes as widespread as smartphones have become over the past decade, that the nature of what is considered 'reality' could change. This has profound implications for, amongst many other things, the concept of ecological validity. What is ecologically valid for people who spend most of their time in an augmented world where the line between what is virtual and what is real becomes harder to define? Perhaps AR technology will offer an opportunity to consider what kind of a sound world we want to live in, and real soundscapes can be designed with potential augmentations in mind in a way that can be healthy and fulfilling for the most possible people.

3.6 Summary

This chapter has explored the concepts of soundscape and environmental noise, and how each of these approaches view sounds in the environment as they relate to human perception. The simplicity of the environmental noise approach has been established as a shortcoming as well as a strength, in contrast to the more nuanced, yet also potentially more complicated, soundscape approach. Particular attention was paid to the various systems of sound categorisation presented in the literature, their utility across different scenarios, and how they might be integrated into an alternative soundscape metric. This was followed by an exploration of some methods by which the soundscape approach has been implemented in listening tests, the potential shortcomings thereof, and the ways in which machine listening systems might be able to mitigate these. The next chapter will therefore present an exploration of machine listening and the various paradigms that have emerged in the field, together

with detailed descriptions of the various algorithms that have been utilised as part of the work presented in this thesis.

4 | Machine Learning for Acoustic Environment Analysis

4.1 Introduction

Research into machine learning for audio applications, or 'machine listening', from the early 1990s up to the mid 2000s mainly focussed on automatic speech recognition (ASR) [8] and music information retrieval (MIR) [97]. These are now mature fields, with robust speech recognition featured in all modern smartphones and smart speakers, and MIR technologies integral to the intelligent playlist algorithms used in many music streaming services [98]. When considering broader sound scenes, Computational Auditory Scene Analysis (CASA) seeks to devise computational solutions to the 'cocktail party problem' - the "ability to listen to and follow one speaker in the presence of others" [99]. Since the focus is the intelligibility of the desired speech signal, other sounds are treated as 'background', to be suppressed and discarded from the analysis. This approach is clearly at odds with the holistic acoustic environment analysis system proposed for this research, in which *all* the sounds that make up the scene are of interest.

ASR research historically approached the problem within the paradigm of emulation of the human hearing system and perception of sound. This included emulating the cochlear response to sound and 'auditory Gestalt' perceptual grouping principles including proximity in frequency and time, and coincidence of onsets and offsets

[8, 100]. One self-imposed limitation of this approach is the restriction of the input to monaural or binaural audio in order to make the work more "biologically relevant" [8]. This ethos is reflected in Wang and Brown's definition of CASA: "the field of computational study that aims to achieve human performance in ASA by using one or two microphone recordings of the acoustic scene" [8].

More recently, the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge and workshop has been established as a forum for machine listening research considering "algorithms that can describe, catalogue and interpret all manner of sounds" [10]. In their paper summarising the results of the first DCASE challenge, Stowell *et al.* state that the "human-centric aims [of CASA] do not directly reflect our goal...which is to develop systems that can extract semantic information about the environment around them from audio data" [10]. This reflects the shift in focus of this work in that the acoustic environment itself is the primary object of attention, rather than speech in particular or any perceptual processes. It might be useful to think of CASA as opposed to CASSE in similar terms as the divide between the soundscape and acoustic environment terminologies discussed in section 3.2. This is not to say that systems concerned with CASSE cannot use perceptually-inspired processing, more to note that "the two research fields do not completely overlap" [54]. Though the most recent DCASE challenges have branched out to more niche application-focussed subtasks including monitoring of domestic activities and detection of birdsong [101], the core aims of CASSE remain acoustic scene classification (ASC) and sound event detection (SED).

CASSE research could be considered in its aims closer to MIR. The DCASE challenge format was inspired by the MIREX (music information retrieval evaluation exchange) challenge [102], which has been running annually since 2005. Some of the MIREX subtasks can be seen as close analogues to DCASE subtasks, 'audio genre classification', for instance being similar in concept to acoustic scene classification, with note tracking roughly similar to event detection, in that, like an event detector, a note tracking system produces onset and offset times along with a label, in this

case indicating pitch and/or instrument source. A key difference arises in the fact that MIR work is often able to make use of extensive metadata, including MIDI sequences, musical score and, in somewhat of a crossover with ASR, lyrics [103]. Although it is possible to use metadata in CASSE research, this will seldom be as specific to the audio stream as in MIR.

4.2 A General Sound Classification Framework

The vast majority of the work done to date in ASC has approached the problem in machine-learning terms as a single-label classification task in which a system has to choose one label from a provided set with which to classify a given input [10, 54, 104]. There have been a plethora of proposals for systems to achieve this for acoustic scene and event recordings, with many variations, but the vast majority use the same general set of steps. A framework formalising these steps has been developed by Barchiesi *et al.* [54], and can be summarised as follows:

A set of audio clips $\boldsymbol{s}_m$ that has been assigned descriptive labels c_m taken from a pool of all possible category labels γ_q is used to train some form of machine learning system, which is subsequently tested on new unlabelled clips $\boldsymbol{s}_{\text{test}}$. In order to provide usable input for machine learning algorithms, input signals are split into frames usually of the order of 20 - 60 ms in length and with consecutive frames typically overlapping between 25% and 75% [105]. The frames then undergo some process $\mathcal{T}$ which produces a sequence of features $\boldsymbol{x}_{n,m}$ (frame n of input m), i.e. $\mathcal{T}(\boldsymbol{s}_{n,m}) = \boldsymbol{x}_{n,m}$. $\mathcal{T}$ can be anything from a simple count-based process to a complex set of transforms, but the key is the processing of raw time-domain frames to obtain a more coarse representation of the sound. This coarsening of the data is an essential step in any machine learning process - the features need to be detailed enough that sounds of different types can be separated and identified (discrimination), yet coarse enough that similar sounds will give similar features (generalisation). Historically, this has been a manual process, but the prevailing ethos behind the current state-of-

the-art convolutional neural network (CNN) models is that the network itself should perform the coarsening of the data, learning appropriate simplifications over time [106] (see Section 4.3.6).

Since acoustic environments are characterised by how they change over time, features extracted over many frames should be analysed together. For this reason, the series of features extracted from a certain class of clip are used to build up statistical models of that class. A machine learning algorithm $\mathcal{S}$ learns a statistical model $\mathcal{M}$ of the training data, i.e. $\mathcal{S}(\boldsymbol{x}_{n,\Lambda q}) = \mathcal{M}$, where Λ_q is the set of clips of class q. After training, new input audio clips $\boldsymbol{s}_{\text{test}}$ are used to test the performance of the models. The same process $\mathcal{T}$ used in training is applied to the new signals to gain the features $\boldsymbol{x}_{\text{test}}$. The features are tested against the statistical model $\mathcal{M}$ using a function $\mathcal{G}$ to get a class label for the new clip c_{test}, i.e. $\mathcal{G}(\boldsymbol{x}_{\text{test}}, \mathcal{M}) = c_{\text{test}} \in \gamma_q$.

4.2.1 Sound Event Detection

Barchiesi *et al.*'s framework was originally written only with regards to ASC, but classifying sound events is a very similar problem computationally, albeit over a different timescale. The aim of a SED system is, however, not only to output labels for isolated audio clips, but to produce a series of labels annotating an input audio recording containing a sequence of audio events. Each annotation gives the start and end points for a given event, along with a name for the event in question [10]. Such systems therefore combine a classification task with a detection task. The original DCASE baseline for this task performed detection by frame-wise classification, then using a threshold on the class likelihoods to create a binary activation sequence for each class [10]. Later systems have used CRNN models to generate this binary activation matrix directly [107]. This is a polyphonic detection paradigm, whereby the output can include annotation of multiple simultaneous sound sources in one time-frame. Many earlier systems provide monophonic output - annotation of a single event per time-frame using sequencing techniques derived from speech recognition [108, 109, 110]. The two approaches are shown in Figure 4.1.

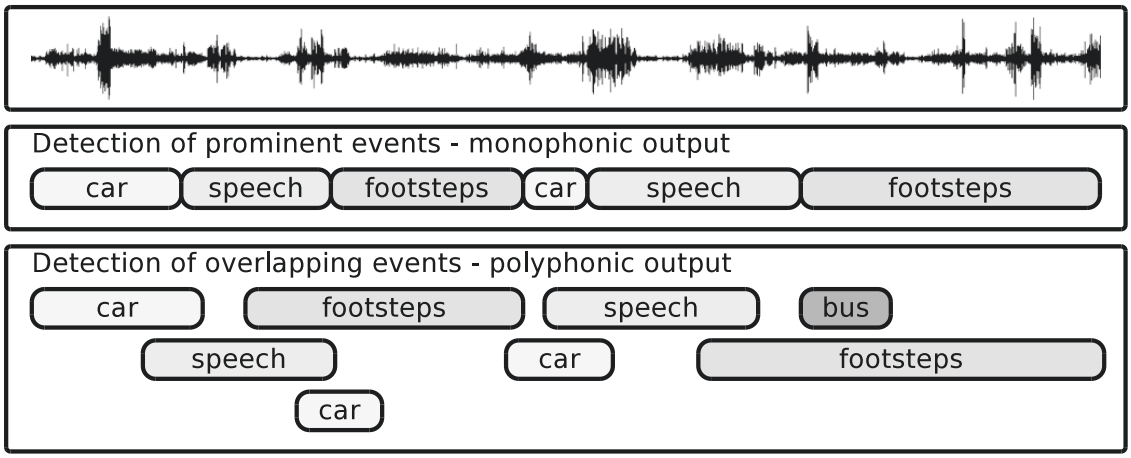


Figure 4.1: The two paradigms for acoustic event detection - monophonic and polyphonic output (from [109]).

The performance of event detection systems is typically lower than that of classification systems. This is likely due to the need to accurately discriminate sounds that may overlap or be partially masked, whilst labelling in such systems can be affected by erroneous estimates for event boundaries. Detection-by-classification systems can also be limited by the fact that the detection segments are unlikely to be matched to the length of audio events, making the problem akin to classification of incomplete samples, further increasing the difficulty [108].

4.3 Features and Classifiers

4.3.1 Mel-spectrogram and MFCCs

A very common process used for $\mathcal{T}$, is calculation of the mel-spectrogram. Mel-Frequency Cepstral Coefficients (MFCCs) are derived by additional processing of the mel-spectrogram. These are also common features, especially in ASR [111]. The mel-spectrogram and MFCCs use the mel scale, which was originally developed to relate physical frequency to the perception of pitch, and is defined as [112]:

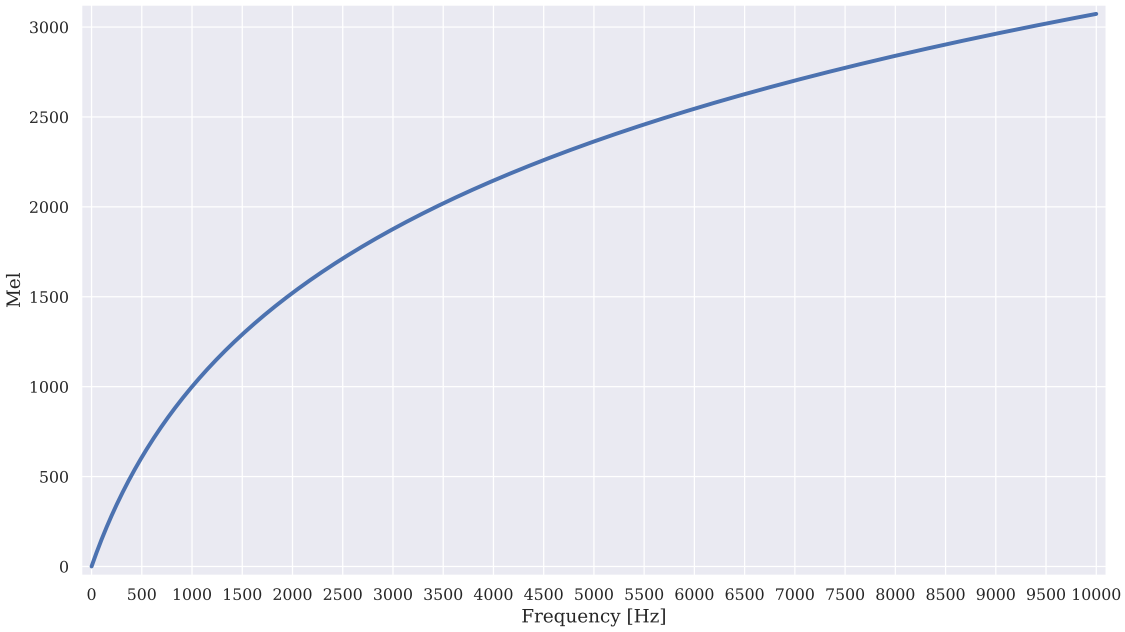


Figure 4.2: The mel frequency scale.

$$\text{mel}(f) = 2595 \log_{10}(1 + f/700) \tag{4.1}$$

The resultant curve, shown in Figure 4.2, is an exponential decay of increasing form, which approximates linear mapping below 1000 Hz, but levels off logarithmically beyond that point [113].

A mel-filter bank is a series of overlapping triangular filters $H_m(k)$ with frequency spacing determined by the mel scale. A bank of M filters is designed as follows [111]:

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ {}^{k-f(m-1)}/_{f(m)-f(m-1)} & f(m-1) \leq k \leq f(m) \\ {}^{f(m+1)-k}/_{f(m+1)-f(m)} & f(m) \leq k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{4.2}$$

where $f()$ is a list of $M+2$ mel-spaced onset and offset frequencies. Figure 4.3 shows the overlapping frequency responses in a mel-filter bank. For clarity the figure shows a filter bank for $M = 10$, but $M = 128$ is a much more common value in practice

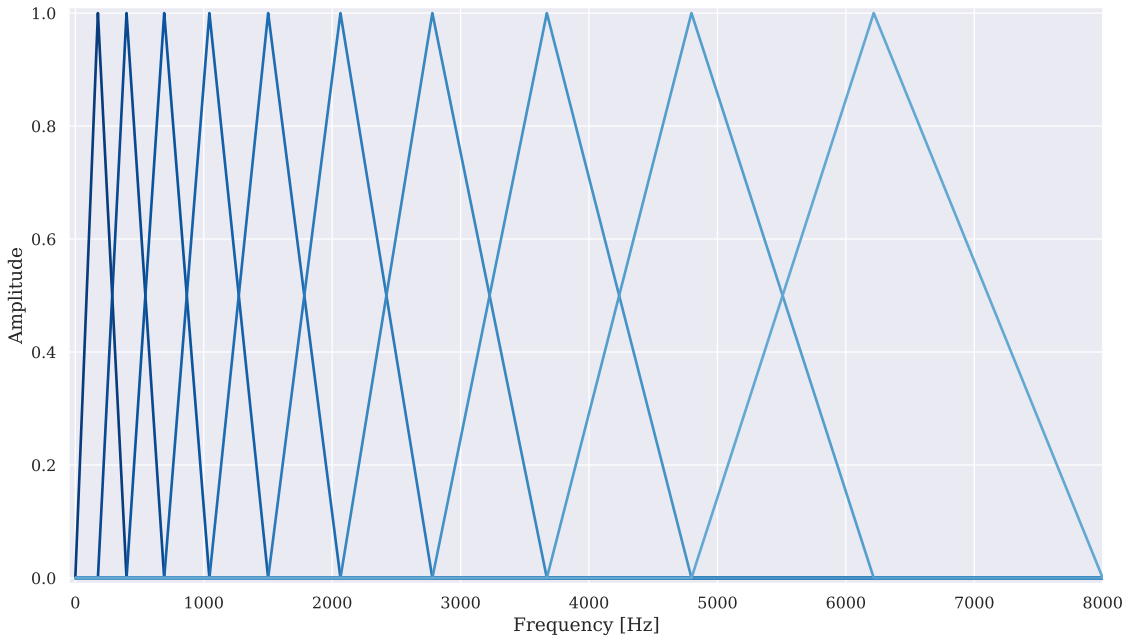


Figure 4.3: Mel-filter bank for $M = 10$ and $f(M + 1) = 8000$ Hz.

for mel-spectrograms and $M = 20$ for MFCCs.

To calculate the mel-spectrogram, the complex frequency spectrum $x(k)$ of each frame of audio is first obtained using a standard short-term Fourier transform (STFT) and the periodogram $P(k)$ obtained as:

$$P(k) = \frac{|x(k)|^2}{N} \tag{4.3}$$

where N is the length of the audio frame in samples. The mel-filter bank is applied to this, the energies in each frequency band are summed, and the logarithm taken:

$$E_m = \log\left(\sum_k H_m(k) \cdot P(k)\right) \tag{4.4}$$

This results in a vector containing logarithmic filter bank energies for a given frame of audio. Across a whole clip these energies make up the mel-spectrogram, and this is commonly used as input to audio recognition systems using CNNs (see Section 4.3.6). Mel-spectrograms make particularly useful input to CNNs as they contain

descriptive information about the whole sound signal, with the mel-scaling coarsening the frequency resolution whilst emphasising those components that are most important to human perception of the sound. Without this coarsening, the CNN would have to be more complex, resulting in increases in required computing power. Altering the number of mel filters also offers a simple way to change the frequency resolution for different applications.

To calculate cepstral coefficients MFCC_l, a discrete cosine transform (DCT) [114] is applied to each frame in the mel-spectrogram:

$$\mathrm{MFCC}_l = \sum_{m=0}^{M-1} E_m \cos\left(\frac{\pi l}{M}\left(m + 1/2\right)\right), \; l = 0, \ldots, M-1 \tag{4.5}$$

MFCCs are a coarse representation of the frequency spectrum that is somewhat analogous to the response of the cochlea, reflecting the origins of this technique in speech recognition. MFCCs offer the advantage over mel-spectrograms of being a more compact representation of the sound signal, saving on computing power at the expense of detail.

4.3.2 Low-Level Features

As should be clear, the mel-spectrogram and MFCCs are high-level features requiring complex processes to extract. Lower-level features include the zero-crossing rate in the time domain [115], calculation of RMS values per frame of audio, and various statistical descriptions of spectral properties [116].

Garrido *et al.* [117] presented a system that uses simple L_{eq} readings across multiple octave bands in order to classify between traffic and leisure sound specifically. The technique was shown to work well in distinguishing these two categories. This system could be viewed as a somewhat expanded version of the approach taken in calculation of the NDSI, described in Section 3.4.2, with ratios between frequency bands not calculated directly, but perhaps inferred by the Support Vector Machine (SVM) model (see Section 4.3.4) used in this work. The lack of consideration for

temporal information in this system also makes it similar to the bag-of-frames approach, which will be described in Section 4.4.1.

4.3.3 Gaussian Mixture Models

There are numerous models suitable for classifying sounds. One statistical model prevalent in many ASR systems and early ASC work is the Gaussian Mixture Model (GMM). A Gaussian distribution $\mathcal{N}$ models the probable density of d dimensional data $\boldsymbol{x}$ with a single peak at mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ [118, 119]:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp\left(-1/2(\boldsymbol{x}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \tag{4.6}$$

A GMM is simply a mixture of K Gaussians, which can be used to model more complex distributions of arbitrary shape with multiple peaks:

$$p(\boldsymbol{x}) = \sum_{n=1}^{K} w_n\,\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \tag{4.7}$$

w_n are the weights for each Gaussian component. GMMs are initialised with their means $\boldsymbol{\mu}$ set at random values from the training data [120]. Figure 4.4(a) shows an example of a GMM for $K = 4$ and $d = 2$ at this stage. In optimising a GMM, the expectation-maximisation (EM) algorithm is used to iteratively alter the parameters of the model to maximise the likelihood that a sample of data generated at random from the mixture of Gaussians could have originated from the data presented to the model [120]. In the *expectation* stage, the posterior probabilities $P(n|\boldsymbol{x}_i)$ of each datapoint $\boldsymbol{x}_i$ belonging to a particular Gaussian component n, also known as the responsibilities r are calculated [118, 121, 122]:

$$r_n^i \leftarrow P(n|\boldsymbol{x}_i) = \frac{\pi_n\,\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)}{\sum_{n'}\pi_{n'}\,\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_{n'}, \boldsymbol{\Sigma}_{n'})} \tag{4.8}$$

The *maximisation* stage consists of updating the model parameters by estimating the maximum likelihood for each datapoint, weighted by the previously calculated responsibilities, as follows:

$$\boldsymbol{\pi}_n \leftarrow \frac{\sum_{i=1}^{I} r_n^i}{I} \tag{4.9}$$

$$\boldsymbol{\mu}_n \leftarrow \frac{\sum_{i=1}^{I} r_n^i \boldsymbol{x}_i}{\sum_{i=1}^{I} r_n^i} \tag{4.10}$$

$$\boldsymbol{\Sigma}_n \leftarrow \frac{\sum_{i=1}^{I} r_n^i (\boldsymbol{x}_i - \boldsymbol{\mu}_n)^2}{\sum_{i=1}^{I} r_n^i} \tag{4.11}$$

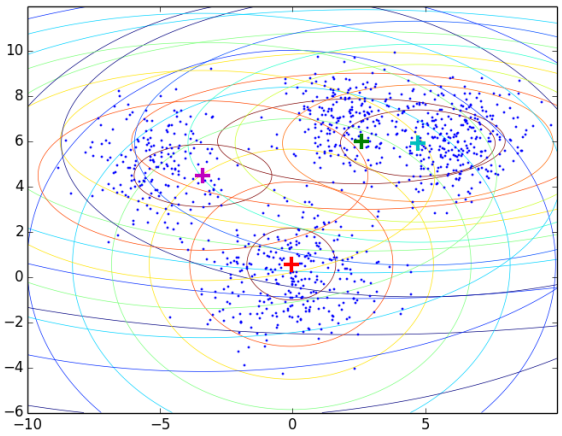
where I is the total number of data points. These two stages are repeated until the parameters converge on optimal values. Figure 4.4(b) shows the example GMM after 100 iterations of the EM algorithm.

In a multi-label classification scenario, there are two methods for using GMMs to classify data. If the distributions of the input features from each class are known in advance to be Gaussian, a single GMM can be used to generate probabilities that incoming test data came from each Gaussian within that model, with the data classified based on the Gaussian giving the highest probability score [120]. In this case, n is set to equal the number of classes expected in the data.

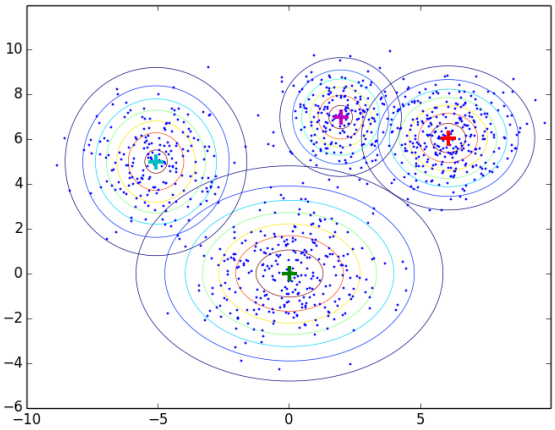
The second method, more common in audio classification work, is to optimise a complete GMM for each class in the training data, so each class is modelled with multiple Gaussians independently. The set of GMMs are used to produce probability scores indicating the likelihood that the input testing data came from each mixture. This is the method used by Aucouturier *et al.* in their bag-of-frames approach (see Section 4.4.1), and the baseline systems used in early DCASE challenges [124, 125]. In this case, n is more difficult to set and often trial-and-error is used. Stowell's smacpy system uses $n = 10$, where the DCASE 2016 baseline system, created as a benchmark by which to test challenge submissions, uses $n = 32$ [125].

4.3.4 Support Vector Machines

Support vector machines (SVMs) are a family of algorithms that can be used for classification or regression. Support vector classifiers (SVCs) work by calculating



(a) Initial state. Four Gaussian components have been initialised with randomised parameters.



(b) The parameters of the four Gaussian components have, over 100 iterations of the E-M algorithm, been optimised to fit areas of density in the data.

Figure 4.4: A four-component GMM before and after optimisation (from [123]).

a hyperplane that separates feature vectors obtained from different classes in the training data [54, 120], whereas support vector regressors (SVRs) calculate a plane to model the function generating a dataset [126].

For an SVC, the dividing hyperplane is defined as the set of vectors $\boldsymbol{x}$ satisfying

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = 0 \tag{4.12}$$

where $\boldsymbol{w}$ is the normal of the hyperplane and b is the bias, which defines the offset of the plane from the origin of the space along $\boldsymbol{w}$. Predictions y_i^{pred} are made by taking the dot product of the feature vectors against the weights as follows:

$$y_i^{\text{pred}} = \begin{cases} 1 & \boldsymbol{w} \cdot \boldsymbol{x}_i + b \geq M \\ -1 & \boldsymbol{w} \cdot \boldsymbol{x}_i + b \leq -M \end{cases} \tag{4.13}$$

where M is the distance margin between the hyperplane and the nearest vectors from each class, which satisfy:

$$\boldsymbol{w} \cdot \boldsymbol{x}_{\text{support}} + b = |M| \tag{4.14}$$

These are the 'support vectors'. The hyperplane maximising M is known as the "maximum-margin hyperplane" and should give the best generalisation to unseen data that is possible given the available training data. Assuming equal distance between the dividing plane and the support vectors from each class, it can be shown that [127]:

$$M = \frac{2}{\|\boldsymbol{w}\|} \tag{4.15}$$

The problem of maximising M therefore becomes a case of minimising $\|\boldsymbol{w}\|$, whilst still separating the classes. In the likely event that the classes are not linearly separable, a plane can still be calculated taking into account the influence of misclassified points, set by a parameter C. The problem of optimising the plane is formally defined as [128]:

$$\begin{aligned} &\text{Minimise } \|\boldsymbol{w}\| + C\sum_{i=1}^{n}\xi_i \\ &\text{subject to } \begin{cases} y_i^{\text{true}} \cdot y_i^{\text{pred}} = y_i^{\text{true}}(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \end{aligned} \tag{4.16}$$

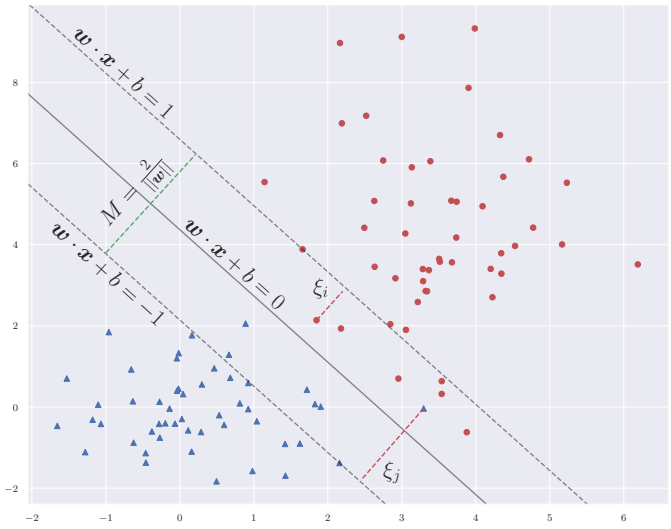
where y_i^{true} are the ground-truth class values (± 1) and ξ_i is the distance of misclassified point i from the boundary of the margin. A large value of C highly penalises errors, and will therefore favour a smaller margin and fewer errors. Conversely, small values of C will allow a larger amount of errors as a trade-off for a larger margin. Figure 4.5 shows two SVCs computed for some dummy data consisting of two randomly-generated clusters. The solid line shows the decision boundary, with the margins shown as dotted lines. The small value of C in Figure 4.5(a) has resulted in a large margin but with a considerable amount of errors, two examples of which are illustrated as ξ_i and ξ_j. The SVC in Figure 4.5(b) uses a larger value for C, and the margin is much narrower as a result. The optimal value of C for a given problem ultimately depends on the unseen test data. By reformulating the margin optimisation problem as the Lagrangian dual it is solvable using standard quadratic problem-solving algorithms [120].

Each SVM can only calculate one hyperplane, so multiple SVCs are needed when discrimination between more than two categories is needed. This is achieved using either a 'one-versus-all' approach, in which each SVC is trained to discriminate between one class and all other classes, or a 'one-versus-one' approach, in which SVCs are needed for every possible combination of two classes.

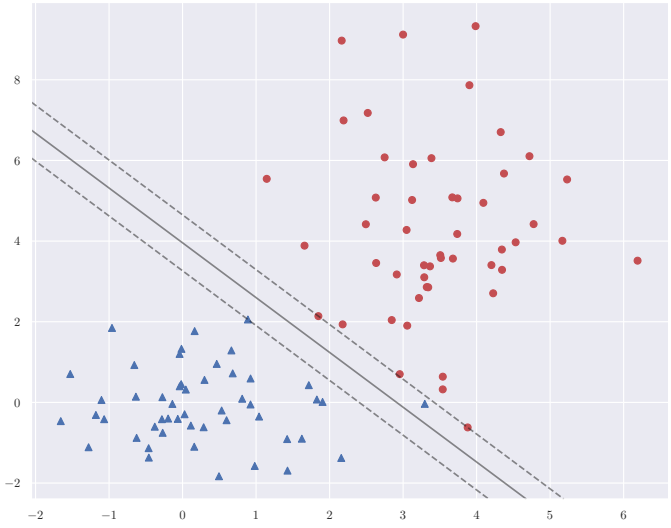
SVRs work very similarly to SVCs, but with a continuous output rather than a binary classification. Equation 4.13 therefore becomes [126]:

$$y_i^{\text{pred}} = \boldsymbol{w}_i \cdot \boldsymbol{x}_i + b \tag{4.17}$$

and the optimisation problem defined in Equation 4.16 is modified for continuous



(a) C = 0.1. This small value has resulted in a larger margin, with more points classified in error.



(b) C = 10. The larger value has resulted in a much narrower margin, which has reduced errors, but may not be as effective for unseen data.

Figure 4.5: Example linear SVCs for different values of C.

output accordingly [126, 129, 130]:

$$\begin{aligned} &\text{Minimise } \|\boldsymbol{w}\| + C\sum_{i=1}^{n}(\xi_i^- + \xi_i^+) \\ &\text{subject to } \begin{cases} y_i^{\text{true}} - \boldsymbol{w}_i \cdot \boldsymbol{x}_i - b \leq \epsilon + \xi_i^- \\ \boldsymbol{w}_i \cdot \boldsymbol{x}_i + b - y_i^{\text{true}} \leq \epsilon + \xi_i^+ \\ \xi_i^-, \xi_i^+ \geq 0 \end{cases} \end{aligned} \tag{4.18}$$

where ξ_i^- and ξ_i^+ indicate the distance of points below the lower margin boundary and above the upper margin boundary respectively, with each being zero on the other side of those planes. ϵ is a tolerance parameter indicating the allowable margin width within which errors are not counted, ultimately affecting the number of support vectors used to define the plane. Lower values of ϵ lead to a higher number of support vectors, and very low values can lead to overfitting. The value of C has a similar effect here as in the SVC, but weighting only the influence of deviations greater than ϵ ($\xi_i^{\pm} > 0$).

For both SVCs and SVRs, a better fit for data where classes are not linearly separable can often be found by mapping the data into higher dimensions using kernel functions k:

$$\sum_{i=1}^{n} \lambda_i y_i^{\text{true}} k(\boldsymbol{x}_i, \boldsymbol{x}_{\text{new}}) + b \tag{4.19}$$

The work in this thesis uses the radial kernel:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_{new}) = e^{-\gamma\,(\boldsymbol{x}_i - \boldsymbol{x}_{\text{new}})^2} \tag{4.20}$$

The radial kernel weights the output for $\boldsymbol{x}_{\text{new}}$ based on the influence of the training examples $\boldsymbol{x}_i$, with γ a scaling factor determining the distance over which the influence of each training examples operate. High values of γ result in a small radius of influence and vice versa [131]. This kernel effectively maps the data into an infinite number of dimensions, possible in linear time by use of the 'kernel trick', in which

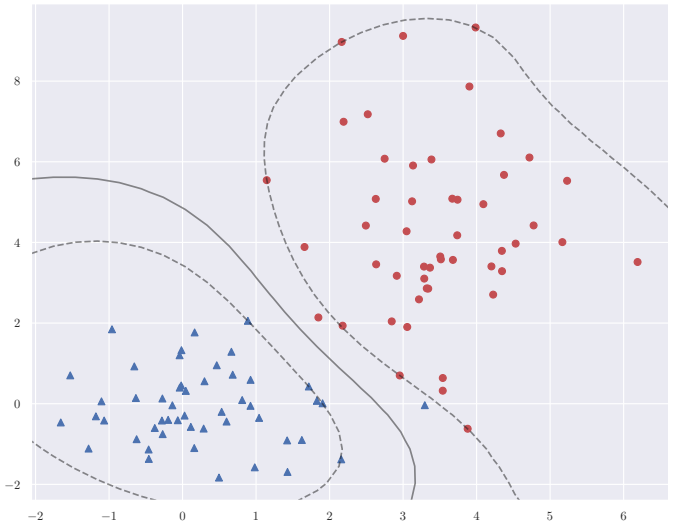


Figure 4.6: An SVC using the radial kernel. The non-linear separation boundary and margins shown are two-dimensional projections of linear separators calculated in higher dimensions.

relationships between data points in higher dimensions are calculated in the original-dimensional space [120]. Figure 4.6 shows an SVC computed using the radial kernel for the same data used in Figure 4.5. Although the decision boundary and margins shown are non-linear in two dimensions, they are in fact a projection of a linear plane calculated for a higher-dimensional mapping of the data by the radial kernel. In this example the data is for the most part linearly separable in its original two-dimensional form, but for more complex datasets, mapping to higher dimensions can improve performance a great deal. When using kernels, high ϵ values cause SVRs to calculate flatter planes when projected back to the original dimensionality.

4.3.5 DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) [132] is an unsupervised clustering algorithm that has been applied to data mining and market research [133], but is not typically used for audio applications. Since the algorithm is used in the source-tracking system detailed in Chapter 7, it will be discussed in

some detail here.

DBSCAN assigns data to clusters based on areas of density in the data. These clusters might correspond to particular classes, though since the algorithm is unsupervised, class labels are not taken into account during training and, if required, must be manually assigned by the user after clusters have been formed. DBSCAN differentiates itself from most other clustering algorithms in that it can identify clusters of any arbitrary shape without making assumptions about the general shape of the data. The GMM, for instance, assumes clusters will be Gaussian.

A given point $\boldsymbol{x}_i$ is said to belong to cluster c if there are a certain minimum number of points *MinPts* within a given radius *Eps*. These parameters interact to determine the density of points the algorithm requires to form clusters. The *Eps* - neighbourhood N_{Eps} of a point $\boldsymbol{x}_i$ within a dataset D is a set of points defined as:

$$N_{Eps}(\boldsymbol{x}_i) \triangleq \{\boldsymbol{x}_j \in D \text{ such that } \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\| \leq Eps\} \tag{4.21}$$

A pair of points is defined as 'directly density-reachable' d_{dir} as follows:

$$d_{\text{dir}}(\boldsymbol{x}_i, \boldsymbol{x}_j, Eps, MinPts) = \begin{cases} 1 & \begin{cases} \boldsymbol{x}_{i+1} \in N_{Eps}(\boldsymbol{x}_i) \\ \left|N_{Eps}(\boldsymbol{x}_i)\right| \geq MinPts \end{cases} \\ 0 & \text{otherwise} \end{cases} \tag{4.22}$$

Points that satisfy the second condition of direct density-reachability are *core* points. For the formation of a cluster there must be at least one core point, and therefore the minimum cluster size is *MinPts*. Points further than *Eps* from each other may be indirectly density-reachable via a chain of directly density-reachable points $\boldsymbol{x}_i, \ldots, \boldsymbol{x}_l$:

$$d_{\text{ind}}(\boldsymbol{x}_i, \boldsymbol{x}_l) = \begin{cases} 1 & \sum_i^{l-1} d_{\text{dir}}(\boldsymbol{x}_i, \boldsymbol{x}_{i+1}) = l \\ 0 & \text{otherwise} \end{cases} \tag{4.23}$$

Dependence on *Eps* and *MinPts* has been omitted for clarity. Clusters are formed based on chains of directly density-reachable points, such that every point in a

cluster is indirectly density-reachable from every core point. *Border* points are the end points of these chains of density-reachability, and therefore represent the edges of clusters. Direct density-reachability is therefore symmetric between neighbouring core points, but asymmetric between neighbouring core and border points. Though border points may not be density-reachable from each other, there will always be connecting points that are density-reachable from both, a condition known as 'density-connectivity' [132]:

$$d_{\text{conn}}(\boldsymbol{x}_l, \boldsymbol{x}_m) = \begin{cases} 1 & d_{\text{indirect}}(\boldsymbol{x}_l, \boldsymbol{x}_o) = d_{\text{indirect}}(\boldsymbol{x}_o, \boldsymbol{x}_m) \\ 0 & \text{otherwise} \end{cases} \tag{4.24}$$

Clusters are therefore defined in the following terms:

$$\begin{aligned} &\forall \boldsymbol{x}_i, \boldsymbol{x}_j : \boldsymbol{x}_i \in c_n \wedge d_{\text{ind}}(\boldsymbol{x}_i, \boldsymbol{x}_j, \mathit{Eps}, \mathit{MinPts}) = 1 \implies \boldsymbol{x}_j \in c_n \\ &\forall \boldsymbol{x}_i, \boldsymbol{x}_j \in c_n : d_{\text{conn}}(\boldsymbol{x}_i, \boldsymbol{x}_j, \mathit{Eps}, \mathit{MinPts}) = 1 \end{aligned} \tag{4.25}$$

where $c_1, \ldots, c_n$ are the clusters formed with respect to *Eps* and *MinPts*. Any points that do not fall into any clusters are designated as noise and are defined as:

$$\boldsymbol{x}_i \in D \,|\, \forall n : \boldsymbol{x}_i \notin c_n \tag{4.26}$$

Figure 4.7 shows a visualisation of the operation of the DBSCAN algorithm on some randomly-generated data using $\mathit{Eps} = 0.8$ and $\mathit{MinPts} = 4$. It can be seen that using these parameters, the algorithm has identified three clusters (denoted by the colours brown, red and purple), with several points designated as outliers, as shown by the crosses. The larger circles in each group represent points the algorithm has designated as core points, with the smaller circles indicating border points. As can be seen from the purple group, it is possible for dense clusters to consist entirely of core points, whereas in the brown group there is only one core point. The diagram indicates one of the continuous paths of density-reachability in the red cluster, from

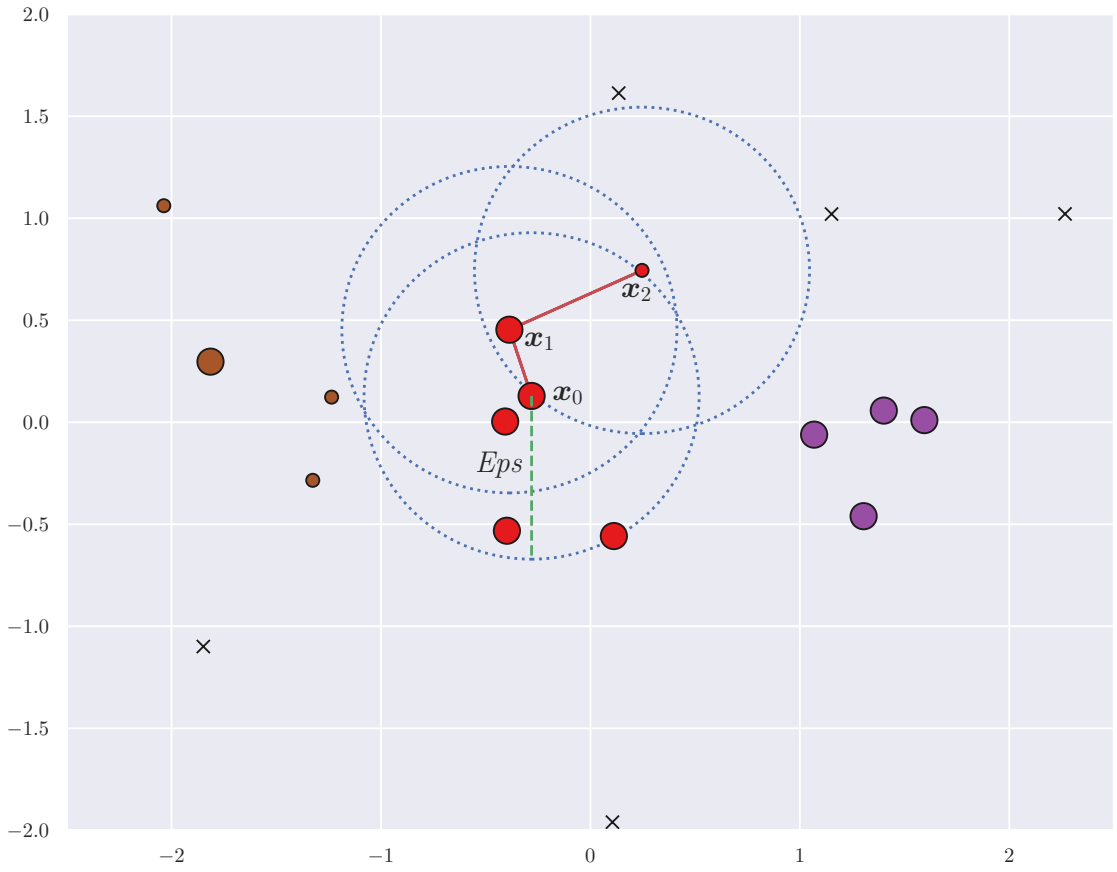


Figure 4.7: Visualising the DBSCAN algorithm. Three clusters have been identified in this data. Red lines indicate one density-reachable path within the red cluster. Dotted circles indicate the *Eps*-neighbourhood of the points at their centres.

an arbitrary core point $\boldsymbol{x}_0$ to the cluster's only border point $\boldsymbol{x}_2$. It can be seen from the dotted circles indicating the *Eps* - neighbourhood for each point in the path that points $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ satisfy $\left|N_{0.8}(\boldsymbol{x}_i)\right| \geq 4$, whereas point $\boldsymbol{x}_2$ does not and has been designated a border point accordingly.

4.3.6 Neural Networks

In recent years, neural networks have increasingly become the dominant models used for many machine learning tasks, including audio applications. The vast majority of the top-performing systems in the DCASE challenge in recent years have utilised

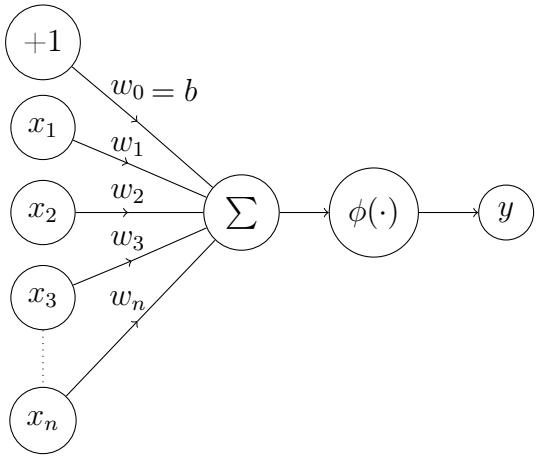


Figure 4.8: A perceptron - the simplest neural network.

convolutional neural networks (CNNs) in various configurations [101, 134]. A short is overview is presented here. A thorough mathematical description is beyond the scope of this thesis and the interested reader is referred to [135] for more detail.

A single-neuron network, or perceptron, consists of four components [135]:

1. A set of weights $\boldsymbol{w}$.
2. A summation stage to add the inputs $\boldsymbol{x}$ as weighted by $\boldsymbol{w}$.
3. An activation function $\phi(\cdot)$ to map the summation to an output y.
4. A bias b which raises or lowers the sum to the activation function.

Figure 4.8 shows the signal flow of a perceptron, where the bias has been formulated as a weight w_0 for a fixed input of +1. Typically, the activation function is the Heaviside step function, which essentially thresholds the summation of the perceptron, producing a binary output:

$$\phi(\cdot) = \begin{cases} 1 & \boldsymbol{w} \cdot \boldsymbol{x} + b > 0 \\ 0 & \boldsymbol{w} \cdot \boldsymbol{x} + b \leq 0 \end{cases} \tag{4.27}$$

It should be reasonably clear that the perceptron functions in a way that is essentially identical to the linear SVC described in Section 4.3.4. The difference here is that, unlike the SVC, a perceptron is not concerned with maximising a margin, but is instead optimised by iteratively updating the weights based on comparison of the network outputs with target values [120, 135]:

$$\boldsymbol{w}^{\text{new}} = \boldsymbol{w}^{\text{old}} + e\boldsymbol{x}_i \tag{4.28}$$

where $e = (y_i^{\text{true}} - y_i^{\text{pred}})$. Typically, the weights are initialised to random values, and, provided that the inputs are linearly separable, each iteration of this equation (known as an 'epoch') will converge on values that yield correct outputs. A perceptron may have multiple neurons to perform multiclass classification, each neuron calculating a separate linear hyperplane to distinguish between classes in a one-versus-all fashion. In this case the weights become a matrix and the update algorithm is:

$$\boldsymbol{W}^{\text{new}} = \boldsymbol{W}^{\text{old}} + \boldsymbol{e}\boldsymbol{x}^T \tag{4.29}$$

Ultimately, single-layer perceptrons are limited in that they are only effective classifiers when classes are separable by a linear function. This fundamental limitation caused research to stall for many years [120]. Where an SVM overcomes this problem using kernels, with neural networks the solution is to add additional layers of neurons, creating a multi-layer perceptron (MLP). Figure 4.9 shows a basic MLP with two fully-connected 'hidden' layers, layers that are not directly visible to the inputs and outputs. Each solid line is associated with a weight, whilst each circular 'node' in the diagram represents a combination of the summation and activation stages which are shown separately in Figure 4.8. A variety of activation functions can be used in place of the Heaviside function, including sigmoid curves such as tanh, and notably the rectifier, which is zero for any value below zero and linear for positive outputs e.g. $f(x) = \max(0, x)$ [106, 135]. MLPs can have any arbitrary topography, with the depth of the network defined as the number of layers and the width defined as the maximum number of nodes in any of its layers.

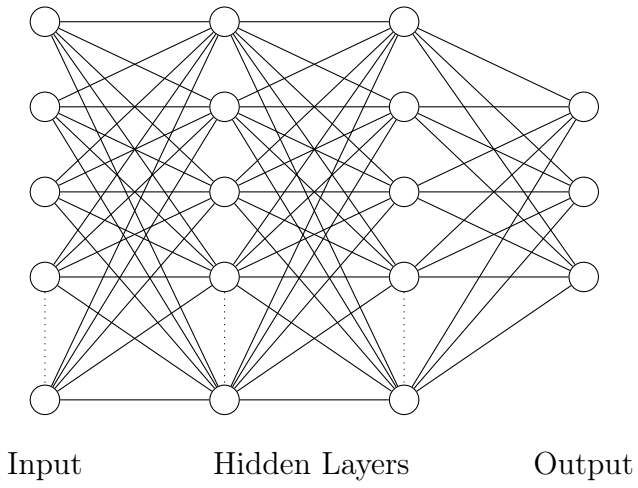


Figure 4.9: Multi-layer perceptron with two hidden layers. Circles are nodes representing a combined summation and activation function, and connecting lines each have an associated weight.

Updating the weights of the MLP and more complex neural networks is done by computing the error at the output layer of the network using a loss function, and adjusting the weights back through the network based on the derivative of this loss in order to iteratively find the minimum error, in a process known as *backpropagation* [135]. Often an update of the weights will be undertaken after presenting a number of examples (a *batch*) and averaging the error. This is known as *batch learning*.

Theoretically, MLPs of varying topographies can learn any given function [135], though this property also means larger networks have a tendency to overfit, becoming very highly adapted to the input data, including any noise it contains, with detrimental effects when tested on unseen data. For this reason, MLP architectures must be tuned to the given problem. MLPs can also be augmented with 'recurrent' layers. In a recurrent neural network (RNN) the output from the previous feature vector is fed back and combined with the current vector. This allows the network to take into account data received prior to the current input, making RNNs especially well suited to time-series data [135].

As either feedforward or recurrent networks become larger, however, the number

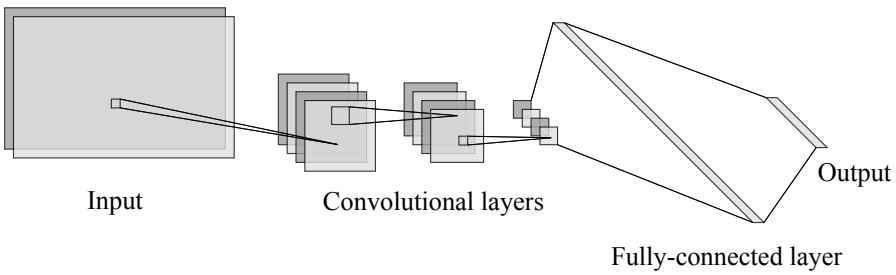


Figure 4.10: Diagram of a simple CNN, featuring three convolutional layers followed by a final fully-connected layer. The convolutional layers automate feature extraction, whilst the fully connected layer has an architecture similar to the MLP shown in Figure 4.9.

of connections, and hence weights that must be computed, grows exponentially. This causes the computation time required to train a network to increase significantly. To solve this problem, the dimensionality of the input features can be manually reduced, perhaps modifying a feature extraction stage to emphasise more salient features [106]. Depending on the problem, this might be very time-consuming. Another alternative is to use convolutional neural networks (CNNs). CNNs consist of a number of convolutional layers followed by one or more fully-connected layers that compute the final output of the network. Typically, CNNs operate on input data represented as a two-dimensional matrix, such as images or time-frequency representations of audio, though any N-dimensional matrix can be used with convolution kernels of corresponding dimensionality. Figure 4.10 depicts a simple CNN with three convolutional layers followed by a single fully-connected layer that computes the output.

A two-dimensional discrete convolution is defined as [136]:

$$x * * h = \sum_{m_1=-\infty}^{\infty} \sum_{m_2=-\infty}^{\infty} h[m_1, m_2] x[n_1 - m_1, n_2 - m_2] \tag{4.30}$$

where n_i are the dimensions of the data matrix x and m_i are the dimensions of the kernel (impulse response) h. This is equivalent to iteratively moving the kernel h

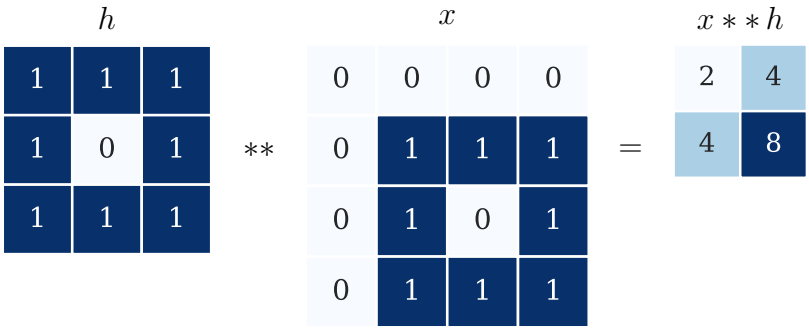


Figure 4.11: Illustration of a basic 2D convolution kernel. The convolution of the kernel h with the input x yields an output matrix which has higher values where the kernel and the input more closely match.

across x, performing elementwise multiplication (Hadamard product) at each stage and taking the grand sum of each result. Figure 4.11 shows a example of a convolution using a basic kernel that could be thought of as a detector of small squares. As can be seen, the output of the convolution is large where the square is present and smaller further away.

Part of the power of CNNs is that the values of the kernels are weights that can be iteratively trained to yield better output accuracy in the same way that the weights of fully-connected layers are trained using backpropagation. There can be multiple 'channels' at the input, perhaps corresponding to different microphone inputs or colour channels in an image. Channels in convolutional layers each represent the output of a different kernel, so multiple salient features can be detected at each stage. The network shown in Figure 4.10 has two input channels (perhaps corresponding to stereo audio), and each convolutional layer has four channels. Convolutional layers often have many more channels than this in practise. Each successive convolutional layer creates increasingly abstract representations of the input data, with the output from the final convolutional layer 'flattened out' and passed to the fully-connected layers. In reducing the dimensionality of the input before presenting it to these fully-connected layers, CNNs are effectively able to automate the extraction of salient

features.

CNNs are the current state-of-the-art in most fields in which machine learning models are employed, and one need only look at recent CASSE research [101, 134] for a testament to their capability and dominance in the field. For audio applications a mel-spectrogram is typically used as input to the network. Many of the network architectures initially tested for audio applications were adapted from the field of image recognition, in particular VGG (Visual Geometry Group) [137]. More recently, however, it has been observed that improvements in performance by newer architectures in the image domain such as ResNet [138] and DenseNet [139] do not necessarily transfer to the audio domain. This is potentially due to the comparative scarcity of data typically available for audio tasks relative to image tasks, and the tendency of deep networks to overfit to smaller datasets. Recent studies have indicated that reducing the size of the convolution kernels, especially in the frequency dimension, can lead to improved performance in ASC [140]. These findings indicate that the temporal structure of sounds are more important for CNN performance than the associated frequency information.

Whilst CNNs provide state-of-the-art performance, their ultimate downside is that it is often difficult to infer exactly which properties of the input data the network is using to draw its conclusions, meaning that the utility of the models is limited if the objective is a detailed exploration of the data itself rather than simply achieving accurate predictive output.

4.4 Some Example Systems

4.4.1 The Bag-of-frames Approach

A very common method for audio classification is using MFCC features with GMM classifiers. Aucouturier *et al.* describe a technique using this combination, known as the 'Bag-of-frames' approach (BOF) [141]. With BOF, the order of events has

no bearing on the model that is ultimately calculated.

The work presented in [141] is a direct adaptation of earlier work using the MFCC+GMM technique in an MIR application for assessing the timbral similarity of different pieces of music [142]. Whilst for the MIR task the authors found that accuracy saturates at around 65% regardless of the different variations on the algorithm used, for ASC the authors claim "near-perfect precision" in identification (over 90% in this study). Given three seconds of sound using this technique, the system was able identify the correct label 91% of the time, whereas humans only managed 35%, needing around 20 seconds of sound for high accuracy. In music, however, this is reversed, with humans shown to give accurate music recognition with as little as 200 ms of audio, whereas the algorithm needed upwards of 60 seconds for similar performance. This leads to an interesting conclusion regarding the BOF method, namely that the "*perceptive saliency* of sound events is modelled as their *statistical typicality*" [141]. This is arguably the exact opposite of human perception, which tends to be attracted to novelty. Humans pay attention to unusual sounds whilst ongoing background sounds tend to be filtered out, facts that inform Schafer's concept of *signal* and *keynote* sounds.

The performance of the algorithm with music signals was tested against its performance using three-minute monophonic recordings made at locations across Paris, including street and market scenes. It was shown that temporally homogenising the input recordings by replacing large portions of the original recordings with repeated subsections ('folding') had much less of an impact on classification of the acoustic environments than the music files. With 50-folding (the first 1/50th of the recording is repeated 50 times), the loss of accuracy in music files is more than 60%, whereas for acoustic environment recordings the reduction in accuracy is only 20% [141].

These findings reveal that the MFCCs extracted from the urban soundscape recordings used here have high statistical homogeneity - most MFCC frames are very similar and so are informationally redundant relative to music signals, which have much more varied MFCC frames. This would no doubt be unsurprising to Schafer,

who termed urban soundscapes "lo-fi", in that broadband mechanical sounds tend to mask the *signal* sounds that carry information [3]. Schafer contrasts this with the "hi-fi" soundscapes of the natural world, which lack significant mechanical sound and offer a greater sense of "perspective". This sentiment is echoed by Wiseman and Wilson [143], who identify the natural world as "information-rich". It should be noted that Aucouturier *et al.* did not evaluate natural or rural soundscapes.

Although these results were promising, almost to the point of indicating that acoustic scene classification was a solved problem, there has subsequently been much criticism of their findings. Lagrange *et al.* [144], re-ran the study using the original dataset and were able to replicate their results, but could not do the same with alternative recordings. Table 4.1 shows Lagrange *et al.*'s results with the original and three alternative datasets. It can be seen that the very high accuracy using the original dataset is not replicated in any other case. Lagrange *et al.* posit that the unusually high accuracy of Aucouturier *et al.*'s original results could be caused by the "overly permissive" structure of their dataset, whereby most of the test audio consisted of different segments of the same longer recordings as the training audio. This is likened to similar misleading accuracies reported in early MIR studies where the testing sets used songs from the same artist or album as the training sets, the so-called 'album effect' [125, 145]. As a result of this, subsequent datasets created for ASC studies have made sure to include "many different locations and situations for each scene class" [10, 47, 125].

4.4.2 Machine Listening Using Spatial Features

The vast majority of machine listening research focuses on monaural recordings, with a few using binaural recordings [8, 55] and fewer using higher numbers of channels [70, 146]. This is potentially due to much of the research inheriting techniques from ASR/CASA, but also influenced by the common focus on applications including use in wearable tech, smartphones and robotics, [54] where use of large microphone arrays would not be feasible. It should be noted, however, that recent developments

Database	**Chance**	**BOF**
Aucouturier (Original)	44 ± 20	97 ± 12
Gustavino	28 ± 10	40 ± 10
Tardieu	33 ± 18	48 ± 20
QMUL	28 ± 13	48 ± 21

Table 4.1: Percentage mean and standard deviation of acoustic environment identification performance (precision-at-rank-5) (After [144]).

in mobile technology may make this less of a limitation in the future.

Multiple-microphone recordings contain spatial information that presents another potential source of features that has been, so far, relatively overlooked. Nogueira *et al.* [55] extract data from stereo recordings using a model of binaural hearing to estimate features including the Interaural Time Difference (ITD) and Interaural Level Difference (ILD), which they add to the standard MFCC feature set. The stated aim of this is to assist in disambiguation between similar scenes, for instance the possible ambiguity between a train station and a train interior. Whereas train sound will be present in both scenarios, at the station the sound might be expected to be more directional in nature. This use of spatial information as essentially a secondary feature is consistent with its theoretical weighting in human ASA [100].

The work of Bunting *et al.* [70, 146, 147] on the ISRIE project utilised directional audio coding (DirAC) analysis (covered in Section 6.2.2) to assist in source separation for B-format acoustic environment recordings, though ultimately the sound separation used only the mono omnidirectional channel. They found that it was much more difficult to separate mechanical sources from one another as their broadband frequency content and continuity over time meant that, even when utilising a time-frequency transform, ω-disjoint orthogonality (a measure of how separated two signals are in frequency and time) was low as the mechanical sources contained

largely overlapping components. Bunting's system was able to separate birdsong from mechanical sound, as the sound level of the birdsong was able to overcome the broadband mechanical noise in the narrow bands in which it was present.

DCASE challenges since 2016 have provided binaural audio datasets for their ASC and SED tasks. The top-performing ASC systems in 2016 and 2018 made use of the extra data by simply extracting more MFCCs from the available channels and combinations thereof, without investigating the potential for new features specific to the spatial format [148, 149]. The majority of systems submitted did not exploit the binaural information at all [150]. The leading SED systems in the DCASE 2016 challenge, presented by Adavanne *et al.* [151] utilise mel-band energies, with time difference of arrival (TDOA, equivalent in this case to ITD) of sound between the pair of microphones calculated for five different mel bands using the generalised cross correlation-phase transformation (GCC-PHAT) method [152]. The authors note TDOA values calculated from short frames tend to be noisy, and so extract TDOA values from three frame lengths, taking the median of these for a single-value estimate. A RNN is used as classifier. Their submission for the following year's challenge [107] once again used the mel-band energies, but phased out the TDOA features, with calculation of spatial properties being offloaded to a CNN. The input to the network was the complex STFT from each frame, following similar work by Chakrabarty and Habets [153], who used phase spectra as input to a CNN for the purpose of broadband DOA estimation in multichannel audio. This method outperformed its predecessor, but the specific spatial properties that contributed to this are not clear.

Despite the assertion at its foundation regarding reduced focus on "human-centric aims", it was only in the 5th edition of the DCASE challenge (2019) that the challenge expanded from the binaural format and a sub-task was set up explicitly investigating the potential for the Ambisonic spatial audio format in SED [134]. Adding the additional requirement for submissions to estimate the direction of arrival (DOA) of sounds, the sound event localisation and detection (SELD) task provided synthetic

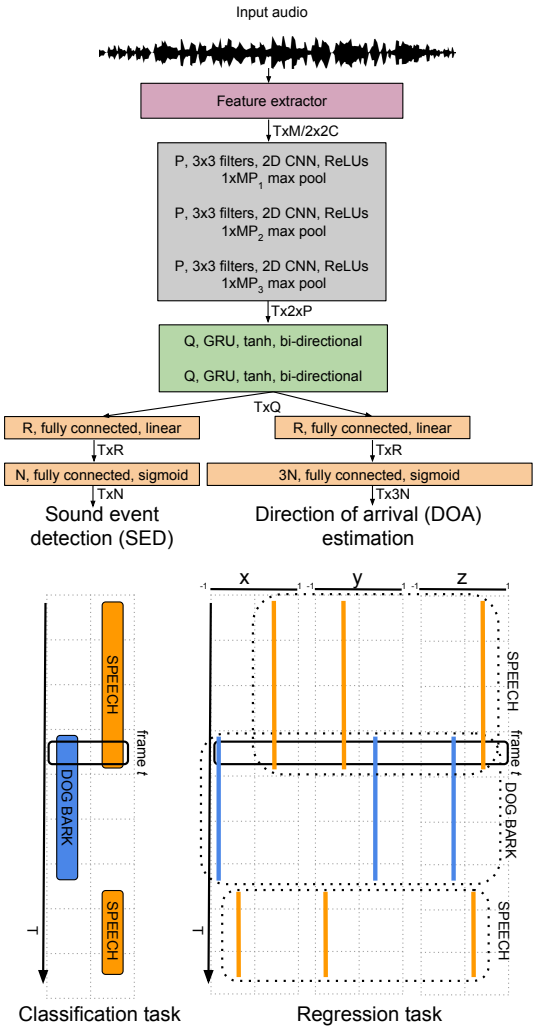


Figure 4.12: SELDnet network architecture (from [154]).

training data in first-order Ambisonic (FOA) B-format and 'four-channel directional microphone' A-format. The baseline system and the majority of submissions for this task use CNNs of various configurations, "thereby avoiding any method- and array-specific feature extraction" [154], though this potentially overlooks any benefits peculiar to the spherical harmonics-based Ambisonic format.

The task built upon SELDnet by Adavanne *et al.* [154], which has an architecture derived from their previous binaural method. SELDnet predicts sound class and activity concurrently with DOA from multichannel audio using magnitude and phase

spectrograms extracted from all four first-order Ambisonic channels as input. The network is designed so all convolutional and recurrent layers are shared between the two tasks, with a final pair of fully-connected layers set to output sound event estimates and another to output DOA estimates. This configuration, shown in Figure 4.12, could be conceptualised as two MLPs performing different tasks using the same features which have been extracted by convolution and recursion. Given that backpropagation will take into account the loss from both outputs, however, the network is theoretically able to learn the DOAs and movements associated with given sound classes. Performance is very good in both SED and DOA estimation when there is only one active source, but becomes more erratic when there are multiple concurrent sources. Mesaros *et al.* note that this is also a problem in human event recognition, "when the acoustic power of the environmental noise is too high compared to individual events, we simply do not hear or recognise them anymore" [110].

Notably, the network is outperformed in terms of DOA error by the MUSIC algorithm [155] in the case of anechoic Ambisonic audio, whilst the earlier DOAnet [156], by the same research team, outperforms SELDnet in all reverberant scenarios. DOAnet performs DOA estimations only and includes an intermediary stage estimating a MUSIC-like spatial pseudo-spectrum (SPS), which is a map of sound power varying with direction. This suggests that estimating a power map might be beneficial for this task. In terms of SED, the earlier MSEDnet [157] tends to outperform SELDnet. This network is essentially similar to the team's DCASE 2016 method, adding convolutional layers but retaining the earlier method's TDOA feature extraction stage. These results suggest that retaining separate models to perform SED and DOA estimation might be more optimal than attempting to combine the two tasks, however beneficial relating DOAs to sound classes might intuitively seem.

Investigation of the potential of crafted (e.g. not inferred by a CNN) spatial features for ASC and SED tasks is one of the main novel contributions of this thesis, as detailed in Chapters 6 and 7. Whilst the end-to-end automated CNN approach

has its benefits and can perform well, the workings of CNNs tend to be rather opaque, so it is preferred to keep the feature extraction manual in order to be better able to assess the merits of various features.

4.5 Summary

In this chapter, the origins of machine listening in speech recognition and music information retrieval have been introduced and related to the application of machine listening in this thesis, namely the identification of everyday sounds. The two main tasks of the DCASE challenge, acoustic scene classification and sound event detection, have been introduced alongside the theoretical framework that underpins these and other single-label classification and detection tasks in machine learning. The bulk of the chapter has been a detailed exploration of features and classifiers, concentrating on those prevalent in ASC and SED approaches from the literature and those which have been used in the work which will be presented in this thesis. The chapter concludes with a summary of some key example machine listening methods from prior work. With the background underpinning this thesis now established, the next chapter will explore the first stage of the original work presented here, namely the creation of the EigenScape database of spatial acoustic environment recordings.

5 | The EigenScape Database

Associated Publications

- M. C. Green and D. T. Murphy, "EigenScape: A Database of Spatial Acoustic Scene Recordings", *Applied Sciences* vol. 7, no. 11:1024, November 2017, doi: 10.3390/app7111204

Contributions

- The EigenScape database, which is the largest publicly available set of acoustic scene recordings in the standardised fourth-order Ambisonic format.
- Data partition setup and software tools to complement the dataset.

5.1 Introduction

In Chapter 3 the key concepts of the soundscape approach, as distinct from the prevailing 'environmental noise' approach, were detailed. Although the soundscape approach has many benefits in terms of its focus on human perception, the relative difficulty of quantification and measurement of acoustic environments using this approach was identified as a major drawback. The ecological validity of lab-based acoustic environment reproduction, and in particular how the experience of an acoustic environment over a long soundwalk might be condensed into a shorter recorded version, was proposed as a potential area for advancement.

Chapter 4 explored the field of machine listening, from low-level detail of certain algorithms to high-level applications including ASR, MIR and latterly, CASSE. The objective of CASSE, namely holistic analysis of acoustic environments in and of themselves, aligns closely with the goals of the soundscape approach. If a reliable system for the automatic annotation of acoustic environments could be created, this would have applications including assisting in generating ecologically-valid acoustic environment condensations. A less-detailed model could, perhaps, improve the derivation of soundecology metrics such as the NDSI, where the current frequency-band based calculation makes it ineffective for use in urban environments, as detailed in Section 3.4.2.

Part of the ethos of CASSE is a move away from the motivation of biological relevance (found particularly in ASR), and to re-focus entirely on analysing audio data with no artificially-imposed limits on the means of doing so. Despite this, partly because of the legacy of techniques inherited from ASR and MIR, and partly because of oft-stated applications in robotics [158], smart devices, and wearable technology, where the mounting of large microphone arrays may not be practical [54, 154], the various tasks in the DCASE challenges have used only monaural or binaural recordings until very recently.

Spatial audio offers many benefits from the point of view of both CASSE and

soundscape:

- As described in Chapter 2, spatial variation is a fundamental feature of sound fields. Acoustic environments might vary spatially in ways that are location-characteristic and distinct from how they vary spectrally.

- Ambisonic recordings can be presented over a loudspeaker array, which could offer increased immersion and ecological validity over a mono or stereophonic presentation. Binaural presentations taking into account a listener's head rotation are often derived from Ambisonic recordings [45, 159].

- Multichannel recordings make source tracking and source separation more feasible, for instance by using beamforming, as described in Section 2.4.2.

This chapter will cover the recording of the EigenScape database of acoustic environment recordings. The database was created as a foundation for the investigations undertaken in this thesis into the utility of spatial information for CASSE, motivated by the soundscape approach. First, there will be a review of existing datasets. This will show points of inspiration for EigenScape, and justify the recording of new data. There will follow a detailed description of the EigenScape recording process, including equipment used, locations visited, publication information, and the subsequent partitioning of the data to facilitate investigation.

5.2 Existing Datasets

To begin an investigation into whether spatial information in acoustic environments can be used to enhance the understanding gained using a machine listening system, it makes sense to start with a relatively high-level analysis. As noted by Torija *et al.*; "categorisation of a soundscape is the first step to evaluate it" [160]. It was therefore decided that the initial exploration in the research that forms this thesis would be an investigation into use of spatial features for acoustic scene classification.

If it could not be shown that spatial features were a useful discriminator for entire acoustic environment classes, their utility for individual acoustic events might be questionable.

5.2.1 DCASE Datasets

For their ASC tasks, the regular DCASE challenges since 2016 have used various iterations of the Tampere University (TUT/TAU) database [125, 161, 162]. The original version of this database consists of binaural recordings of 15 different environment classes: lakeside beach, bus, cafe/restaurant, car, city centre, forest path, grocery store, home, library, metro station, office, urban park, residential area, train, and tram. Multiple examples of each were recorded at locations in Finland using the Soundman OKM II Klassik/studio A3 electret in-ear microphones and Roland Edirol R09 wave recorder. All original recordings were between 3-5 minutes in length and were segmented into 30-second clips for a total of 104 segments, making up 52 minutes of audio per class. A development set up was provided assigning 75 % of the data to a development set and the remaining 25 % to an evaluation set, making sure that there was no overlap in recording location between each set, in order to avoid the 'album effect' [144] covered in Section 4.4.1. Later versions of the database reduced the number of classes to 10 but expanded recording locations across Europe, increasing the amount of data for each class available to 144 minutes in 2018 [163] and 240 minutes in 2019 [164]. The TUT/TAU database is rigorously produced and provides an ideal model for ASC tasks, but is limited to binaural channels, making derivation of full-3D spatial features somewhat difficult.

The original DCASE challenge in 2013 used a set of 11 first-order Ambisonic (FOA) recordings for its SED task [10]. These were, however, limited to scripted recordings made in office environments, reflecting its intended purpose for SED, rather than the broad range of location classes required for ASC. Furthermore, only stereo versions of these recordings were released publicly. More recent DCASE challenges have used synthesised FOA scenes for their sound event localisation and

detection (SELD) tasks [165, 166, 167]. Though these datasets were synthesised using impulse responses from real rooms, with later versions using many source-receiver positions to simulate smooth source movement, the number of simultaneous sound sources is limited to two, meaning these scenes have only a fraction of the complexity of most real-world environments. Additionally, source sounds were taken randomly from the DCASE 2016 isolated sound events [168] and NIGENS general sound events [169] datasets, and spatialised using random DOA positions and trajectories, meaning that no potentially characteristic spatial features of events in real-world acoustic environments were modelled. For a spatial ASC task, then, a synthesised set such as this is not appropriate.

5.2.2 Spatial Audio Datasets

There are existing datasets featuring spatial recordings of real-world locations. The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND) database was presented in 2013 [170], with the purpose of aiding development of noise reduction algorithms for source separation. DEMAND is presented as a freely available alternative to paid datasets such as the AURORA [171] and NOISEX [172] databases. Six different location classes are included: Domestic, Office, Public, Transportation, Nature, and Street, with three different examples of each. Recordings are five minutes in length, giving 15 minutes total per class. This is a substantial amount of data, but potentially too small a dataset to provide effective classifier training and testing, especially compared with the amount of data available for each class in the TUT/TAU datasets. Further, the setup used to record DEMAND was a custom-built planar array featuring 16 microphones arranged in a grid. This offers a great deal of potential for spatial information, but elevation data might be more difficult to derive than azimuth, and techniques developed using data from this array might not be generalisable to other recordings in the way that the standardised Ambisonic format allows.

Perhaps the closest precedent to the type of dataset required for spatial ASC is

the work of Stevens *et al.* [20, 45, 80], who used the STM450 SoundField microphone array [173] to make 10-minute first-order Ambisonic recordings of acoustic environments at eight locations in and around Leeds and the North York Moors National Park, England. Locations were selected to give a broad range of urban and natural environments for the purpose of lab-based listening tests investigating soundscape perception. Assessing this data with a regard to use in ASC, the locations are broadly grouped into three classes: rural, suburban and urban. There are two recordings of rural locations and three each for suburban and urban, yielding 20-30 minutes of audio per class. This could be enough for a small pilot study into spatial ASC, but the uneven amount of data per class could make creating balanced test setups somewhat difficult. A much larger database in this format would be ideal. EigenScape was therefore conceived as essentially an amalgamation of the recording approach of Stevens *et al.* and the breadth of locations and data available in TUT/TAU.

5.3 Specification

In reviewing the existing available data, it became clear that a new set of recordings was needed to facilitate a comprehensive investigation into spatial features for ASC. The new dataset was created on the following criteria:

- Samples of a broad range of urban and natural environments.
- On the order of 50 - 100 minutes of audio for each class for comparability with at least the initial 2016 version of the TUT/TAU dataset.
- Equal amounts of audio for each class for balanced train/test partitioning.
- Ambisonic format to facilitate spherical harmonic processing and for transferability of results in a common format.

Figure 5.1: The Eigenmike spherical microphone array.

5.3.1 Equipment

The array chosen to record the new database was the mh Acoustics Eigenmike [32, 33], some detail of which has already been covered in Section 2.4.1. Figure 5.1 shows the array, capable of recording up to fourth-order Ambisonic format.

In a two-part investigation, Bates *et al.* compared the performance of various Ambisonic microphone arrays, including the Eigenmike and SoundField MKV, assessing both perceptual audio quality and directional accuracy [174, 175]. In these tests, the Eigenmike fared poorly in terms of perceptual audio quality and was rated as "dull" compared to other microphones. It is surmised this may be related to the large number of microphone capsules in the array contributing a greater amount of noise in the output signals than lower-order arrays with fewer capsules. In directional tests, however, the Eigenmike significantly outperforms the other arrays. This is despite the fact that, for parity, only the first-order channels from the Eigenmike were considered in these tests. These results indicate that whilst the SoundField microphone is probably a better choice for recordings to be used in listening tests such as in [20], the Eigenmike is a good choice for the work in this thesis, given

that the primary concern is machine (rather than human) listening. The Eigenmike is capable of recording far more detailed spatial information than any first-order array, whilst remaining practical to transport and use on-location. Given the choice of array, the new database was named *EigenScape*.

Recordings were made using the Eigenmike interface box [34] and EigenStudio recording software [176] running on an Apple MacBook pro [177] at 24-bit/48 kHz resolution. Audio was recorded simultaneously as raw signals from the 32 microphones and in Ambisonic format, encoded by EigenStudio using ACN channel ordering [178] and SN3D normalisation [28]. EigenStudio also automatically applies the high-pass filters required to compensate for array self-noise in the higher-order channels, as outlined in Section 2.4.1. The recording gain was set in the EigenStudio software to +25 dB. This high gain was necessary as the majority of recording locations did not yield a strong recording level using lower gains. The only exception to this was one train station location where very high level engine noise caused severe clipping at +25 dB, so a gain of only +5 dB was used, as noted in the database metadata, included in Appendix A. Whilst a systematic gain discrepancy, for instance if all examples of one class were recorded at a consistently different gain level, would represent a significant source of bias in the dataset, as there is in fact only an isolated discrepancy in a single recording, any bias should be minimal.

For the majority of the recordings, the Eigenmike was placed within a Rycote modular windshield [179]. Although the windshield was not designed for microphone arrays with such a large diameter as the Eigenmike, care was taken to mount the array securely and the shield was effective in reducing wind noise. An initial number of recordings used a custom-made windshield, but this was replaced with the Rycote as setup time proved too long. One indoor recording used no windshield. It is thought that the recording discrepancies incurred in windshield use should be negligible in comparison to the wide range of sound sources and acoustics present in the recorded environments. This should especially be the case when coarse features are extracted from the audio for use in a machine listening system. For such a sys-

Figure 5.2: The EigenScape recording setup at Redcar Beach, North Yorkshire (recording of Beach-04).

tem to be at all effective, it should be robust to the small spectral changes incurred by the use of different windshields, as well as in ambient sound level between scenes. Indeed, another database created before EigenScape, DARES-G1 [180] used completely different recording equipment for indoor and outdoor locations, and this was judged to have "minimal influence on the quality of the database". Nevertheless, all small recording discrepancies are noted in the EigenScape metadata (Appendix A).

To make the recordings, the Eigenmike was mounted on a standard microphone stand set to around head height. A Samsung Gear 360 camera [181] was mounted to the same tripod beneath the microphone to record video data, with a view to assistance in annotation of sound events where the source might be ambiguous, or potentially to use in future immersive listening tests using VR. Ultimately the video data was not used in the work presented in this thesis. Figure 5.2 shows the full recording apparatus.

Class	Typical Sounds
Beach	Ocean waves, wind
Busy Street	Heavy traffic, pedestrian crossings, speech
Park	Birdsong, speech, children playing
Pedestrian Zone	Speech, footsteps, street music
Quiet Street	Light traffic, birdsong
Shopping Centre	Speech, background music, reverberation
Train Station	Train engines, announcements, speech
Woodland	Birdsong, occasional footsteps

Table 5.1: EigenScape Classes

5.3.2 Locations

Location classes in EigenScape were based upon the classes featured in the TUT/TAU database, with a focus on open public spaces. The eight classes included in EigenScape are listed together with some typical sounds in Table 5.1. These classes were chosen to give a good variety of acoustic environments found in or near urban areas and range between those dominated by mechanical sound (Busy Street, Train Station), to those featuring largely natural sound (Woodland) and some in between (Park, Quiet Street). Figure 5.2 shows a typical example of a Beach location. Images of further locations are available in Appendix A.2.

Eight different examples of each class were recorded, for a total of 64 recordings. All recordings are exactly 10 minutes in length, giving 80 minutes per class - just under 11 hours in total for the whole database. In practise, a little over 10 minutes was recorded at each location, with the very beginning and end trimmed to remove the noise incurred by the activation and deactivation of the recording equipment. This facilitated segmentation of the recordings into clips of equal length, giving many options for the length that do not leave any audio left over (e.g. 20×30s segments,

60 × 10s segments, etc.), so the recordings can be easily split into training and testing sets. Recordings were made across the north of England in May 2017 at locations somewhat determined by ease of access from the AudioLab at the University of York. Figure 5.3 shows the location of all the recordings with a legend of icons indicating the classes, whilst Figure 5.4 shows a closer view of an area in the city centre of York where a higher concentration of recordings were made. Appendix A.2 contains maps of the other city centre recording locations, whilst Appendix A.1 contains a complete list of EigenScape recording locations. This includes information on the date and time of each recording together with notes on any discrepancies. Consent from individuals is not required when making recordings in public spaces in the UK [182], but permissions from all relevant local authorities and premises management were sought and obtained where possible. Some locations would not allow tripod-based recordings, so the microphone stand was collapsed and held as a monopod. All relevant documentation, including risk assessment, letter of support (including a method statement), and a written permission obtained from Scarborough Council are included in Appendix A.3.

Apart from one single scene (Beach-08), all recordings were made between the standard working hours of 9 am - 5 pm. This was partly for practicality, but also to reduce discrepancies to the acoustic environments that might be caused by varying activity levels according to the day/night cycle. Every effort was made to avoid introducing sound sources to the scene by either the equipment or the experimenter. In some of the locations with heavy footfall, conversations between the experimenter and curious passersby have been recorded. This was quite unavoidable in some of the busier locations, but since conversation tended to be a typical part of such environments, this is not too anomalous and should certainly not affect feature extraction. Discretion is recommended if these recordings are to be used in a listening test.

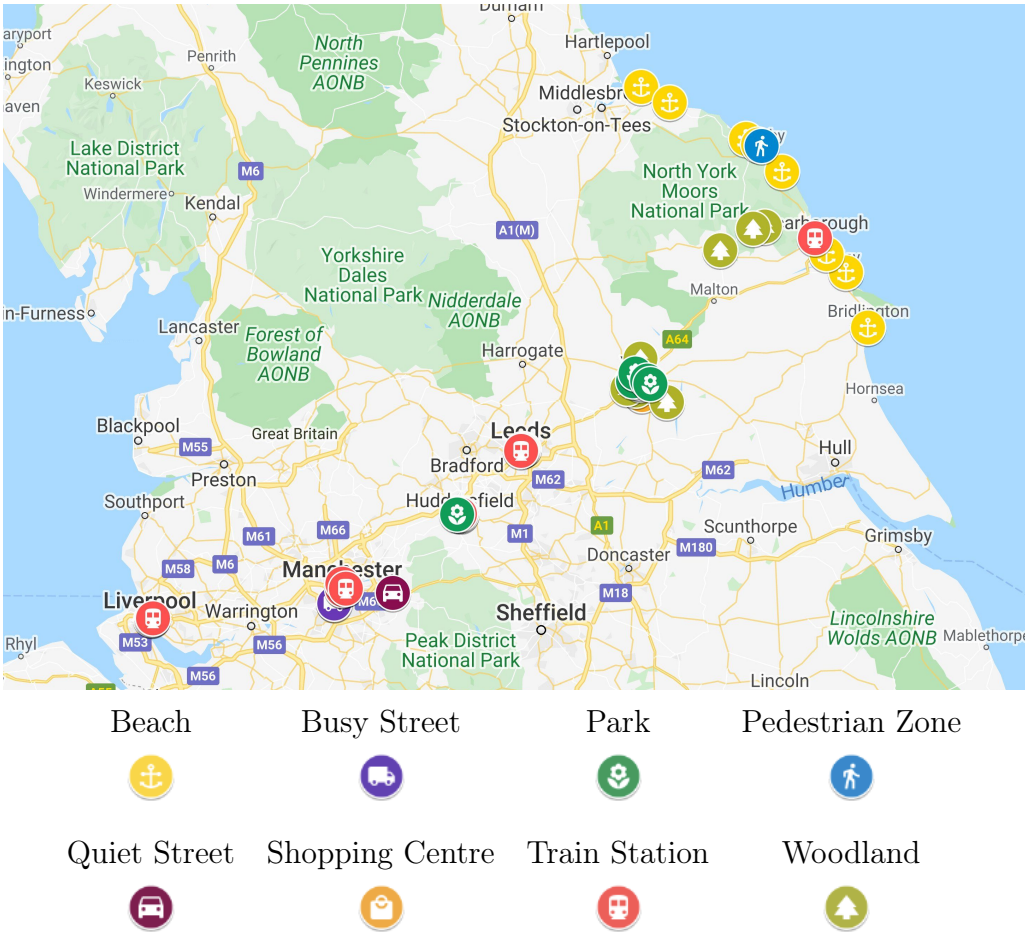


Figure 5.3: EigenScape recording locations across the north of England.

5.3.3 Publication

Apart from being an essential foundation for the work in this thesis, EigenScape was also envisioned as a platform for other researchers interested in the topics of spatial audio and soundscapes. The complete dataset has therefore been made freely available in Ambisonic format for download online via the CERN-administered Zenodo platform [183], under the Creative Commons Attribution 4.0 license [184]. The data is presented as WAV files grouped in a series of compressed ZIP files, organised by class. Since each recording is 10 minutes of 25 audio channels at 24-bit/48kHz resolution, EigenScape contains almost 140 GB of data in total. As this could potentially be taxing on disk space and take an extended period of time to download on

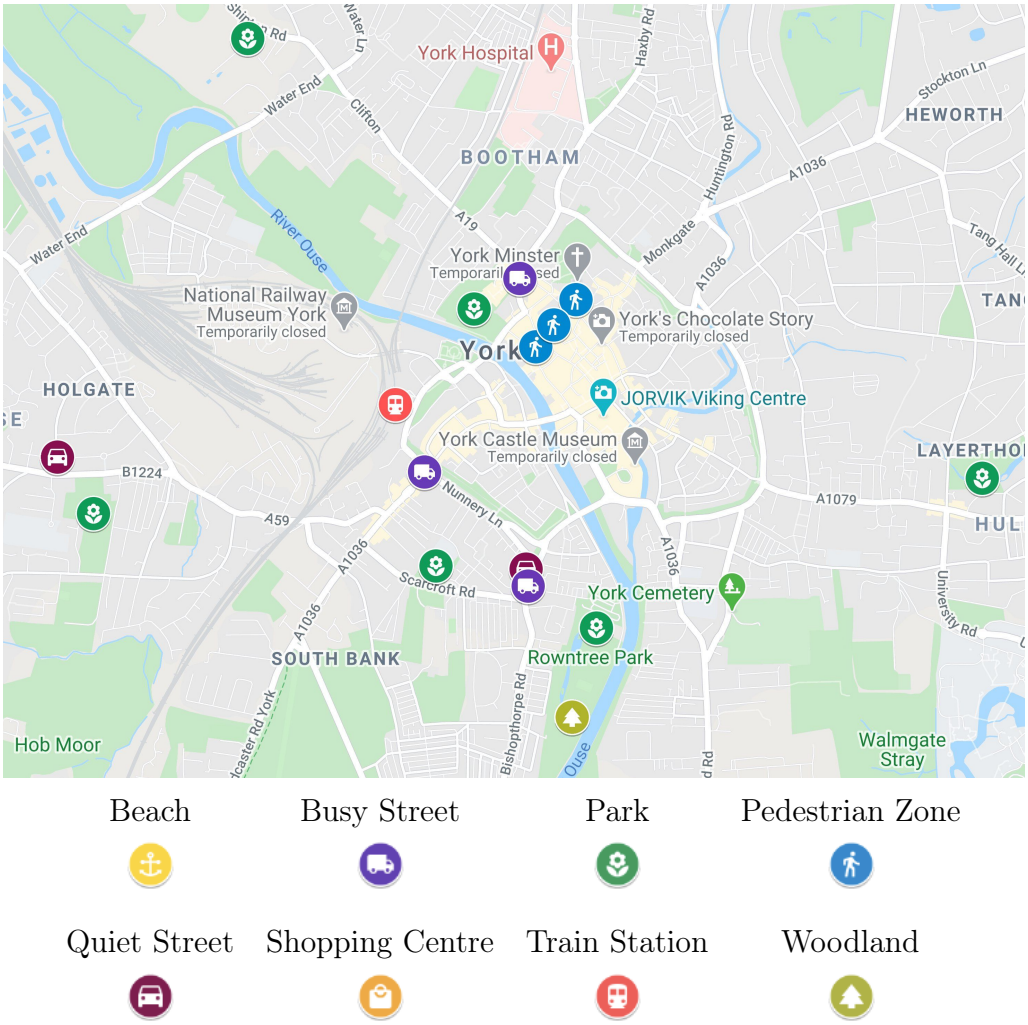


Figure 5.4: EigenScape recording locations in York.

slower internet connections, a second version of the dataset was compiled for easier access. The cut-down version contains all the full-length recordings, but limited to the FOA components (the first four tracks) and using the free lossless audio codec (FLAC) compression format. This second version of EigenScape is only 12.6 GB of data, but still allows for spatial audio analysis and reproduction, albeit in reduced spatial resolution. FLAC is also the UK Data Service recommended audio data format [185]. The raw 32-channel microphone data has not been released.

Figure 5.5: Fourfold cross-validation setup in EigenScape.

5.4 Cross-Validation

Although EigenScape contains 80 minutes of audio per class, it is still a small dataset by modern machine learning standards. Computer vision datasets in particular can contain many thousands of examples of each class [186], ensuring reduced variance and less likelihood of a model overfitting. Smaller datasets may not capture a broad enough range of statistical properties for each class, so outliers have an outsized influence and models have a greater tendency to overfit.

One method that can be employed in increasing the validity of results from a model trained on a smaller data sample is cross-validation [121]. With cross-validation, the model is trained and tested multiple times on different subsets of the available data. The data is partitioned such that each data point is included in the test set only once. With very small datasets, sometimes leave-one-out cross-validation is used. This is a special case where the test set consists of only a single datapoint each time, with the training set being the entirety of the rest of the data. This would not be appropriate to use with EigenScape, as it would violate the requirement that clips from the same longer recordings should not cross over between the training and test sets. Instead, k-fold cross-validation, in which the full dataset is partitioned into k equally-sized subsets, was used. $k-1$ of the subsets are used for training, with the remaining subset used for testing. Training and testing is repeated k times with each subset used for testing in turn. The mean performance of the model across all folds is calculated as the final result.

Ordinarily, partitioning of the available data is done entirely at random. How-

ever, to maintain the separation of longer recordings between sets in EigenScape, entire 10-minute clips should be assigned to either the training or testing set. These can subsequently be partitioned into smaller segments as required. EigenScape was specifically designed to allow two, four, or eightfold cross-validation in this manner. Figure 5.5 shows an example of 4-fold cross validation, as could be done using the EigenScape data. The choice to include eight examples of each scene class was made to facilitate these cross-validation setups whilst putting a practical and reasonable limit on the amount of data to be collected by a single researcher. Whilst inclusion of more data would increase the diversity of the dataset, this would cause the amount of work involved to increase significantly. The dataset was, however, recorded across six different cities and large towns and at many locations between. This offers a range of diverse locations comparable to the "six large European cities" across which recordings were made for the TUT 2018 dataset [163].

A set of software tools were published alongside EigenScape to facilitate easy creation of test setups including partitioning of the data into folds, and segmentation of the longer audio clips [187].

5.5 Summary

This chapter has detailed the collection of the EigenScape database of spatial acoustic scenes. A review of already-existing databases was presented, and whilst none of these databases were completely appropriate for the work presented in this thesis, elements from them have influenced the direction of EigenScape. The resultant new database is essentially a fusion between the approach taken with the TUT/TAU databases [125, 161, 162, 163, 164] and the spatial audio recordings of Stevens *et al.* [20, 45, 106], motivated by the soundscape approach.

A detailed specification of EigenScape was presented, including information on the recording equipment, location classes and specific locations shown on a map, together with details on the public release of the data online. Finally, information

on cross-validation was presented, including how EigenScape was planned to consider this, with a view to avoiding the 'album effect' inflation of results as reported by Lagrange *et al.* [144].

Now that this chapter has established the groundwork, the following chapters will detail the original research facilitated using the EigenScape data, beginning with an investigation into the characterisation of acoustic scenes by their spatial properties.

6 | Acoustic Scene Classification Using Spatial Features

Associated Publications

- M. C. Green and D. T. Murphy, "Acoustic Scene Classification Using Spatial Features,", in *Detection and Classification of Acoustic Scenes and Events Workshop*, Munich, Germany, 2017

- M. C. Green and D. T. Murphy, "EigenScape: A Database of Spatial Acoustic Scene Recordings", *Applied Sciences* vol. 7, no. 11:1024, November 2017, doi: 10.3390/app7111204

- M. C. Green, D. T. Murphy, S. Adavanne, T. Virtanen, "Acoustic Scene Classification Using Higher-Order Ambisonic Features", in *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2019

  – Contributions of co-authors limited to development of CNN classifiers. All analysis performed by the author of this thesis.

Contributions

- Results indicating that acoustic environments can be characterised by their spatial properties independently of their spectral properties.
- Evidence that similarity of scenes' spatial properties do not always coincide with similarity in their spectral properties.
- Several classification models developed using features derived from Ambisonic audio.

6.1 Introduction

EigenScape is the first database of its breadth to have been recorded in high-order Ambisonic format, and was specifically designed to facilitate investigation into spatial features for acoustic scene classification, as a precursor to more detailed study of acoustic environment properties. This chapter will present the results of several classification systems trained using features extracted from the EigenScape data. Initially, features were derived from only the first-order Ambisonic channels, with classification accuracies using a GMM trained on these features compared to those achieved using the standard MFCC features. Later, spatial features were derived from the complete fourth-order Ambisonic information available in the dataset, with these results compared to those achieved using the first-order features only. These tests were repeated in an additional study that was a collaboration between the author of this thesis and researchers at TUT [188]. The role of the TUT researchers was the creation and training of CNN classifiers in place of GMMs, whilst the analysis of the results was conducted by the author of this thesis.

6.2 Spectral and First-Order Ambisonic Features

6.2.1 MFCC Implementation

To establish a baseline classification performance against which spatial features can be compared, the first set of features to be extracted from EigenScape are MFCCs. The MFCC function from the librosa library [116], is used to extract these features, retaining the first MFCC as in [125], from the 0th-order omnidirectional channel of each EigenScape recording. The standard behaviour of the librosa MFCC function is to resample the input audio to 22.05 kHz, followed by extraction of MFCCs from frames of 2048 samples with 25 % overlap across 20 frequency bands. All of these parameters are retained apart from the resampling frequency, which is altered to

24 kHz. This reflects a 50% reduction in the 48 kHz sampling frequency used for EigenScape, whereas the librosa standard is a 50% reduction from CD-quality audio at 44.1 kHz. The MFCCs therefore represent frequencies up to 12 kHz.

6.2.2 Directional Audio Coding

For the first stage of the spatial ASC investigation, it was decided to use features derived from only the FOA channels, with more complex higher-order features investigated as a second stage. The directional audio coding (DirAC) technique [189, 190, 191] is used to derive coarsened spatial features from the raw audio data. Developed based on the spatial impulse response rendering (SIRR) algorithm [17, 192], DirAC was intended to facilitate reconstruction of a 3D sound field from monaural audio based upon extracted spatial parameters. In, for instance, a teleconferencing system, this would allow only a single channel of audio to be transmitted with the spatialisation defined by metadata, a much reduced data rate compared to transmitting a complete four-channel FOA signal.

To limit the metadata to a manageable data rate, several assumptions regarding human perception were made. The most important of these is "humans decode at one time only single cues per each critical band from the ear canal signals" [190], evidence for which is presented in [193]. DirAC therefore makes use of perceptually-motivated features, and this makes a comparison with MFCCs especially interesting. The frequency-dependent perceptual basis of DirAC provides an appropriate coarsening of the data for machine listening purposes. This, combined with the relative simplicity of the method, as detailed below, makes DirAC an ideal candidate for FOA spatial features.

Before feature extraction, the EigenScape audio clips are resampled to 24 kHz, limiting the maximum recorded frequency to 12 kHz in parity with the process used to extract MFCCs. This resampled audio is then filtered into 20 mel-spaced frequency bands calculated to match exactly the 20 frequency bands used by the librosa MFCC function, using a bank of finite impulse response (FIR) filters. Intensity

vectors $\mathbf{I}$ are calculated from this filtered Ambisonic signal as:

$$\mathbf{I} = \mathbf{PU} \tag{6.1}$$

where $\mathbf{P}$ contains the 20 mel-filtered versions of the zeroth-order Ambisonic channel equivalent to the pressure component of the sound, and $\mathbf{U}$ is a three-dimensional matrix containing the filtered versions of the three first-order channels, representing the velocity component. The resultant matrix $\mathbf{I}$ contains instantaneous intensity vector estimates for each frequency band. DOA estimate vectors run in the opposite direction to these intensity vectors. Azimuth ϕ and elevation θ angles can be derived from this trigonometrically:

$$\phi = \arctan\left(\frac{\mathbf{I}_3}{\mathbf{I}_1}\right) \tag{6.2}$$

$$\theta = \arccos\left(\frac{-\mathbf{I}_2}{||\mathbf{I}||}\right) \tag{6.3}$$

where $\mathbf{I}_1$, $\mathbf{I}_2$ and $\mathbf{I}_3$ are the filtered first-order channel matrices contained within $\mathbf{I}$. Values for diffuseness ψ are calculated as [190]:

$$\psi = 1 - \frac{||\mathbf{I}||}{c\mathbf{E}} \tag{6.4}$$

where $\mathbf{E}$ is the instantaneous energy density [190]:

$$\mathbf{E} = \frac{\rho}{2}\left(\frac{\mathbf{P}^2}{Z_0^2} + ||\mathbf{U}||^2\right) \tag{6.5}$$

and Z_0 and ρ are the acoustic impedance and mean density of air, as explored in Section 2.2.1.

Features are extracted based on rolling mean values for $\mathbf{P}$ and $\mathbf{U}$ calculated in frames of 2048 samples, with 25% frame overlap, again matching the MFCCs. This gives a set of DirAC parameter estimates for every 64 ms. The output from each frame of the DirAC analysis is a set of azimuth, elevation and diffuseness estimates

for each of the 20 frequency bands, resulting in a 60-dimensional feature vector of FOA spatial information.

6.3 Higher-Order Ambisonic Features

The second stage of this study explores whether features utilising information in the higher-order Ambisonic (HOA) channels might improve upon the classification performance of systems using the FOA features. Given the inherent trade-off in machine learning between discrimination and generalisation, it is not clear that this will be the case. The increased spatial precision of HOA could mean that the extracted features represent details more specific to individual scenes than those general to the scene class.

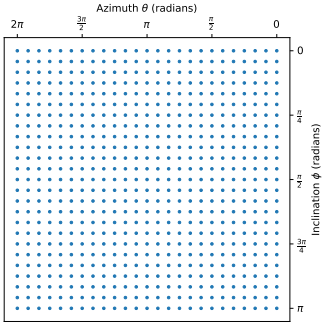
One of the aims in the selection of HOA features is some equivalence with the FOA features in order to make comparisons between the two more meaningful. In practise, this means finding methods for estimating frequency-dependent DOA and diffuseness incorporating the HOA channels. The most immediately obvious candidate for this task is the HOA extension to DirAC, which calculates parameters in sub-sectors of the complete spherical space [194]. This technique, however, outputs separate DOA and diffuseness estimates for each individual sector, so these values are not directly comparable with those obtained from FOA DirAC.

Instead, DOA estimates are derived from a steered-response power (SRP) map created using spherical harmonic beamforming. SRP is a technique of sampling the sound field power in all directions, analogous to the "instantaneously-rotating directional microphone" discussed in Section 2.4. Diffuseness features are extracted using the covariance matrix eigenvalue diffuseness estimation (COMEDIE) algorithm [195]. The next two sections will detail these techniques.

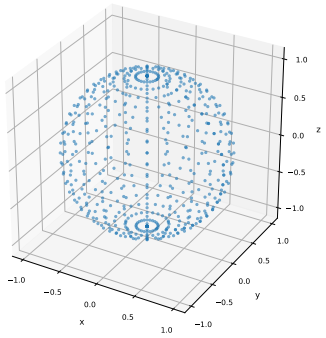
6.3.1 Distributing Sample Points on a Sphere

In order to obtain a representative sample of the sound power as it varies across space when creating an SRP map, it is necessary to define a set of sampling points. Despite the spherical geometry of the search space, choosing an appropriate scheme for this task is not as simple as it might seem. Figure 6.1 shows a selection of sample distributions in a flat space and wrapped to spheres. Perhaps the most initially obvious choice, a regular grid, will in fact cause the 'poles' to be sampled with much higher resolution than equatorial areas when wrapped to the sphere. This increased detail would cause a great deal of computation time to be spent processing largely redundant data, as sound is much less likely to originate near polar regions in most scenes. The increased density of sample points at the poles will also bias the sound power readings towards them - it will look like sources near the poles contain more power simply as they are sampled more densely. This problem can be solved using an appropriate quadrature weighting [196], but if this step can be avoided, computing power can be saved.

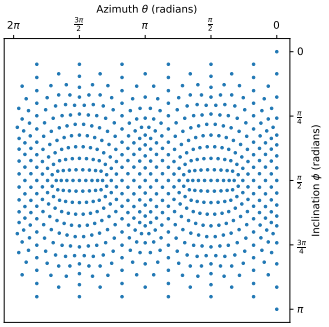
It is a fundamental law of geometry that there are a finite number of ways to evenly distribute points on the surface of a sphere, corresponding to the vertices of the platonic solids [197]. The regular dodecahedron, with 20 vertices, provides the greatest number of points. Clearly, 20 points will not provide sufficient resolution for a representative SRP map. A common way to circumvent this theoretical limit is to interpolate between the vertices of a regular icosahedron, equivalent to the central points on the faces of the dodecahedron [146, 147, 196]. Figure 6.1(c) shows the pattern produced at an interpolation factor of 8. When projected to the surface of a sphere (Figure 6.1(d)), these points provide far more uniform sampling than the regular grid, with only slight irregularities around the original icosahedron vertices. The main disadvantage of this approach is that the number of available points P are at fixed intervals depending on the interpolation factor i:



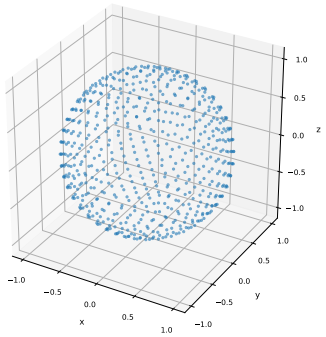
(a) Regular (625-point)



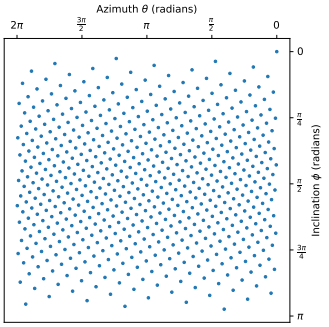
(b) Regular Grid on Sphere



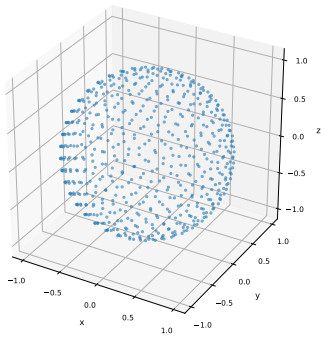
(c) Geodesic (642-point)



(d) Geodesic Grid on Sphere



(e) Fibonacci (600-point)



(f) Fibonacci Grid on Sphere

Figure 6.1: Various schemes for sampling spherical co-ordinates.

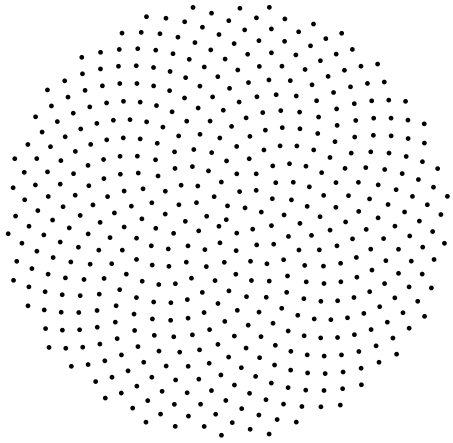


Figure 6.2: 500-point 'sunflower' distribution generated using Equation 6.7.

$$P = 10i^2 + 2 \tag{6.6}$$

The pattern shown in Figures 6.1(c) and 6.1(d) has 642 points, whilst the nearest levels of interpolation above and below this give over 100 points greater and fewer respectively. This limits options when aiming for a particular angular resolution. The popular Lebedev quadratures [198] are also limited to certain values, though it should be noted that the distance between these is generally much smaller than using icosahedron interpolation.

Another alternative is the Fibonacci sphere [199]. This takes the distribution proposed by Vogel to simulate the arrangement of sunflower seeds [200] and spreads the pattern across the spherical surface. The original two-dimensional formulation is defined in polar co-ordinates as follows:

$$\begin{aligned} \phi_p &= \frac{2\pi p}{\varphi} \\ r_p &= \sqrt{p} \end{aligned} \tag{6.7}$$

where φ is the golden ratio $\frac{1+\sqrt{5}}{2}$ and p is an integer index for each point. Figure 6.2 shows a spiral pattern generated using Equation 6.7. For spherical sampling, the expression to calculate radial distance r is replaced by the following, which determines elevation:

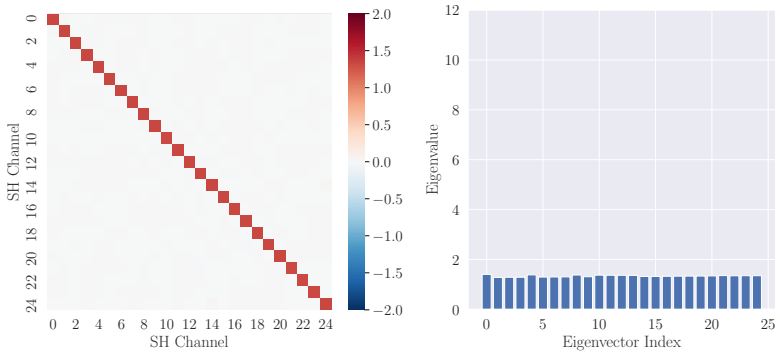
$$\theta_p = \cos^{-1}\left(\frac{-2p}{P-1}\right) \tag{6.8}$$

Figures 6.1(e) and 6.1(f) show a 600-point Fibonacci spiral. This sampling scheme has the advantage of having a very uniform distribution except for slight under-sampling at the poles [196], and, unlike icosahedron interpolation or the Lebedev quadratures, allows for any arbitrary number of points, which is very beneficial when defining the resolution of the search grid. For these reasons, in this study the SRP map was created using a 300-point Fibonacci spherical sampling scheme. The other distributions detailed here were not tested as part of this work. It was considered that the discrepancies in point density using regular sampling would likely affect DOA estimates in an adverse manner, whilst the aforementioned control offered over the specific number of sampling points made the Fibonacci spiral preferred over the geodesic grid.

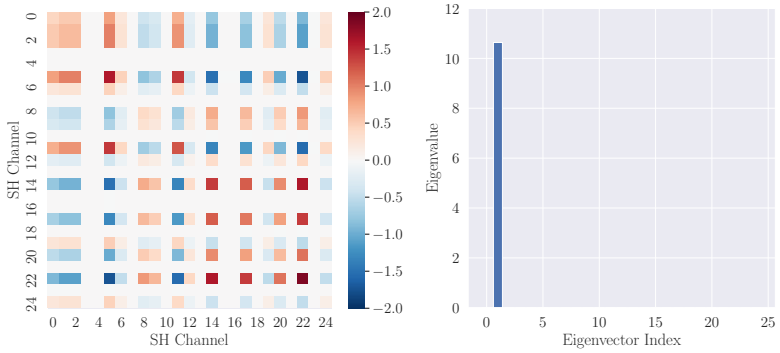
The SRP map was created using CroPaC beams (see Section 2.4.3), and weights were applied to the k dimension of the beamformer output (Equation 2.26) to obtain azimuth and elevation estimates from the same 20 mel-spaced frequency bands as the FOA features. The maximum-amplitude point of the power map from each band was taken at the DOA estimate.

6.3.2 COMEDIE Diffuseness

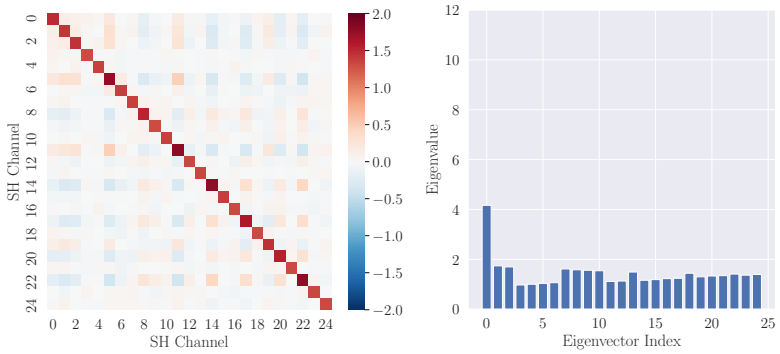
Diffuseness in the HOA channels was estimated using the covariance matrix eigenvalue diffuseness estimation (COMEDIE) algorithm [195]. The method was developed by Epain and Jin following an investigation of the covariance matrices of the spherical harmonic signals for several synthesised sound fields. It was found that in a perfectly diffuse sound field, all eigenvalues of the matrix are equal, whereas in a sound field featuring only a single plane wave, only one eigenvalue is nonzero. More complex sound fields yield more irregular spectra that lie between these two extremes. Figure 6.3 shows covariance matrices and their respective eigenvalues for simulated diffuse and single plane wave sound fields, and a combination of the two.



(a) Diffuse field



(b) Plane wave



(c) Plane wave plus diffuse background

Figure 6.3: Spherical harmonic signal covariance matrices and their eigenvalues for various synthesised sound fields (after [195]).

COMEDIE makes use of this phenomenon to calculate figures for diffuseness based on the mean deviation γ of eigenvalues ν from their mean value $\bar{\nu}$ [195]:

$$\gamma = \frac{1}{\bar{\nu}} \sum_{i=1}^{(N+1)^2} |\nu_i - \bar{\nu}| \tag{6.9}$$

This is used to calculate diffuseness by:

$$\psi = 1 - \frac{\gamma}{\gamma_0} \tag{6.10}$$

where γ_0 is the value of γ for the perfect single plane-wave soundfield, derived in [195] as:

$$\gamma_0 = 2\left[(N+1)^2 - 1\right] \tag{6.11}$$

Epain and Jin also introduce the concept of diffuseness profiles. In testing the COMEDIE diffuseness calculation, it was observed that, owing to the varying spatial resolution across SH orders, diffuseness estimates for the same sound field could vary depending on the orders used. For sound fields where one source dominates, or that are very diffuse, diffuseness estimates across orders tend to be similar, resulting in a flat profile. On the other hand, diffuseness values that decrease with increasing SH order are an indication of a sound field consisting of several uncorrelated sources incident from different directions. This scenario tends to read as more diffuse at lower orders, owing to the reduced spatial resolution. There is, therefore, useful information regarding the true nature of the soundfield to be obtained by calculating diffuseness across all available orders. Using diffuseness profiles helps to disambiguate "whether diffuseness arises from the presence of a diffuse noise background or from the presence of multiple yet countable uncorrelated sources distributed in space" [195].

Diffuseness profiles including values for all four available orders were calculated as part of the set of HOA spatial features. As with the FOA DirAC diffuseness values, these were calculated across 20 mel-bands, resulting in 80 diffuseness values

in total. The complete HOA feature vector includes 20-dimensional vectors for both the azimuth and elevation estimates with this 80-dimensional diffuseness vector, giving 120 dimensions in total.

6.4 Method

The EigenScape database is split into four folds for cross-validation following the scheme outlined in Section 5.4. Features extracted from the training audio are used to train a bank of GMMs (one for each scene class) using the E-M algorithm (see Section 4.3.3). Each GMM used ten Gaussian components, matching those used in the original DCASE basline classifier [124]. It was decided to initially use these simple GMM models in order to keep the focus on the broad discriminative information contained within the features rather than the power of the model. It is reasoned that if a GMM could produce good classification accuracy using the spatial features, this would provide strong evidence for the effectiveness of the features themselves.

Test recordings are cut into 30s segments, giving 40 segments in total per class, per fold. Features from the clips are presented to the GMM bank, with each GMM outputting a probability score indicating the estimated likelihood that new frames came from the same class as the training frames for that model. Scores are summed across frames from the entire 30s segment, with each segment classified according to the model giving the highest total probability score across all frames. The performance metric used is a simple percentage accuracy (e.g. proportion of correctly-assigned labels), with the mean value taken across all four folds.

To act as a baseline level against which performance with the new spatial features can be compared, a separate set of GMMs are trained using the MFCC features. Separate GMM banks are also trained using individual FOA features and combinations thereof, with one set trained using a concatenation of both the MFCCs and the complete set of FOA features, and another trained on the MFCCs and diffuseness

features alone. The approach with the HOA features mirrored that taken with the FOA features.

FOA and HOA classifiers are kept completely separate, with no classifiers trained with combinations of FOA and HOA features. This is because it was felt that since HOA features require the use of information contained in the FOA channels, mixing the two sets of features would in a sense represent duplication of data. This is most evident when considering the COMEDIE diffuseness profiles, which are explicitly multi-order features - adding FOA DirAC diffuseness to these is unlikely to add any extra discriminative information. On the other hand, the concept of 'diffuseness profiles' raises the possibility of 'DOA profiles' compiled in a similar manner. However, whilst calculating diffuseness across different Ambisonic orders can yield additional information about the scene, as discussed in Section 6.3.2, it is less clear what would be gained from including multi-order DOA estimates. Whilst it is conceivable that the first-order DOA estimations obtained using DirAC could contain information complementary to higher-order SRP-derived estimates, lower-order estimates calculated using SRP are likely to be simply noisier versions of the higher-order estimates.

6.5 Results and Discussion

6.5.1 FOA

Figure 6.4 shows the mean classification accuracies attained across all scenes using various subsets of the FOA features. The complete set of DirAC features yield an accuracy of 64%. MFCC features, on the other hand, yield a 58% accuracy, a figure in line with MFCC-GMM performance reported in the literature [10, 144]. The only spatial feature to perform worse than MFCCs are the azimuth estimates, giving a 43% average accuracy. Performance using elevation features is very similar to that achieved using MFCCs. Using diffuseness features improves on this slightly.

The poor performance when using azimuth features is possibly due to the fact that azimuth estimates are affected by the microphone array's orientation relative to typical elements of the recorded scene. If, for instance, the scene in question is the BusyStreet class, in one recording the road could be passing directly in front of the array, and in another the road might be on the left hand side. This was not controlled for at the recording stage and the recorded soundfields were not rotated *post hoc*. Elevation and diffuseness estimates should, by contrast, be invariant to horizontal array rotations.

The complete DirAC features outperformed the MFCC baseline by a larger margin than either the elevation or diffuseness features alone, despite including azimuth features that were shown to perform poorly. An additional bank of GMMs trained excluding the azimuth features achieved 69% overall accuracy, which is the best performance using FOA spatial features. The 7% standard deviation indicates that the features extracted from the dataset are fairly consistent across folds. The accuracies achieved using these elevation/diffuseness (E/D) features are shown broken down by scene class in Figure 6.5. Every class is identified with a mean accuracy above 60%, except Beach, which has a mean accuracy of only 8%. The poor classification of the Beach scenes is such an outlier that discounting these particular results would raise the overall mean accuracy by 9%. One reason for this could be the dominant ocean wave sound at Beach locations. This is by nature an enveloping sound that is both broadband and diffuse, and could yield indistinct features difficult to separate from other scenes.

Three of the classes were identified with averages above 80%, with one, PedestrianZone, classified at 97% accuracy with a standard deviation of only 4%. The standard deviation values represent the fluctuation in performance between folds, giving some indication of the variability of data within a given class. As such, the low standard deviation of the PedestrianZone accuracy implies that the PedestrianZone recordings have very similar sonic characteristics, yielding very consistent features. By this metric, QuietStreet, ShoppingCentre and Woodland show the most

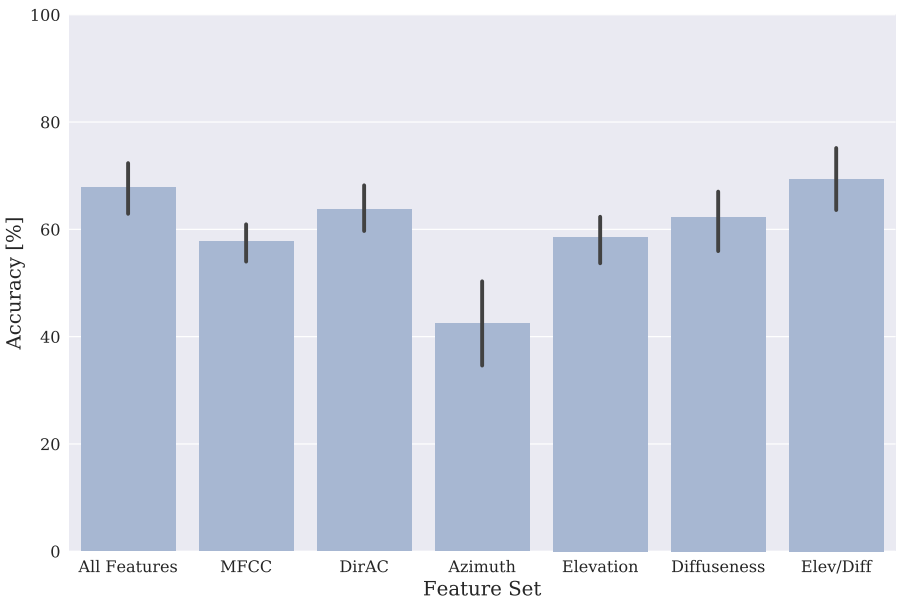


Figure 6.4: Mean and standard deviation classification accuracies for different FOA feature subsets.

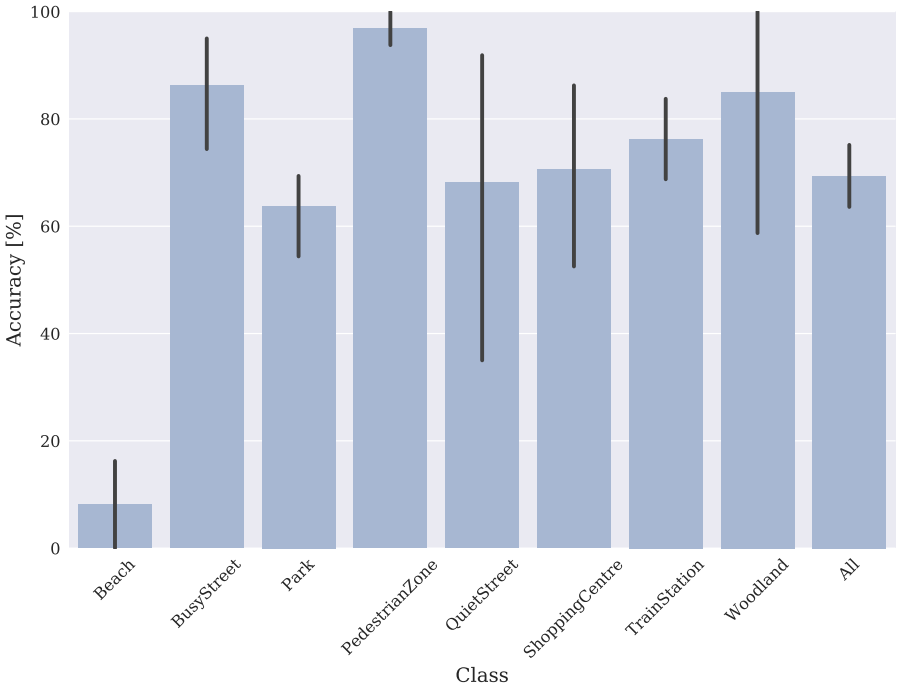
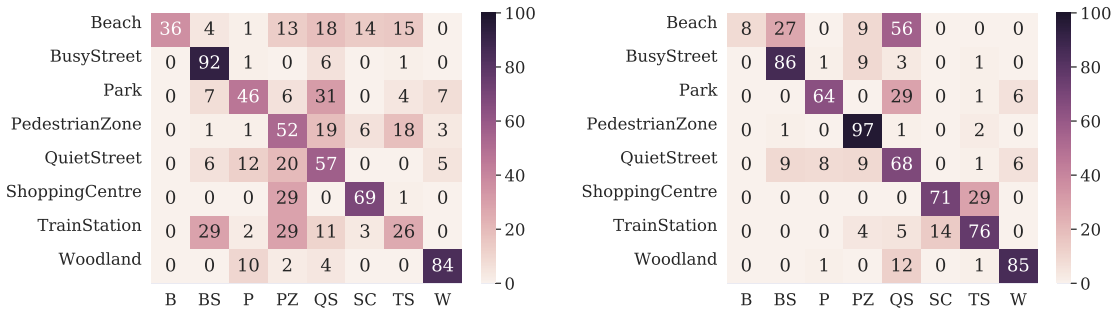


Figure 6.5: Mean and standard deviation classification accuracies for each scene class using the elevation/diffuseness feature combination.



(a) MFCC Classifier

(b) E/D Classifier

Figure 6.6: Confusion matrices for GMM bank classifiers trained using MFCC and FOA elevation/diffuseness features. Figures indicate classification percentages across all folds.

variability, with the remainder of the scenes being moderately variable.

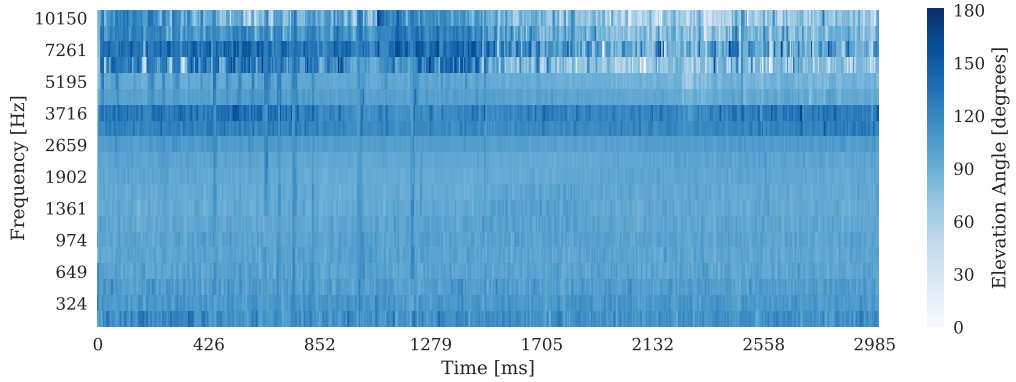
Further insight can be gained by studying confusion matrices shown in Figure 6.6. The rows of these matrices indicate the true classes, and the columns show the labels returned by the GMM bank classifier. It is immediately clear that the E/D matrix has a much more prominent leading diagonal, whereas confusion is more widespread in the MFCC classifier. This indicates that the spatial E/D features outperform the spectral MFCC features in the majority of cases. The only scene class where the performance of the MFCC classifier significantly improves on the E/D classifier is Beach, though the 36% accuracy is hardly satisfactory.

It is especially interesting to look at the specific misclassifications between scenes. Whilst the misclassifications of Beach scenes by the MFCC classifier are spread fairly evenly across several other classes, the E/D classifier most commonly misclassifies Beach scenes as either BusyStreet or QuietStreet. It is possible that this is due to broadband noise from passing vehicles creating patterns in the spatial features similar to those of ocean waves. Corroboration for this interpretation is visible in Figure 6.7, which shows elevation data extracted from Beach, QuietStreet and TrainStation recordings. It can be seen that the Beach and QuietStreet scenes have large areas where elevation estimates are broadly uniform at around 90° across both

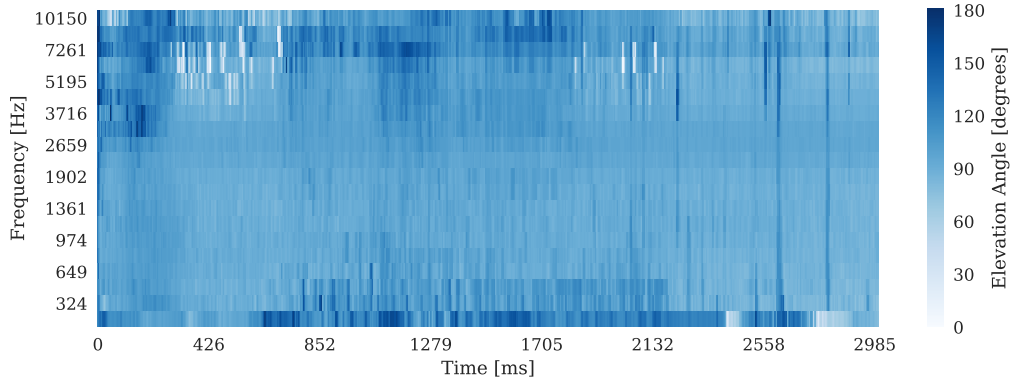
time and frequency, whereas the elevation data from the TrainStation clip is far more erratic.

It is also interesting to consider cases where the E/D classifier has considerably outperformed the MFCC classifier. PedestrianZone, classified 97% accurately by the E/D classifier, but only 52% accurately by the MFCC classifier, is particularly noteworthy. These results suggest that the spatial properties of pedestrian zones are much more unique to them than their spectral properties, which seem to have a degree of overlap with the spectral properties of both quiet streets and train stations. This observation can be explored further by considering those classes which are significantly confused by both classifiers. Sometimes specific misclassifications coincide, such as with the Park class, which is most commonly misclassified as QuietStreet by both spatial and spectral classifiers. This is possibly due to both of these locations being open areas of relative quiet in the midst of urban areas, mostly consisting of natural sound with occasional human sound, and characterised by the presence of a low-level background urban 'hum' keynote notably absent from the Woodland recordings.

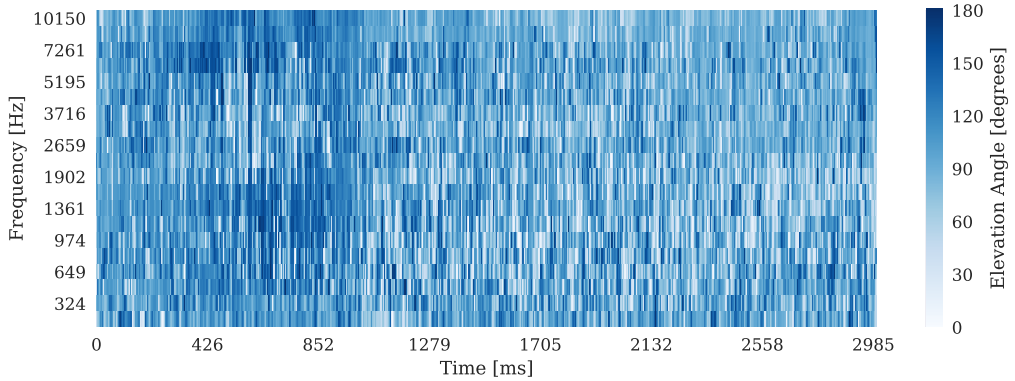
Perhaps more interesting are the instances where specific misclassifications do not correspond. The ShoppingCentre class, for instance, is most commonly misclassified by the MFCC classifier as a PedestrianZone. This is most likely caused by the prominent human sounds present in both of these location classes. The E/D classifier, on the other hand, most commonly misclassifies ShoppingCentre clips as TrainStation, and does not misclassify any clips as PedestrianZone. This could be due to the similarity in acoustics at these locations. Both classes are often large indoor spaces with lots of reverberant surfaces, which could generate a common signature in terms of elevation and diffuseness values. The divergences in these misclassification patterns lead to the interesting observation that the *spectral* similarity and *spatial* similarity of scenes do not necessarily coincide. This is a key finding that provides an indication that the spatial properties of acoustic scenes can be used for classification in a way that is complementary to their spectral properties.



(a) Beach



(b) QuietStreet



(c) TrainStation

Figure 6.7: Elevation estimates extracted from 30s segments of Beach, QuietStreet and TrainStation recordings.

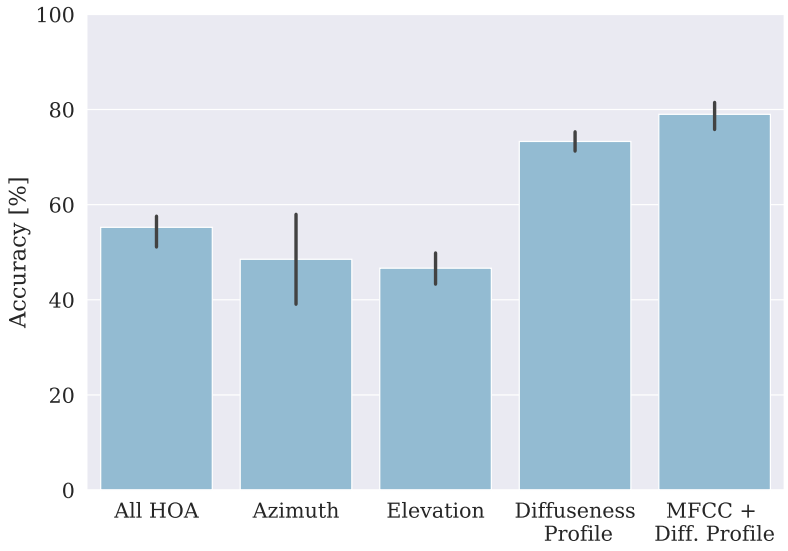


Figure 6.8: Mean and standard deviation classification accuracies for HOA features.

These ASC performances using FOA features are important for a number of reasons, most significantly that they are the first results to conclusively show that the spatial properties of sound scenes can be used as discriminative features in an acoustic scene classification system. They also provide validation for the EigenScape dataset in terms of its suitability for ASC and soundscape work. The good, but not perfect, overall accuracy shows that the database provides a satisfactory variety of recordings and has successfully overcome the potential issue of the 'album effect' [125, 145] described in Section 4.4.1. These results provide a good benchmark against which the performance of the HOA features can be compared.

6.5.2 HOA

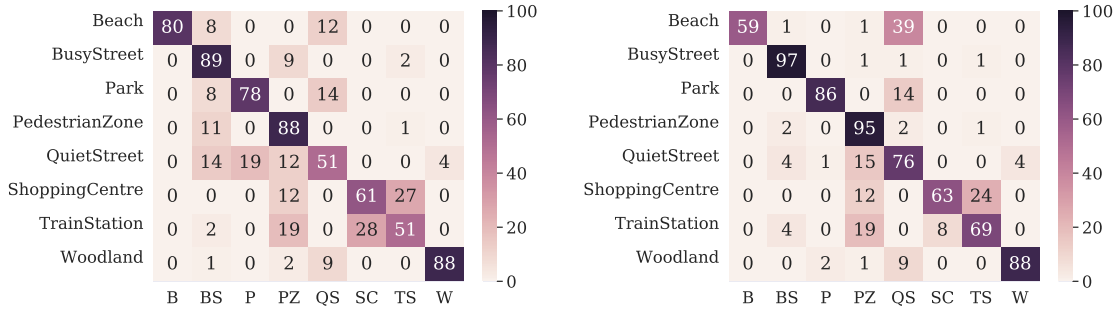
Figure 6.8 shows the mean and standard deviation classification accuracies across all folds for different subsets of the HOA features. Comparing these results to those from the FOA features shown in Figure 6.4, it can be seen that performance when using the complete set of HOA features is somewhat reduced compared to the DirAC features, down from 64% to 55%. Breaking this down into feature subsets, the azimuth performance has improved slightly from 43% up to 48%. This is, surprisingly,

a slightly better overall performance than was achieved using the elevation features in HOA, which have a markedly worse accuracy, at 47%, than their FOA counterparts at 59%. This is a reversal of the pattern seen using FOA direction-of-arrival features. Given that the standard deviation of the azimuth accuracies is somewhat larger than that of the elevation accuracies, this is not a conclusive indication of HOA azimuth features outperforming elevation features.

Diffuseness profiles, on the other hand, give a performance increase of around 10% relative to DirAC diffuseness, with the low standard deviation also indicating consistently good results across folds. This provides evidence that there is information in the HOA channels that can be more discriminative than is available in FOA, and supports the assertion in [195] that diffuseness profiles can disambiguate between sound fields that could look very similar at lower orders.

The large reduction in accuracy when using the HOA direction-of-arrival features relative to FOA is surprising. One reason for this might be that the SRP method used to calculate the HOA azimuth and elevation estimates uses a quantised grid of directions (as specified in Section 6.3.1), whereas the intensity-vector technique used in DirAC means that the FOA estimates are not limited in this way. Also possible, but perhaps less likely in light of the improved performance using diffuseness profiles, is that the increased spatial resolution in the HOA data could be adversely affecting the generalisability of the features between scenes of the same class. It is also possible that since the azimuth and elevation estimates derived from the SRP maps essentially assume a dominant point source in each frequency band, they do not capture the complete character of these real-world sound fields particularly well. The extremely narrow beams achieved by CroPaC would exacerbate all of these issues.

Figure 6.9(a) shows the confusion matrix for the classifier trained on diffuseness profile features. Perhaps the most striking aspect of these results is the drastic improvement in the classification of Beach scenes relative to the FOA E/D classifier, up from 8% to 80%. This is further evidence of the greatly increased discriminative



(a) Diffuseness Profile Features

(b) MFCC + Diff. Profile Features

Figure 6.9: Confusion matrices for GMM bank classifiers trained using HOA diffuseness profile features, and those same features together with MFCCs. Figures indicate classification percentages across all folds.

power of the diffuseness profile features, which disambiguate between Beach and Street scenes using the additional information available in HOA.

Following the finding with the FOA results that the discriminative information contained in spatial and spectral information could be complementary, it was decided to train an additional GMM bank using a combination of the MFCC and diffuseness profile features. This classifier achieved an average accuracy of 79%, and the result is shown for comparison in Figure 6.8. This is the best accuracy achieved using GMMs. Figure 6.9(b) shows the confusion matrix for this classifier. Relative to the GMMs trained using diffuseness profiles only, there are improvements across all classes apart from Beach, which is reduced to 59% accuracy. Whilst still an improvement on the very poor FOA Beach classification, this result shows that inclusion of MFCC features as well as the diffuseness features does not necessarily result in improvements in performance in every case.

It should also be noted that the accuracies for the ShoppingCentre and TrainStation classes are slightly reduced relative to the FOA E/D classifier, both with and without the inclusion of the MFCC features. Looking at the specific misclassifications in both matrices shown in Figure 6.9 it can be seen that a great many mirror those observed when using the FOA E/D features (Figure 6.6(b)), indicating that

the characterisation of the scenes remains fairly consistent across features despite the different extraction methods used. These include the confusion of QuietStreet with BusyStreet, Park and PedestrianZone, and the mutual confusion between ShoppingCentre and TrainStation. The TrainStation class is, however, misclassified as PedestrianZone 19% of the time in HOA, whereas in FOA E/D this figure was only 4%. This suggests that the DirAC elevation information was key for the FOA classifier in disambiguating these two scenes. It might improve performance even more to have included these alongside the HOA features.

6.5.3 CNN Classification

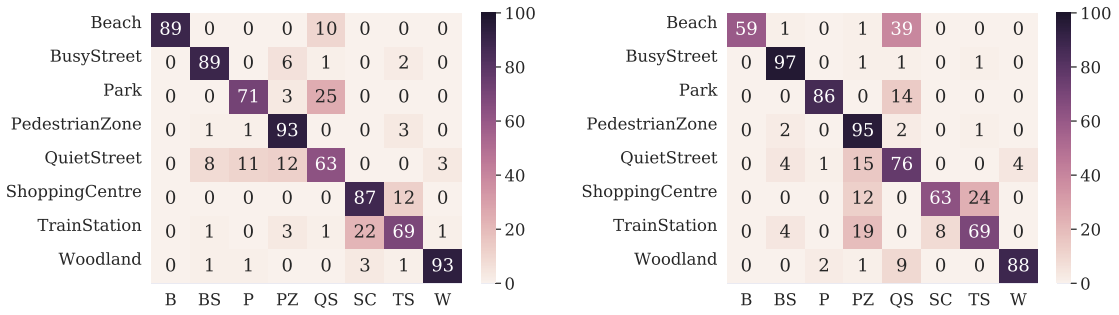
Subsequent to the work with GMM classifiers, the investigations described so far in this chapter were repeated using a CNN [188]. The CNN architecture is tuned as a set of hyperparameters including the number of layers $\in \{1, 2, 3, 4, 5\}$, the number of filters per layer $\in \{16, 32, 64, 128, 256\}$ and dropout [201] in all layers $\in \{0, 0.15, 0.3, 0.5\}$. It was found that the optimum number of filters for all features was 64 per layer, with an optimal dropout of 15% for all layers. Most feature sets required 5 convolutional layers for optimal results, whereas subset combinations using diffuseness features only required 3 layers. This in and of itself suggests that the diffuseness features characterise the classes better than the other features, requiring less manipulation from the convolutions to achieve good results. The CNNs are trained for 500 epochs using the Adam optimisation algorithm [202], with training set to stop early if performance does not improve for 50 epochs.

Figure 6.10 shows a comparison of the mean classification accuracies using GMM and CNN classifiers using both FOA and HOA feature subsets. Note that the slight discrepancies in the GMM accuracies shown here compared to those in Figures 6.4 and 6.5 are due to these results being derived from another training run of the GMMs, which will have used different random initial values. It can clearly be seen that the CNN outperforms the GMM for all feature subsets across both FOA and HOA. This is hardly surprising given the relative complexity of the models

Figure 6.10: Mean classification accuracies for GMM and CNN classifiers using various subsets of FOA and HOA features.

and the proven record of CNN classifiers in DCASE ASC challenge tasks [203]. With the CNN, use of HOA channels increases accuracies in all subsets except elevation. For the feature subsets giving the very best results, however, the difference in performance is negligible, with the HOA elevation + diffuseness yielding only a 0.2% increase over their FOA counterparts and MFCC + diffuseness yielding an extra 0.8%. This suggests that the CNNs are able to derive additional discriminative information from the FOA features that the GMMs cannot.

In line with the GMM results, the very best accuracy is achieved using MFCC + diffuseness features. This gives extra weight to the observation that spectral and spatial features are complementary. In fact, where the CNN outperforms the GMM by a mean of approximately 14% across the other feature subsets, the CNN accuracy using MFCC + diffuseness features, 82%, is only 3% higher than the 79% accuracy achieved using the GMM. This suggests that this particular combination of features captures the scene information particularly well and can be used for effective classification without complex processing. Since a simpler 3-layer CNN was used to produce these results, it seems unlikely that a more complex classifier could improve on this performance to any great degree, perhaps suggesting that



(a) CNN (b) GMM

Figure 6.11: Confusion matrix for best-performing CNN classifier using MFCC + diffuseness profile features, with GMM matrix for comparison.

these results represent a saturation point for these features from the EigenScape database. It is possible, however, that if more data were available than better results could be achieved. The best-performing ASC system in the 2019 DCASE challenge used data augmentation to artificially create additional training data from spectral features, achieving a classification accuracy of 86.7% [204]. It would perhaps be an interesting next step to investigate incorporation of that technique with the combination of spatial and spectral features presented here.

Figure 6.11 shows the confusion matrix for the top-performing CNN classifier, with the GMM confusion matrix from Figure 6.9(b) repeated for easy comparison. The specific patterns of confusion are again broadly similar. The GMM outperforms the CNN considerably for the BusyStreet, Park, and QuietStreet classes and also slightly for the PedestrianZone class. In fact, if the Beach accuracy is discounted, the GMM would slightly outperform the CNN overall. Whereas the GMM bank trained solely on diffuseness profiles was able to classsify Beach scenes with 80% accuracy (see Figure 6.9(a)), adding the MFCC features causes a reduction to 59%. The fact that the CNN accuracy is not affected in this way is an indication of the ability of the CNN to disgard specific features that impede accuracy, where the GMM does not.

Future work could involve a closer investigation into the Beach scene features

to determine exactly which aspects of these features make them more difficult to classify relative to the other scenes. It would also be beneficial to develop rigorous statistical tests to confirm the significance of these results.

6.6 Summary

This chapter has detailed the process of investigating the potential of spatial features for acoustic scene classification using recordings from the EigenScape database. This had the dual purposes of ascertaining the viability of spatial features for characterising acoustic scenes, and validating the database in terms of providing a good variety of recordings whilst avoiding the album effect. DirAC techniques for parameterising first-order Ambisonic recordings were described as an initial method of spatial feature extraction. Following this, CroPaC SRP maps (including detail on spherical sampling schemes) and COMEDIE diffuseness profiles were described as methods for obtaining features utilising higher-order Ambisonics.

Results were first presented from a GMM classifier using FOA spatial features. Accuracies using the combined elevation and diffuseness features improved on those obtained using MFCCs. This provided the first evidence that spatial features are a viable way to characterise and classify acoustic scenes. Training the same classifier on HOA features yielded an improvement in performance when using diffuseness profiles, though not the direction-of-arrival features. The best-performing GMM classifier used a combination of the MFCCs and the diffuseness profiles. Results from a CNN classifier improved on those from the GMM for all feature subsets, though this was not always the case for individual class accuracies. The most pronounced improvements were when using FOA features, suggesting that the convolutional process is able to derive more information from these features than the GMM. Once again the best performance was achieved using the MFCC and diffuseness features combined, though in this case the CNN only outperformed the GMM by a small amount. This suggests that this combination of features captures the properties

of the scenes particularly well, and does not require sophisticated processing for effective classification.

In practical terms, these results indicate that the incorporation of spatial features together with spectral features has the potential to greatly enhance ASC performance. A potential area of future investigation is whether features capturing diffuseness properties can be derived from recordings made using equipment more portable than the Eigenmike array, perhaps from binaural microphones or the horizontal plane Ambisonic recordings achievable using affordable consumer-facing recorders such as the Zoom H2N [205].

Relating this back to the wider goals of the soundscape approach, the results presented here are the first confirmation that acoustic scenes can be categorised according to their spatial properties as well as their spectral properties. This is a crucial finding in building towards the understanding that is needed to produce ecologically valid lab reproductions of acoustic environments. If the spatial and spectral properties of a simulated scene could be shown to closely match those extracted from real examples, this similarity could be used as an indication of a level of validity. The results also indicate that building detailed statistical models of the spatial distributions of sources in different acoustic scene classes is a valid approach to take to provide parameters for acoustic environment simulation. In light of this, the next chapter will explore a system for individual source tracking in Ambisonic audio, building upon the SRP technique used for HOA direction-of-arrival estimates in this chapter.

7 | Sound Event Localisation and Trajectory Prediction

Associated Publications

- M. C. Green and D. T. Murphy, "Sound Source Localisation in Ambisonic Audio Using Peak Clustering" in *Detection and Classification of Acoustic Scenes and Events Workshop*, New York City, USA, 2019

Contributions

- Novel system for source tracking in Ambisonic audio using spherical harmonic beamforming in combination with DBSCAN clustering.
- Investigation of trade-off between maximum spherical harmonic order and performance, with results indicating second-order Ambisonics represents a good point of compromise.

7.1 Introduction

The work presented in the previous chapter showed that acoustic environments can be characterised by their spatial properties, and that these features can be used to accurately predict the class of acoustic environment in a recording. This was a crucial milestone in the project and was the first published evidence that spatial properties could be a fruitful area of investigation for acoustic environment analysis. As the models of sound scenes created by an ASC system contain high-level statistics regarding the spatial and spectral distributions of sounds in different environments, such models could be useful as part of a system designed to ecologically validate synthesised acoustic scenes, perhaps producing metrics on the similarity of synthesised scenes to their real-world counterparts.

The features used in the previous chapter provide general information about sounds in space at different frequency ranges, but there is no information on specific sound sources. For many applications of this research, it will be very useful to have a description of the soundscape, including semantic labels for sound sources, typical directions of arrival, and trajectories of motion. Other applications of a system providing this information include audio surveillance [206], tracking targets in the military [207], and robotics [208].

The work in this chapter therefore builds upon the previous chapter by investigating how spatial audio techniques can be used to go beyond classification of an entire scene and begin to detect and identify individual sound sources from the mixture.

7.2 System Specification

As discussed in Section 4.4.2, the process of detecting the onsets and offsets of sound sources in Ambisonic recordings, together with their Direction of Arrival (DOA) and trajectory of movement is known as Sound Event Localisation and Detection

(SELD). SELD is defined to include three stages [154]:

1. Identifying onset and offsets of active sound sources.

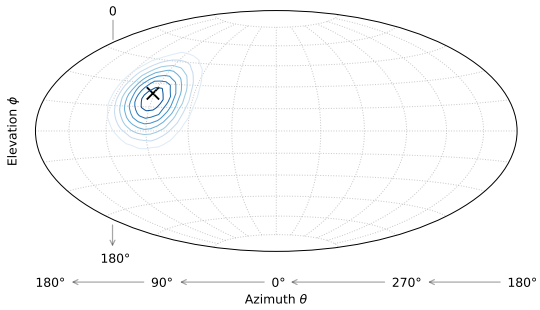
2. Direction-of-arrival estimation during each active frame.

3. Labelling the sound with descriptive text to identify the source.

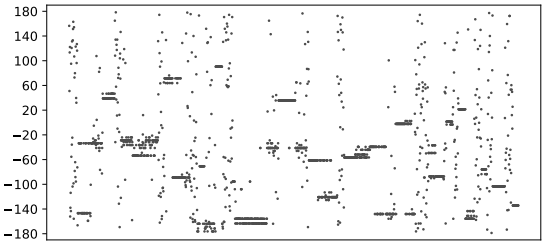
The potential of combining localisation and detection into a single model is that information on the spatial location of a sound could help disambiguate labelling, and vice versa. For instance, an ambiguous mechanical sound might be more easily labelled 'aeroplane' given localisation at high elevation. Alternatively an unambiguously-labelled aeroplane sound might assist localisation in the case of a strong ground reflection that would otherwise confuse the system. A model trained on this joint problem could theoretically learn these types of relationships. However, as outlined in Section 4.4.2, models trained on this joint problem, such as [154], tend not to perform quite so well as those designed for each problem separately, or, in the case of localisation, as well as traditional signal processing approaches. More recent work has indicated that performing 'sequence matching' on the outputs from separate SED and DOA predictors can yield increased performance relative to models trained on the joint problem [209], and as such may be a better way to utilise the DOA-label relationships outlined above.

The approach presented in this chapter performs the first two stages of SELD using steered-response power beamforming combined with some relatively simple ML models. Given this approach, adding labelling would necessitate the addition of essentially an entirely separate classification model. It was therefore considered that at this stage it would be more illuminating to focus solely on localisation and tracking, perhaps with a view to the eventual inclusion of this method as part of a complete SELD system using sequence matching.

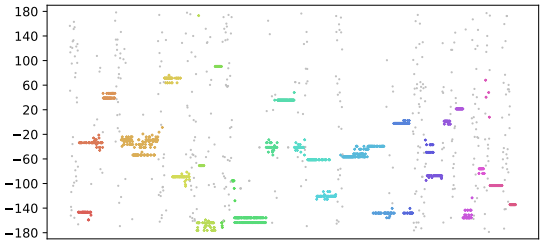
Similar to the approach in the ASC work, the performance of the system using different Ambisonic orders will be compared. Classical machine learning algorithms



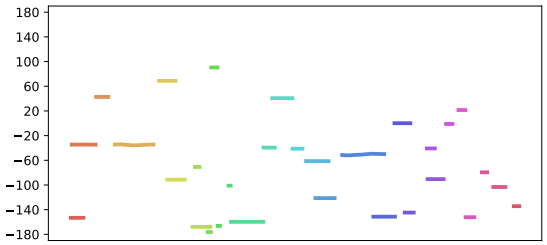
(a) SRP map for an audio frame with a single source.



(b) Array of peaks identified in SRP maps across all frames of a clip.



(c) Clusters of peaks identified by DBSCAN algorithm.



(d) Trajectories estimated by SVR based on identified clusters.

Figure 7.1: The stages of the sound event trajectory prediction system, with the ground-truth reference. The peak array diagrams shows only predicted Azimuth positions for clarity.

were again used in order to focus on the information in the data rather than the power of the classifier. Figure 7.1 shows the four main stages of the proposed sound event trajectory prediction system. The following sections will explore these in detail.

7.2.1 Steered-Response Power Maps

The first stage is the creation of SRP maps calculated in the same way as described in Section 6.3. Owing to the generally reduced ASC performance when using said features, however, several changes were made. Given that one of the reasons for the reduction in performance for ASC might have been the directional quantisation,

the SRP maps for this system used 600-point Fibonacci spherical sampling for increased resolution, rather than the 300-point scheme used previously. The work in this chapter uses both PWD and CroPaC beams across all four available spherical harmonic orders, rather than jumping from first order to fourth order as previously. SRP maps were calculated according to equation 2.26 using frames of 256 samples with no overlap. Note that, unlike in the ASC work, for this application the summation across k is retained so the SRP maps are representative of the full available frequency spectrum. In the example shown in Figure 7.1, PWD beamforming has been used to create SRP maps for a synthesised example file with at most two overlapping sound sources.

7.2.2 Peak-Finding

The next step towards source identification within the SRP maps is peak-finding. In the ASC work, this was achieved very simply by taking the highest-amplitude point from the SRP map in each frequency band. This method is fine when the interest is the general spatial and spectral distributions of sound sources, however it is not suitable when the aim is tracking multiple simultaneously-occurring sound sources which may have similar frequency content. It is not certain that a sound source that is dominant at a certain frequency in one frame of audio will also be dominant in subsequent frames. This could result in wildly erratic arrays of peaks, which would negatively affect the subsequent clustering stage.

True peak-finding in a sampled spherical function represents a challenge in that the wraparound of the sphere at the edges of the data (i.e. $f(\theta, 2\pi) = f(\theta, 0)$) must be accounted for, along with the fact that the sphere has not been sampled using a regular grid. Peak-finding was therefore performed using the *dipy* library [210], which was originally developed for the analysis of medical MRI data, also sampled spherically. This library overcomes the aforementioned issues by requiring the sampling directions as input data as well as the power map. Two parameters determine the behaviour of *dipy*. The first is *rel_pk*, which is used to calculate a

threshold below which peaks are discarded:

$$\begin{aligned} \text{threshold} &= \wedge + rel\_pk \cdot R \\ \wedge &= \max\left[0, \min\left[Z(\theta,\phi)\right]\right] \\ R &= \max\left[Z(\theta,\phi)\right] - \min\left[Z(\theta,\phi)\right] \end{aligned} \tag{7.1}$$

with $Z(\theta,\phi)$ again calculated according to Equation 2.26. The second parameter is *min_sep*, the minimum angular distance allowed between peaks. If two or more peaks are found within this distance of one another, only the largest is retained. This helps to avoid groups of peaks being identified where a sound might be incoming from a broad area, but will also cause peaks to be discarded where narrower sources may pass close together, and as such represents a trade-off between these two considerations.

There is no parameter included in the algorithm which specifies a maximum number of peaks that may be returned, but in creating this system, memory was allocated for a maximum of 20 peaks per frame. This could easily be extended if required for a certain application, but in practise the limit was rarely approached. The output of this stage is an array containing the angles of detected peaks along with a time specified in seconds. The SRP map pictured in Figure 7.1(a) features a single sound source with a black cross indicating the detected peak. The output array of peaks is shown in Figure 7.1(b).

7.2.3 Clustering

The array of peaks is then processed in order to estimate which peaks belong to the same sound sources. The DBSCAN algorithm (see Section 4.3.5) is used to intelligently create clusters of peaks based on proximity in space and time. Onsets and offsets for each sound source are predicted based on the timings of the earliest and latest-occurring peaks that have been grouped into each cluster. DBSCAN was chosen as it is formulated to create clusters of any arbitrary shape within noisy data, with a provision for excluding noisy data points from belonging to any cluster. As

indicated in Equation 7.1, the peak-finding algorithm uses a threshold that changes relative to the range of power levels present in each frame of the data. Consequently, for frames where there is no active sound source, many peaks will be found in the background noise, a behaviour apparent in Figure 7.1(b) where the peaks found in the background noise can be seen as vertical 'stripes'. The effectiveness of DBSCAN at excluding these peaks from consideration is seen in Figure 7.1(c), where those peaks that have not been assigned to clusters are rendered in grey. Whilst DBSCAN does not explicitly take into account temporal structures, there is precedent for its use as an anomaly detector in time-series data [211, 212]. Its application here essentially represents the opposite of this aim for each sound source cluster.

The wraparound of the spherical data once again presents an issue in that, should a particular sound move across the wraparound point, there would be a discontinuity in the spherical co-ordinates that would likely cause DBSCAN to place the peaks before and after the crossing into different groups. To circumvent this, prior to clustering, the spherical co-ordinate of each peak is mapped to a Cartesian co-ordinate on the unit sphere, transforming the clustering space from 3D $[\,t\ \theta\ \phi\,]$ to 4D $[\,t\ x\ y\ z\,]$, a technique also used in [154]. The spatial dimensions of the data were normalised to zero mean and unit variance, as is standard in machine learning. The time dimension was not normalised, as this results in a collapse of the time dimension which causes DBSCAN to cluster peaks occurring at similar spatial co-ordinates but originally separated by large amounts of time. The result of DBSCAN clustering is shown in Figure 7.1(c). The colour-coding indicates peaks that have been judged by the algorithm to belong to the same cluster, with the greyed-out peaks judged to be noise.

7.2.4 Regression

Each cluster that has been identified by DBSCAN is used to fit a set of SVRs, which model trajectories based upon the available data. Individual two-dimensional regressors are used for each spatial dimension, modelling x, y, and z separately

against t. This was necessary as the scikit-learn SVR implementation used does not support multi-variate regression [130]. The outputs of each of these regressors were concatenated to give three-dimensional positions for each time step. This process has the effect of both smoothing the data to reduce the 'jitter' that can occur as adjacent sample points are instantaneously identified as the peak in neighbouring frames, and to fill in missing points as the cluster may not include peaks for every frame. The output from these regressors is calculated for each time step between the first and last-occurring points in each cluster. Since the regressors are fitted to to normalised data, their predictions must be re-scaled back to the original spatial ranges. Finally, the Cartesian co-ordinates are converted back to spherical co-ordinates, and this is the final output of the system. Figure 7.1(d) shows the final azimuth angles output from the SVRs.

Fixed values for C and ϵ (see Section 4.3.4) of 1×10^{-3} and 0.4, respectively were used to fit the SVRs, with a value of 10 for the radial kernel's γ parameter. These values were determined by trial-and-error to return sufficiently smooth trajectory predictions, reducing jitter whilst not completely flattening out source movement. The setting of these values does, however, represent a key limitation in the use of SVRs for this stage, as values that work well for one setting might not necessarily be appropriate for all scenarios. Whilst in this study this did not present a problem (see the description of the dataset in Section 7.3.1), in a complex sound scene, a rate of movement representing noisy jitter in one type of source might reflect actual movement in another. Thus, parameters ensuring accurate smoothing for one source might obscure fine detail in another. Careful tuning of the C and ϵ parameters would therefore be required for each scenario.

7.3 Method

7.3.1 Dataset

This system was originally envisaged with the intention of further investigating the recordings in the EigenScape database. Using real data presents a problem, however, when the performance of the system is to be evaluated. In order to assess performance, ground-truth data is required with which to compare the output of the system. For real recordings such as those in EigenScape, the ground truth would have to be created by manual labelling of the sound sources in the recording with onset and offset times and direction-of-arrival angles. It has been shown that for onset/offset labels, at least five separate annotators are required for an average annotation approximating actual ground truth [213]. Whilst feasible, this work would be very time consuming and as yet has not been undertaken for EigenScape. Annotation of DOA angles presents a further challenge. Indeed it is not clear how one would approach annotation of DOA in a real recording of a scene of normal complexity.

For these reasons, the system was instead tested using an expanded version of the TUT sound events 2018 Ambisonic dataset [214]. This is a set of synthesised Ambisonic scenes, each of up to 30 seconds in length, with sounds placed at static locations at intervals of 10° in the full range of azimuth angles, with elevation angles limited to ±60°. Scenes are synthesised with three levels of polyphony, denoted *OV1*, *OV2* or *OV3* for a maximum of one, two, or three simultaneous sounds, respectively. Using synthesised data means that the onsets, offsets, and positions of the sound sources are precisely known without requiring annotation. The clear disadvantage of using this data is that it is not representative of real world scenes. On the other hand, the precise quantisation of the synthetic data means that variables can be controlled and so the performance of the system given varying numbers of overlapping sources can be reliably assessed. Validating the effectiveness of any source tracking system

using synthetic data is necessary before any conclusions can be drawn following the application of the system to real recordings where ground truth DOA values are not readily known.

The TUT dataset is available in both anechoic and reverberant versions. In this work, the anechoic version was used as it was reasoned that it would be useful to gather data on the performance of the proposed approach in idealised conditions before adding confounding factors such as reverberation. In this way, any limitations in performance can be considered to be inherent in the approach itself, rather than a result of environmental conditions. Furthermore, the reverberant version of the dataset was created using RIRs from only one small room [215], and this was considered to be unrepresentative of either the open outdoor or large indoor spaces typical of urban soundscapes and recorded in the EigenScape dataset.

Since the original version of the TUT dataset was synthesised in first-order Ambisonic format only, this work uses a new version re-synthesised to fourth-order Ambisonics as a set of overlapping plane waves according to Equation 2.19. The original variety of everyday sounds from the DCASE 2016 Task 2 dataset [168] are used. The TUT dataset consists of 240 training clips and 60 testing clips for each level of polyphony. The new approach proposed in this chapter does not require training data, so only the test clips are re-synthesised and used for assessment. All examples are resampled to 16 kHz, as would be necessary with real Eigenmike recordings. It should be noted that for beamforming using this synthetic data, the value of the $b_n(kr)$ term in Equation 2.23 is 1, as there is no microphone array scattering in synthesised scenes.

7.3.2 Metrics

Performance was assessed using two frame-wise DOA metrics employed in judging the SELD systems submitted to DCASE 2019 task 3 [216]. Firstly, *DOA error*, which indicates the average error between predicted and actual DOA angles, defined as [217]:

$$\text{DOA error} = \frac{1}{\sum_{t=1}^{T} D_E^t} \sum_{t=1}^{T} \mathcal{H}(\mathbf{DOA}_R^t, \mathbf{DOA}_E^t) \tag{7.2}$$

where $\mathbf{DOA}_R^t$ and $\mathbf{DOA}_E^t$ are lists of reference and estimate DOAs for frame t, and D_E^t is number of estimates in $\mathbf{DOA}_E^t$. $\mathcal{H}$ denotes the Hungarian algorithm [218], which is used to match predicted angles to reference angles by assessing and optimising pairwise costs based on angular distance. A system which predicted the DOAs of all identified sources exactly would have a DOA error of 0. This metric, however, does not give any information regarding the identification of sound activity.

The second metric is therefore an indication of the proportion of frames where the estimated number of active sounds matches the actual number of active sounds described in the reference. This is named *frame recall* (FR), and is defined as:

$$\text{FR} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(D_R^t = D_E^t) \tag{7.3}$$

where D_R^t is the number of estimates contained in $\mathbf{DOA}_R^t$ (i.e. the ground truth number of active sources present in frame t) and $\mathbb{1}$ is a binary indicator function returning 1 if the condition is met, otherwise returning 0. A system which perfectly predicted the number of sources in every frame would therefore have an FR of 1. This metric could be considered too strict, as it does not make any distinction between small or large errors in polyphony. If, for instance, a prediction was made of two active sources when in fact there were three, this would be penalised to the same degree as an estimate of one active source. On the other hand, this means that a high FR score indicates very good performance indeed.

Good performance according to one of these metrics will not necessarily translate to good performance according to the other. A system achieving 0° DOA error could in fact be a system producing no output. A successful system should therefore aim to maximise frame recall whilst minimising DOA error.

7.3.3 Optimisation

The approach taken with this system is similar in essence to that proposed by Bunting [147], in which the plastic self-organising map (PSOM) [219], a neural network technique, is used to cluster peaks estimated using DirAC. As Bunting describes, "this grouping can be considered analogous to clustering input data according to physical locations". Whereas in the work presented in this chapter, peaks are transferred from spherical to Cartesian co-ordinates, in Bunting's work the PSOM algorithm is instead reformulated to work in spherical co-ordinates. Bunting's system works conditional on ω-disjoint orthogonality, which stipulates that simultaneous sound sources must not overlap in frequency and time in order to remain separately identifiable by the clustering algorithm. Given that in the system presented here, sound power is summed over the frequency domain, the requirement of ω-disjoint orthogonality cannot be applied. Nevertheless, a similar condition that does apply in this case is the fact that in order for two simultaneous sources to be separable by this new approach, at any given point in time those sources must not pass within a certain spatial distance of one another. This distance is in practise conditional on the DBSCAN *Eps* and *MinPts* parameters interacting with the *rel_pk* and *min_sep* parameters of the peak-finding algorithm, together with the resolution of the spherical sampling.

The *hyperopt* library [220] was used to run 1000 iterations to find optimal values for these four parameters. This was done using all 60 test files for each available Ambisonic order and level of polyphony. *Hyperopt* employs the tree parzen estimator (TPE) algorithm [221] to find optimal value combinations over time. This is a more focussed optimisation process than a random search and allows greater fine-tuning of performance for a given number of iterations. The algorithm was set to optimise for Frame Recall. Following preliminary tests to find appropriate ranges, the search space was set as follows:

- $\{Eps \in \mathbb{R} \mid 0.1 \leq Eps \leq 1.25\}$

- $\{MinPts \in \mathbb{Z} \mid 3 \leq MinPts \leq 10\}$
- $\{rel\_pk \in \mathbb{R} \mid 0.0 \leq rel\_pk \leq 1.0\}$
- $\{min\_sep \in \mathbb{Z} \mid 0 < min\_sep < 90\}$

The effect of changing the grid resolution was not investigated, as it would have taken a lot more time to calculate multiple SRP maps for every scene.

A key question to be answered by this investigation is how critical the setting of the hyperparameters is to achieving optimal results. If wildly different values are required for different levels of polyphony, this would be of limited utility in real world scenarios, as in practise the level of polyphony in real environments is unlikely to be known.

7.4 Results and Discussion

7.4.1 Overall Performance

Figure 7.2 shows the DOA Error and FR of the optimised iterations across all Ambisonic orders and levels of polyphony, with previously-published results from SELDnet [154] (discussed in Section 4.4.2) included for comparison. It can clearly be seen that performance on the *OV1* clips is excellent regardless of N or the beampatterns used, with almost perfect FR and DOA errors in the range of 2.5 - 3°. This is unsurprising as single sources with absolutely no interference of any kind should not be difficult to locate, even using beams with low spatial resolution. For the *OV2* clips there is a clear pattern of improving performance with increasing N for both PWD and CroPaC iterations. CroPaC performance is slightly better than PWD performance for all orders in terms of DOA error, but in terms of FR only improves on PWD for *N2*, with slightly reduced FR relative to PWD in *N3* and *N4*. There is a much larger gap between the *OV2/N1* and *OV2/N2* PWD results than between the higher orders. FR for *OV2/N4* is 0.85 with a DOA error of 2.9°,

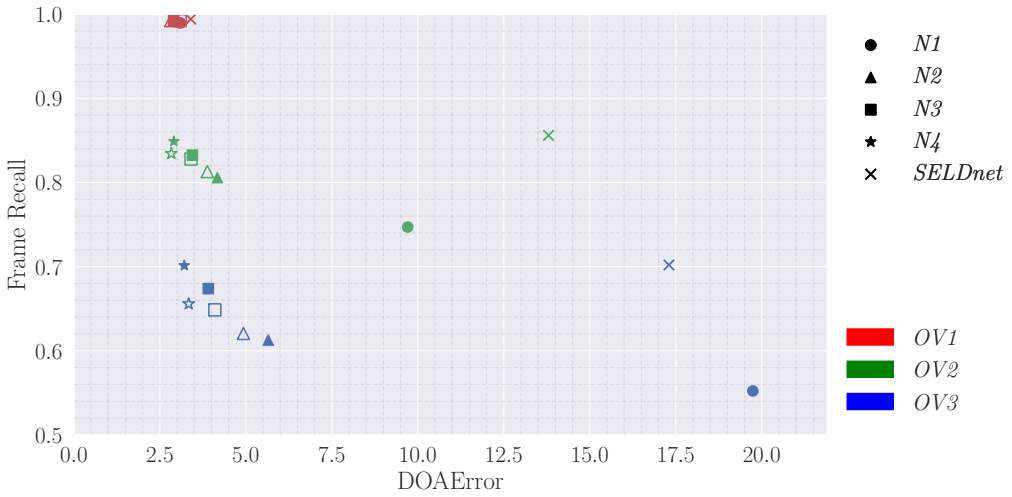
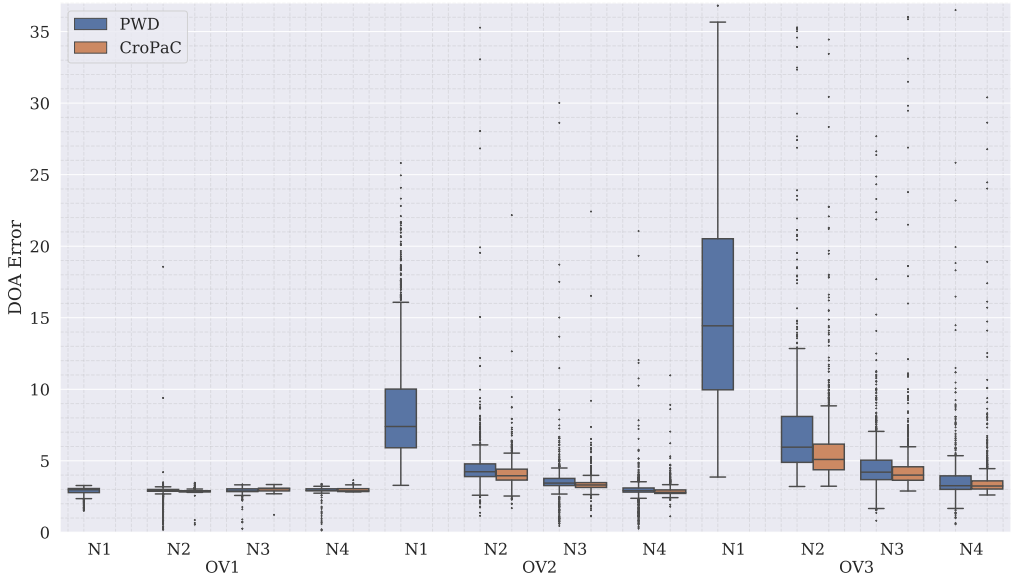


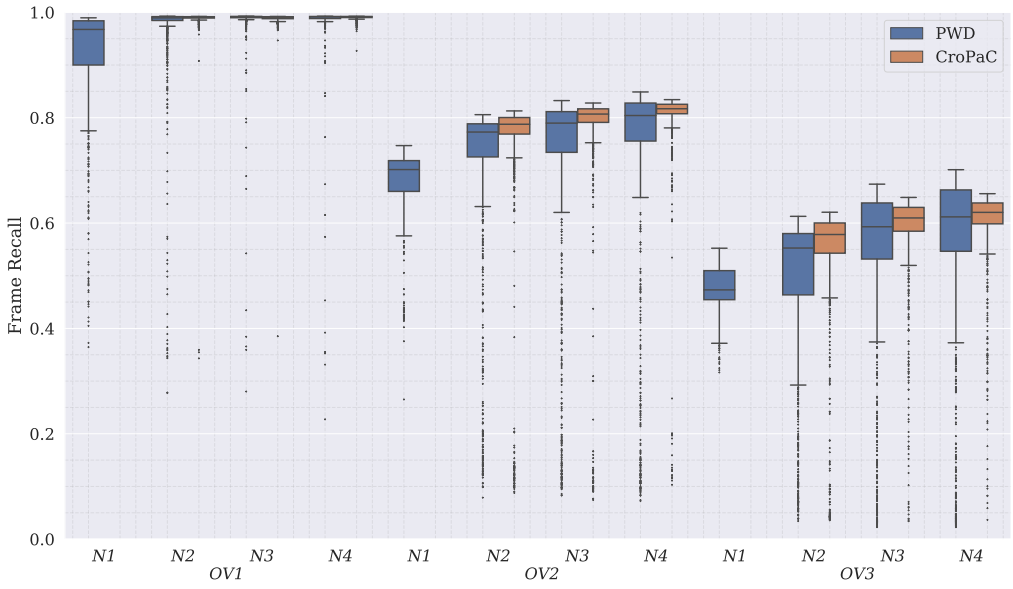
Figure 7.2: DOA Error and Frame Recall scores for iterations optimising FR. Filled markers indicate iterations using PWD beamforming and unfilled indicate CroPaC. Results from the SELDnet system [154] are included for comparison.

whereas for *OV2/N2* FR is 0.81 with a DOA error of 4.2°. These are remarkably similar compared to the *OV2/N1* FR of 0.75 and DOA error of 9.7°. This pattern recurs more prominently in the *OV3* results. The DOA error for *OV3/N1* is 19.7°, whilst increasing the order to *N2* reduces the DOA error to 5.7°, an improvement of 14°. The difference between *N2* and *N4* is only 2.5°. The increase in FR between orders at *OV3* is more linear, though still shows diminishing returns as N increases. At *OV3*, use of CroPaC beams only improves on the PWD results for *N2*. CroPaC performance for *N3* and *N4* is worse than PWD for both metrics. In general, when these results are compared to results from SELDnet, DOA error is smaller with this new approach except for *OV3/N1*. SELDnet generally outperforms the new system in FR, except using PWD at *N4*.

Figure 7.3 shows the distribution of results returned by the 1000 iterations of the system for each OV and N using various hyperparameter combinations. It should be noted that due to *hyperopt*'s use of the TPE algorithm, these are skewed towards results using hyperparameters set near to optimal values. This is reflected in the large number of visible outliers, though they are a small minority of the

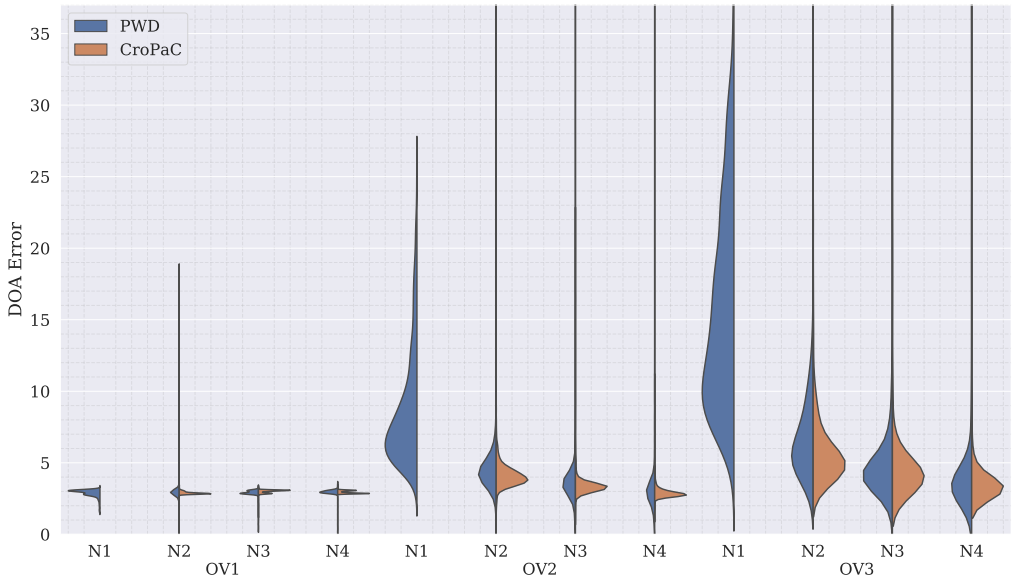


(a) DOA Error

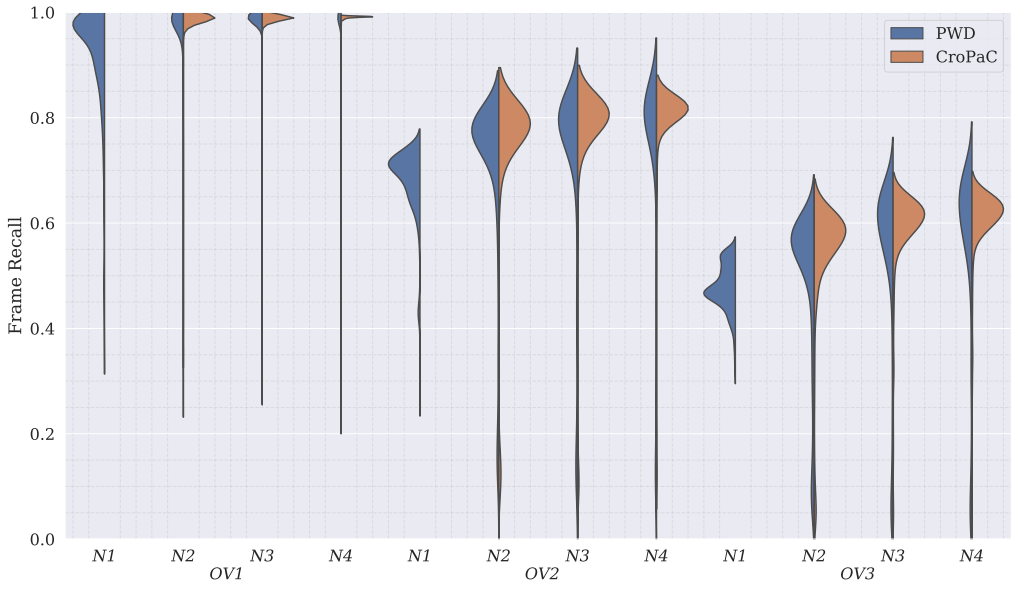


(b) Frame Recall

Figure 7.3: Box plots showing distributions of (a) DOA error and (b) Frame Recall for both PWD and CroPaC-based iterations across all 1000 *hyperopt* iterations.



(a) DOA Error



(b) Frame Recall

Figure 7.4: Violin plots showing distributions of (a) DOA error and (b) Frame Recall for both PWD and CroPaC-based iterations across all 1000 *hyperopt* iterations.

1000 iterations. Figure 7.4 shows violin plots of this same data, demonstrating that, despite the aforementioned skew, in most cases the distributions approximate normality in the region around modal values. Based on this, it is reasoned that the effect of the outlier data should be small enough that it remains meaningful to discuss the median and interquartile range (IQR) values displayed on the boxplots in Figure 7.3. The mean values are excluded from this discussion, however, as these will be more affected by the skew.

Looking first at the PWD results in Figure 7.3, it can clearly be seen that increasing the Ambisonic order causes a marked decrease in the median and IQR of the DOA error for *OV2*, with even larger differences at *OV3*. Similar to the patterns indicated in Figure 7.2, the effect is most pronounced between *N1* and *N2*, with smaller gains at higher orders. This general pattern of improvement at higher orders is also visible in the FR results, though again it is not so marked as for DOA error. FR declines consistently given increasing polyphony. The IQR for PWD at *OV1* decreases between *N1* and *N2*, but is consistent at all orders for *OV2*. It increases between *N1* and *N2* for *OV3*, remaining consistent with *N2* for all higher orders.

Comparing PWD with CroPaC, it can clearly be seen that, in general, CroPaC results have an improved median score and lower IQR for both metrics. This indicates that for an iteration using some combination of hyperparameters within the ranges specified in Section 7.3.3, one can typically expect better results from an iteration using CroPaC beams than one using PWD. This is reflected in the violin plots shown in Figure 7.4, in which it can clearly be seen that the distribution peaks are generally narrower for CroPaC than PWD. However, as indicated in Figure 7.2 and the top whiskers in Figure 7.3(b), the very best performances attained using PWD exceed those achieved using CroPaC.

In terms of both best and median scores, results indicate that there is a larger performance gap between iterations using first and second-order Ambisonics than those using higher orders. For DOA error, there is also a large reduction in the

IQR between first and second-order, which again is not so pronounced for higher orders. Increasing N incurs various penalties in terms of cost. The higher the order used, the greater the amount of storage required for the recorded data, and the more computationally complex the beamforming stage becomes. For instance, the filesize of a ten-minute clip from the EigenScape database in fourth-order is 2.16 GB, whereas for the first-order version this is 345.6 MB. The number of microphone capsules required for adequate sampling of real-world spatial audio also increases. For these reasons, there is a clear incentive to keep the value of N as low as possible for systems to be deployed at scale. The large increase in performance between *N1* and *N2* indicates that using second-order audio may be worth the increased cost, with the limited performance gains at higher order suggesting that second-order represents a good point of compromise for this application. Use of second-order Ambisonics also enables the calculation of CroPaC beams.

The fact that FR seems to decrease linearly with increasing polyphony is interesting, especially in light of the fact that the test audio was anechoic. Coupled with the diminishing returns in terms of performance improvements with increasing Ambisonic order, which indicate an asymptotic trend towards a maximum performance level which is less than perfect, these results could show the existence of an upper limit imposed either by the dataset or more fundamentally with this type of approach. The close alignment of the best FR results achieved here with the results from SELDnet provide additional weight to this observation. Further improvements to FR might require an alternative method for producing the power map than the SRP method used here or the spatial pseudo-spectra generated by the neural networks in [154, 156].

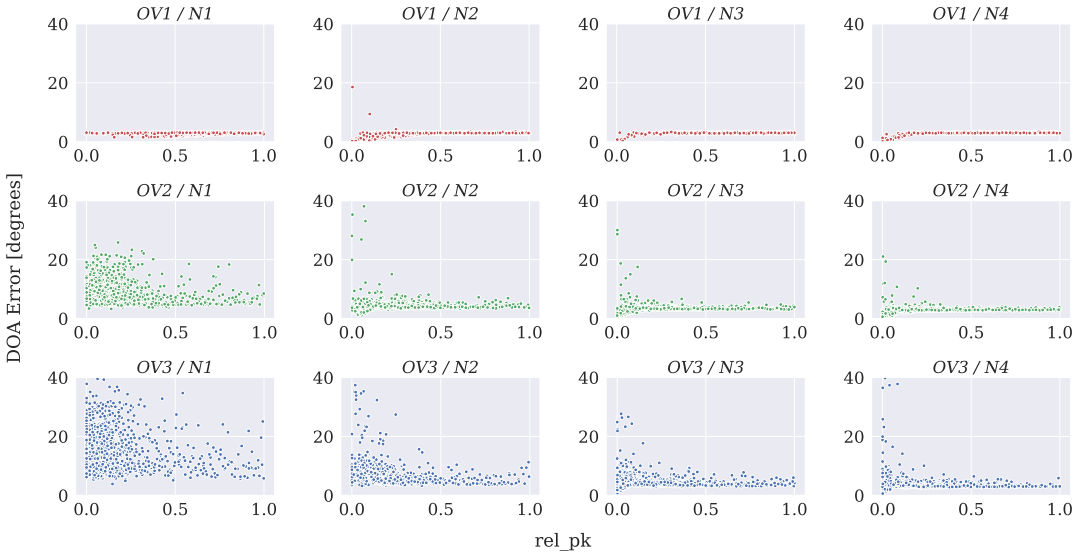
The lowest DOA errors achieved at each OV for the optimised iterations shown in Figure 7.2 are all very similar, lying between 2.5 and 3°, an observation that is mirrored in the lowest median values shown in Figure 7.3, which are near these values. This corresponds closely with the average angular distance between pairs of points in the 600-point Fibonacci spiral, which is 2.72°, and may indicate that

the system could achieve even lower DOA errors if a finer grid pattern was used. On the other hand, the fact that results up to fourth-order continue to improve provides some evidence that, in fact, the sampling scheme used is capturing increased detail from fourth-order Ambisonics, so is in some sense out-resolving third-order Ambisonic signals at least. It would be interesting in a future study to test using sampling schemes of varying resolution to find the minimum required density to achieve maximal results at each order, thus minimising the required computing power.

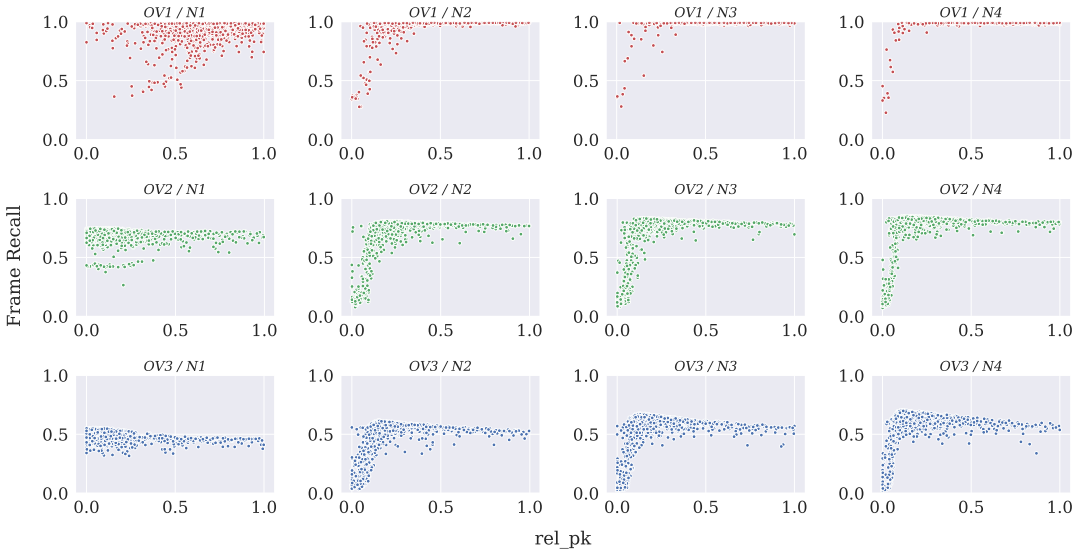
Looking closely at Figure 7.3(a), it can be seen that there are several outlier iterations achieving very low DOA error values, especially using PWD beams. For instance, the lowest DOA error achieved by a PWD iteration for *OV2/N4* was 0.26°. The FR of this iteration was, however, at 0.14, very low. This pattern of very low DOA error at the expense of FR was observed across all of the low-end outlier PWD-based iterations. Upon closer investigation, it was found that these iterations used low values for the *rel_pk* parameter discussed in Section 7.2.2. Using very low values for *rel_pk* likely results in clusters of peaks being identified in the general directions of active sound sources, as opposed to the single peaks returned when the parameter is set more appropriately. The subsequent regression stage may be able to take advantage of these clusters, interpolating to find a truer DOA for the source lying in the central region of these peaks. Unfortunately, using low values for these hyperparameters also results in an increase in spurious peak identification, which causes the reduction in FR.

7.4.2 Parameter Tuning

Figure 7.5 shows scatter charts of performance metrics for PWD iterations, varying with *rel_pk* values for all N and OV. The spread of results at each value of *rel_pk* is caused by variations in the other variables, nevertheless, there are some clearly visible trends. DOA error for lower *rel_pk* values (Figure 7.5(a)) tends to be more variable, with some iterations having very low DOA error. As *rel_pk* is increased,



(a) DOA error values for all PWD instances, varying with *rel_pk*.



(b) Frame Recall for all PWD instances, varying with *rel_pk*.

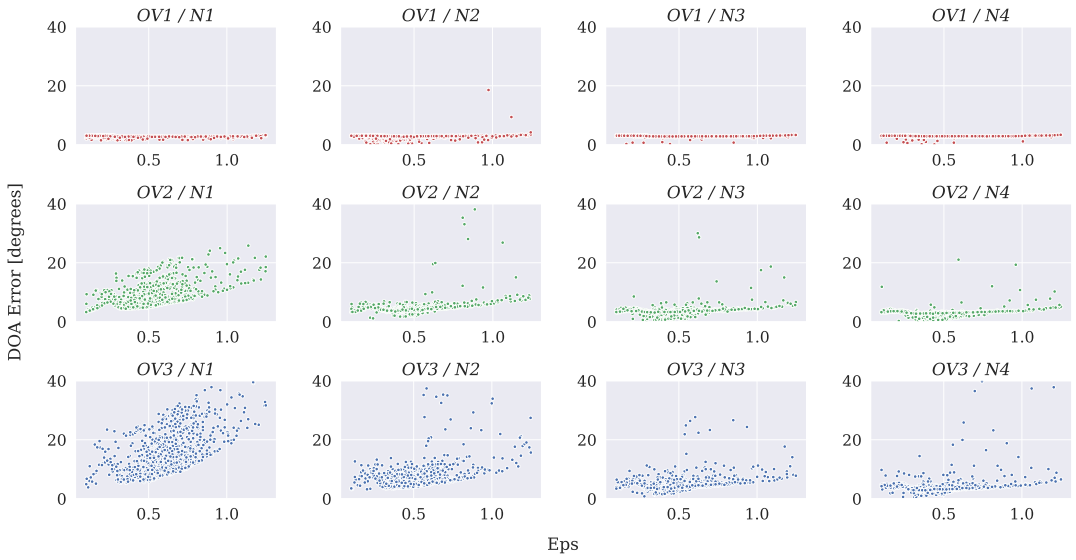
Figure 7.5: Performance metrics for instances using PWD beams, varying with the *rel_pk* parameter of the peak-finding stage.

the spread reduces and error tends to level out to around the 3° value observed in the best iterations. This pattern is most visible when N and OV are both 2 or above, whereas for *OV1* the low values tend to give more uniformly low DOA error at all orders without the increased spread of results, and for *OV2/N1* and *OV3/N1* the spread remains relatively high throughout the range of values.

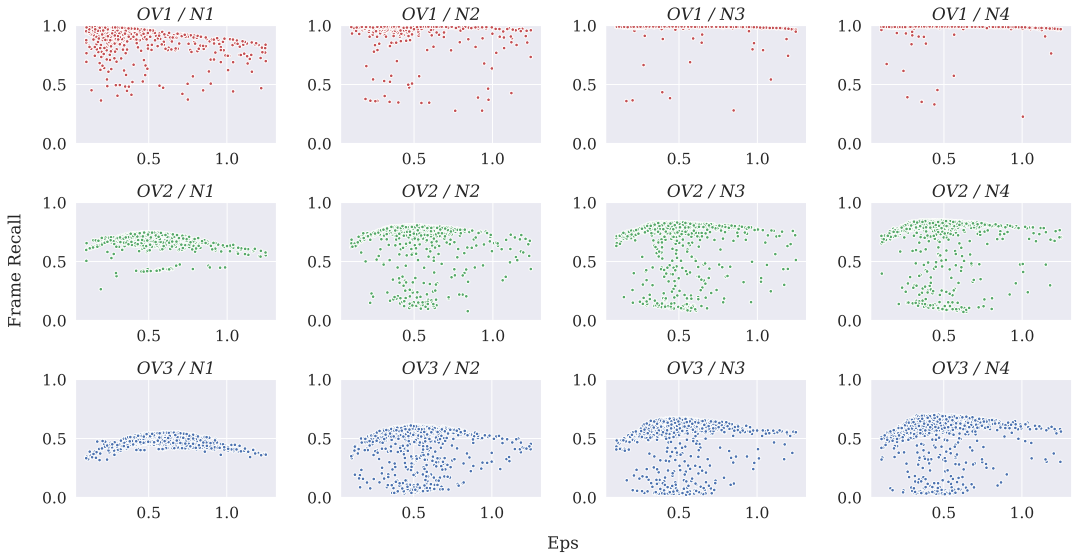
The FR results shown in Figure 7.5(b) are similar in that when N and OV are both 2 or above, results follow a consistent pattern, all showing a distinct 'knee' where maximum performances are achieved at *rel_pk* values of around 0.1, with a shallow decline as the value increases beyond this point. These charts also have implications for the question of the criticality of hyperparameter settings given different levels of polyphony. It is evident from Figure 7.5 that a reasonable setting for *rel_pk* is a value between 0.1 and 0.2. This value appears to be somewhat consistent where OV and N are 2 or greater. Since real acoustic scenes are almost guaranteed to have some polyphony it seems prudent to consider these somewhat more representative of real scenes than *OV1*. These findings add weight to the conclusion that second-order Ambisonics represents a good compromise in spatial resolution.

Investigating how performance metrics are affected by *Eps* again reveals definite trends. Figure 7.6 shows performance metrics for instances using PWD, varying with *Eps*. It can again be seen that trends are fairly similar for $N \geq 2$ iterations with *OV2* or *OV3*. Figure 7.6(b) shown arched patterns in FR values, with a peak at *Eps* values between approximately 0.3 and 0.5. These maxima correspond, however, to points of largest spread of results, with some very poor FR scores forming an opposite minima. It is likely these results are from iterations using the low *rel_pk* values explored previously, as upon comparison with Figure 7.6(a) it seems that these regions tend to coincide with some iterations with very low DOA error. Apart from this, the DOA error tends to increase slightly with increasing *Eps* above a value of around 0.5.

Appendix B.1 contains scatter charts similarly relating the other two variables, *MinPts* and *min_sep*, to the performance metrics for PWD instances. Studying



(a) DOA error values for all PWD instances, varying with *Eps*.



(b) Frame Recall for all PWD instances, varying with *Eps*.

Figure 7.6: Performance metrics for instances using PWD beams, varying with the *Eps* parameter of the clustering stage.

Figures B.1 and B.2 it is clear that changing the values of these variables has very little influence on the results. FR and DOA error values and variability remain consistent across the ranges of both variables, apart from perhaps a slight downward trend in FR with the larger values of *min_sep*. Since FR and DOA error are frame-wise metrics, it is perhaps not surprising that these are not affected by *MinPts* as they do not take into account information from beyond their immediate frame of reference. Lower values of *MinPts* increase the likelihood of the algorithm returning a series of small clusters rather than one large cluster representing a complete sound source from beginning to end. These metrics will not penalise this, as for an individual frame it makes no difference whether an identified point has been assigned to a group with reasonable onset/offset times or not. Using an alternative event-wise metric might provide further insight. In the present study no such metric was used, so it is difficult from these results to come to any conclusions with regards to optimal values for these variables.

Appendix B.2 contains scatter charts with the performance metrics varying with all parameters for the instances using CroPaC. It is apparent when comparing Figures B.3 and B.4 with Figures 7.5 and 7.6 that the trends visible when varying *Eps* and *rel_pk* are essentially the same as for the PWD system, though are grouped somewhat more closely, reflecting the lower IQRs observed using CroPaC. Based on these results and those discussed previously, use of CroPaC in systems deployed in the field can be recommended if the requisite computing power is available. Lower IQR indicates a higher degree of consistency in results with less dependence on hyperparameter tuning, which are both desirable features.

7.4.3 *Eps* and Physical Distance

In this approach, *Eps* values represent a distance in the four-dimensional space in which the spatial Cartesian co-ordinates have been normalised to zero mean and unit variance. Since the time dimension was not normalised and is represented as a value in seconds, the distance between two adjacent time steps is 0.016s. This is an

order of magnitude lower than the peak *Eps* values, so it can be inferred that for temporally nearby frames the spatial distance dominates. Reversing the normalisation process, it can be calculated that *Eps* values between 0.3 and 0.5 represent, in adjacent frames, physical angular distances between 6.4° and 10.1°. This corresponds very closely with the minimum angular separation of 10° specified in the dataset used. It is therefore unsurprising that values above 0.5 result in gradual performance reductions, as these will allow clusters to be formed consisting of two or more separate sound sources, in turn causing the regression stage to incorrectly interpolate between peaks belonging to multiple sources.

Considering further the relationship between *Eps* and the time dimension, it is clear that the contribution of time to the distance between peak points is exactly equal to the time difference measured in seconds. Over larger amounts of time than a few frames, the time distance begins to dominate the calculation. This has the effect of reducing the maximum angular distance within which points will be included in a cluster. With an *Eps* value of 0.5, for instance, spatial points up to around 10° apart in adjacent time frames will be grouped. As the time distance increases this angle will reduce until, at a temporal distance of 0.5s, only points with exactly the same spatial co-ordinates will be grouped. Beyond this time separation, even peak values at the same spatial locations will be assigned to different clusters.

As previously mentioned in Section 7.2.3, normalising the time dimension resulted in large reductions in performance as peaks at similar spatial locations were grouped despite being separated by large amounts of time. In fact, normalising the time dimensions results in a reduction of temporal step size by an order of magnitude. Hence, in this case the contribution of the time dimension to the distance between points remains small even over large periods of time, which results in the aforementioned erroneous clustering. These considerations raise the possibility that the degree of expansion or contraction of the time dimension could be a useful parameter in this approach to be tuned further depending on the application context, and in particular the expected temporal intervals between source activities.

7.5 Summary

This chapter has proposed and investigated a new approach to source localisation and tracking using spherical harmonic beamforming, DBSCAN clustering and support vector regression to identify sound sources in Ambisonic recordings. Instances using both PWD and CroPaC beamforming were tested using audio synthesised up to fourth-order Ambisonics and featuring three levels of polyphony. The *hyperopt* library was used to test 1000 iterations for each beampattern, Ambisonic order, and level of polyphony, in order to find optimal values for various hyperparameters. It was shown that for PWD instances, use of second-order Ambisonics improved performance significantly over first-order. Using higher-order channels improved performance over that of second-order, but to a much smaller degree, leading to the conclusion that second-order Ambisonics might represent a good point of compromise between performance, in terms of localisation accuracy, required computing power, and the cost of hardware. Given this result, it could be interesting for a future project to return to the spatial ASC work of Chapter 6 to see whether a similar pattern is evident in ASC accuracies, as that work only compares results from first and fourth-order Ambisonics, without considering the intermediate orders.

It was found that whilst the very best-performing PWD instances tended to give better results than equivalent CroPaC instances, the median performance using CroPaC was better and IQR consistently lower than using PWD. Since a lower IQR indicates a degree of robustness to the variation of hyperparameters, it would be a good idea to use CroPaC beamforming in real-world systems, computing power permitting.

Results also showed that, for the most part, where variations in parameter values were shown to affect significant trends in the performance metrics, optimal values were relatively consistent for iterations using second-order Ambisonics and above for two or three simultaneous sound sources. These settings could therefore be recommended as good starting points should a system such as this be deployed in

the field. The clear next stage of development would be the incorporation of a source labelling stage, which would make this a fully-fledged SELD system.

Whilst this chapter built on the findings of Chapter 6 by construction of a system for analysing acoustic scenes in greater detail, the next chapter will build upon the EigenScape dataset in a different way by building it into an application designed for real-time monitoring in the field, incorporating ideas from both soundecology and augmented reality audio which were explored in Chapter 3.

8 | Machine Learning for Soundscapes in Practise

Associated Publications

- M. C. Green and D. T. Murphy, "Environmental Sound Monitoring Using Machine Learning on Mobile Devices", *Applied Acoustics*, vol. 159, pp. 107041, February 2020, issn: 0003-682X, doi: 10.1016/j.apacoust.2019.107041
  - All of the work in this paper, including development of the iOS app, was by the author of this thesis.

Contributions

- Novel app combining AR audio and machine learning for soundscape monitoring.
- Method of estimating contributions of human, natural, and mechanical sound sources based on ASC techniques.
- Adaption of NDSI metric to attempt to describe perceptual *pleasantness* based on the estimated contributions of natural and mechanical sound sources.

8.1 Introduction

In the previous two chapters, different approaches have been proposed and tested for the characterisation of Ambisonic recordings of acoustic scenes, first in terms of classifying the whole scene and then tracking of individual sources. Whilst both of these studies yielded interesting results, neither were systems that are practical for deployment in the field at present. The work in this chapter takes another approach, creating an iPhone application that utilises the EigenScape data and builds on the acoustic scene classification work from Chapter 6. The app conducts real-time monitoring of acoustic environments in the field, incorporating ideas from the soundscape approach that have informed much of the work in this thesis.

The primary goal was the assessment of whether an intuitive measurement system could be created for environmental sound monitoring on a mobile device, using machine learning to produce meaningful readings beyond L_{Aeq}. A secondary objective was the exploration of whether augmented reality (AR) technology could be used in conjunction with ML to assess interventions to real environments that could alter their soundscapes. The app was therefore developed according to two criteria:

1. Provide an intuitive interface for the measurement of acoustic environment properties inspired by the soundscape approach.

2. Use AR technology to allow users to test the effects of environmental alterations on said soundscape measurements.

To this end, an app was created allowing users to place and move virtual objects with both visual and auditory components. The resultant augmented acoustic scene can be heard by the user and is also passed to a machine learning component for analysis. This was made possible by using the Sennheiser AMBEO smart headset (ASH) [95], shown in Figure 8.1. As outlined in Section 3.5.3, the ASH combines high-quality binaural mcirophones and earbuds into one headset to enable

Figure 8.1: The Sennheiser AMBEO Smart Headset

augmented audio. The rest of this chapter details the construction of the app, named *Soundscape AR*, followed by its testing at several locations in York city centre.

8.2 App Development

8.2.1 Core ML Model Creation

To keep the app development practical, it was decided to focus on classifying acoustic scenes according to the high-level perceptually-motivated sound categories detailed in Section 3.4.3, namely human, natural and mechanical sound. The ML model employed in the app is created along the lines of ASC as explored in Chapter 6 rather than SELD, as in Chapter 7. Although the usual aim of an ASC system is to assign class labels to incoming audio clips, the scene classifiers used here are repurposed to provide estimates for the prevalence of sounds from each perceptual category.

The Core ML library [222] is used to incorporate a model into the app to analyse the audio incoming from the ASH. Models are again trained using the EigenScape data for all eight available scene classes in a manner similar to that described in

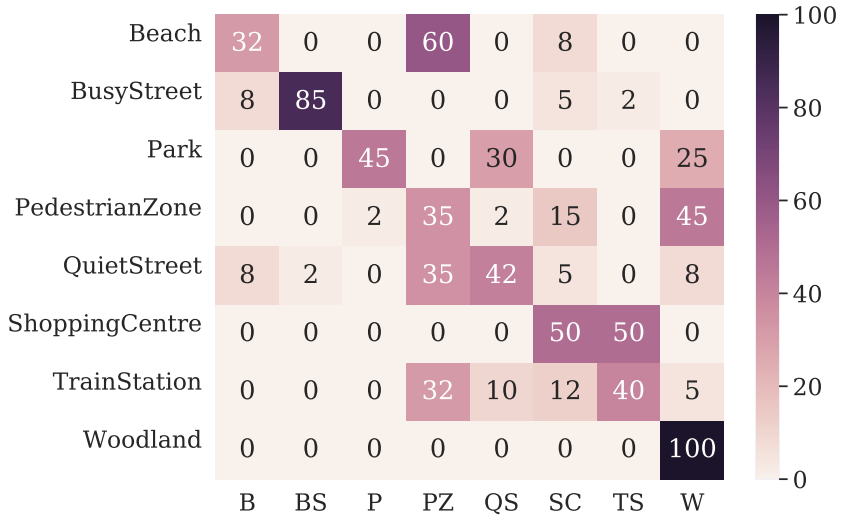


Figure 8.2: Confusion matrix for SVC-MFCC classifier (*aubio*-extracted MFCCs).

Chapter 6, but with several important differences owing to compatibility with the mobile hardware and software. Whilst the previous work in this thesis makes extensive use of Ambisonics, only binaural audio was available from the ASH, so the model was created using MFCC features. MFCCs in the baseline system used in Chapter 6 were extracted using the *librosa* library, but since *librosa* will not run in real-time on iOS (the operating system used on Apple iPhone devices [223]) this was substituted for the *aubio* library [224]. The GMMs used in Chapter 6 are incompatible with CoreML, so were substituted for SVCs. Since in this work the real-time on-location sound is effectively the test audio, the SVCs are trained using all the available EigenScape data, rather than partitioned folds as previously. Given that Ambisonic audio was not available from the ASH, it would have been possible to augment the EigenScape data with additional recordings from other databases, most notably the extensive binaural recordings available in the TUT datasets used in DCASE [161, 162, 163, 164]. These binaural recordings were not included so the results presented here could be more easily related to those in Chapter 6.

Figure 8.2 shows the confusion matrix for the SVC classifiers. Whilst BusyStreet and Woodland scenes are classified well, accuracy across all scenes is substantially

lower than that achieved by the GMM-MFCC baseline classifier in Chapter 6. Although better accuracy would be desirable, since precise scene classifications are not the primary concern for this model, this did not present a significant problem. Based on these results, the BusyStreet classifier was repurposed to rate sounds for mechanical content, with the Woodland classifier chosen for natural ratings. It is assumed that these scenes will largely consist of sounds from the perceptual categories their classifiers have been chosen to represent, an assertion which is reinforced by the results of listening tests reported in [80].

Perhaps the most obvious scene model for human sound would be PedestrianZone. PedestrianZone scenes, however, were largely misclassified, with 60% of Beach clips labelled as PedestrianZone despite these recordings being typically absent of human sound. Another scene class with a large human sound component is ShoppingCentre, which is identified 50% correctly. Since the majority of clips mislabelled as ShoppingCentre are TrainStation and PedestrianZone scenes, which both have relatively prominent human sound, it was decided to use the ShoppingCentre classifier for human sound ratings. As with the GMMs in Chapter 6, each of these SVCs outputs a value indicating the estimated probability that incoming features from each frame of audio come from a scene of the same class they were trained on.

It might have been a valid approach to combine the recordings from several scenes assumed to consist of mainly a single perceptual category, for instance adding the TrainStation data to the BusyStreet data with a new label of 'mechanical'. It was felt, however, that this would rely on too many assumptions that would ideally need to be investigated in listening tests to validate the method, and that the single-class approach taken here was sufficient for this initial study.

For the ASC system in Chapter 6, the GMM returning the highest cumulative probability for all frames across a 30-second clip is used to generate a label for that clip. In this application, Platt scaling [225] is used to generate probability estimates from SVC decision values, a function provided by scikit-learn [226]. These probabilities are used as ratings for each perceptual category and are presented to

the user. In other words, estimates for human, natural, and mechanical components are obtained by measuring the similarity of audio in the user's environment to the ShoppingCentre, Woodland, and BusyStreet recordings, respectively.

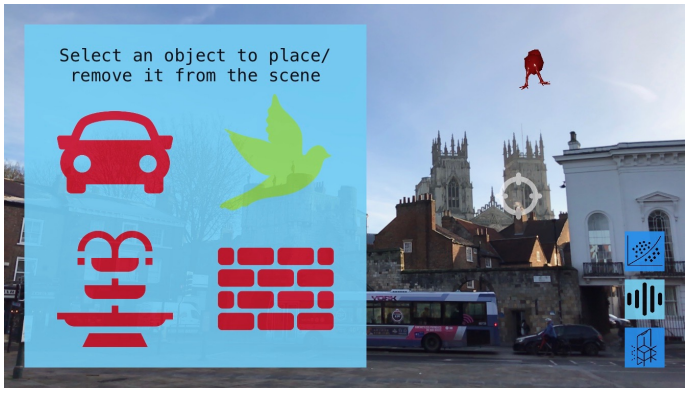
8.2.2 User Interface

Figure 8.3 shows the main views of the app. The main view (Figure 8.3(a)) shows the device's live camera feed and any active virtual objects in view. The app features three sub-windows which perform various functions that can be accessed using three small buttons in the lower right of the interface. The first of these is the AR status window. Before ARKit is able to properly track the environment, which is necessary to ensure consistent placement of virtual objects, detection of the floor or other flat horizontal surface is required. The AR status window indicates the detection of this plane by way of a text indicator turning green. Once this occurs, the window is redundant and when it is closed the user can proceed to place virtual objects.

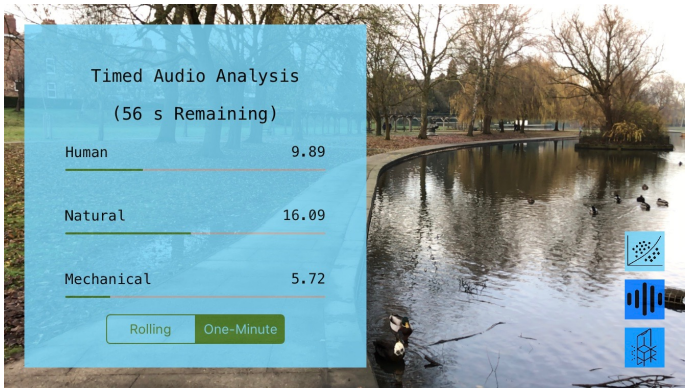
Figure 8.3(b) shows the AR objects window. Four virtual objects are available for placement, a bird, car, water fountain, and acoustic barrier. 3D visual models for each object were taken from the open source repository free3d.com [227], with sounds taken from freesound.org [228] (see Figure 8.4 for spectrograms). The AR objects window represents these as icons in red or green depending on whether each object is active. By tapping on the icons, objects are activated and placed in the scene at a location on the floor plane which the user can specify by aiming crosshairs that are also visible in this view. Placed objects can be tapped and dragged to change their position, and can optionally be moved to hover above the floor plane, a feature useful for realistic positioning of the bird object, as shown in Figure 8.3(b). Tapping on the icon for an active object removes the object from the scene, and the icon will turn red to indicate this.

The key view for environmental sound monitoring is the audio analysis window, shown in Figure 8.3(c). This window features three indicators showing probabilities from the machine listening system. The Core ML object outputs probabilities for

(a) Main app view showing virtual fountain, barrier and bird objects, with interface buttons on lower right.

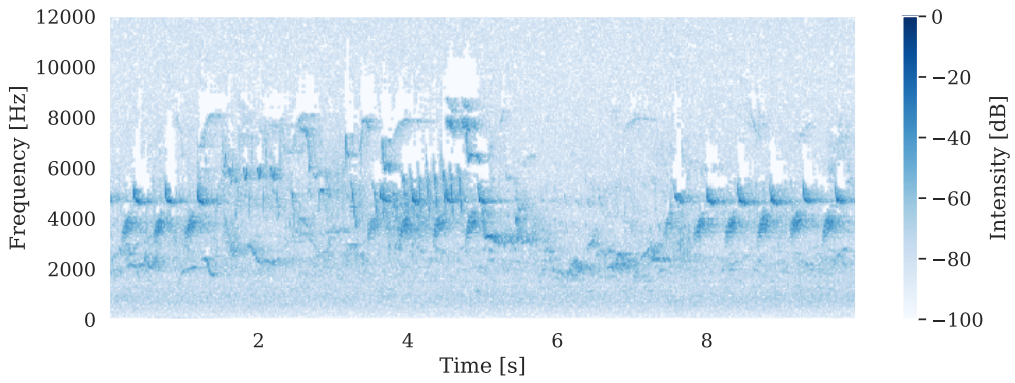


(b) Object selection window showing target crosshairs and active bird object.

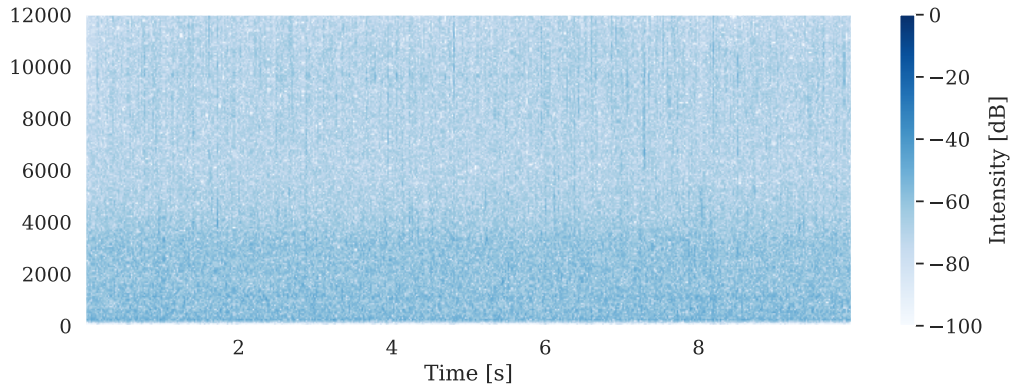


(c) Audio analysis window showing one-minute averaging in progress.

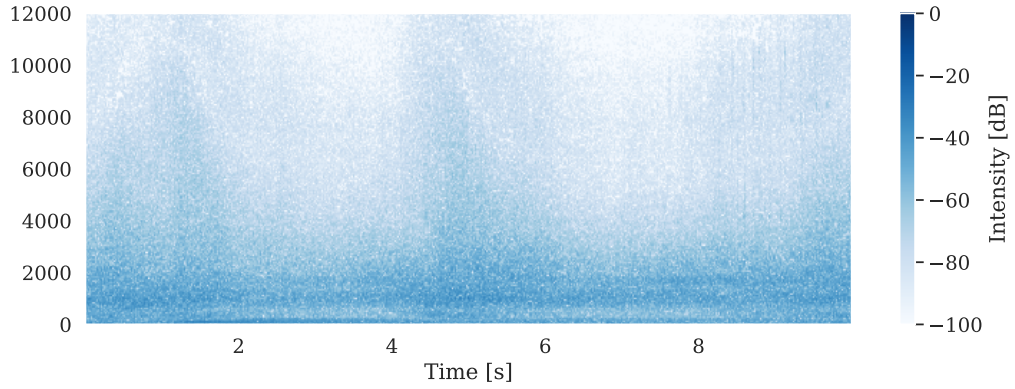
Figure 8.3: Various views of the Soundscape AR interface.



(a) Bird object.



(b) Fountain object.



(c) Car object.

Figure 8.4: Spectrograms showing ten-second segments of the sounds associated with the virtual bird, fountain, and car objects.

every frame of audio, but such instantaneous values can fluctuate rapidly and are very difficult to interpret. These values are therefore smoothed for display using a one-second rolling mean, and this is the default shown to the user. A second mode available in this window is a one-minute mean recording mode, which saves one minute of framewise ratings with the display held at the end to show the mean value for this period of time. Both of these modes are designed to be similar to those used in SPL meters to survey L_{Aeq}, as it is felt that some familiarity with existing approaches might be beneficial if alternative metrics such as these are to find more widespread adoption.

8.2.3 AR Audio Sources

To synchronise AR audio with visuals, custom objects were created in Apple SceneKit [229] to couple 3D graphics with audio sources spatialised binaurally using HRTFs. Default SceneKit objects feature a built-in audio player with "3D audio" [230], which tracks the device rotation to pan an object's audio appropriately. In testing, however, it was found that this used standard stereo panning only. Apple's iOS audio framework [96], on the other hand, contains an object called the AVAudioEnvironment node, which includes the option of using high-quality HRTF rendering. The custom object created for this app therefore adds an extra audio player to the standard SceneKit object. With the 'position' of the audio player set to equal the visual position of the node, the output audio is spatialised appropriately. Audio amplitude is attenuated based on distance according to the inverse square law. The AVAudioEnvironment default is to begin the amplitude rolloff only at distances greater than one metre. This was altered so rolloff is continuous across the whole distance range.

8.2.4 AR Acoustic Barrier

As well as AR audio sources that add sounds to a real scene, a virtual acoustic barrier object was created to simulate the effect of adding a barrier to the scene. The object attempts to selectively filter the real-world sound picked up by the ASH before this is relayed to the listener as part of the complete augmented audio mix. Given the limitation of binaural microphones relative to the flexibility offered by HOA recording and the limited processing power available on a smartphone, this was achieved somewhat crudely using dual low-pass filters (LPFs) from the standard bank of effects included in AudioKit. The output from these is blended with the dry ASH signals and panned with respect to the angle between the listener and the virtual barrier object. As sound impacting a barrier is diffracted around it, the level of attenuation is contingent on the path length difference between the direct sound and the route over the barrier. Whilst the distance between the smartphone and the virtual object is precisely known, there is presently no practical way to measure the distance between the barrier object and the various real-world sound sources, or even to tell whether these sources originate from in front or behind the virtual barrier. The cutoff of the two-channel LPF, representing the amount of high-frequency attenuation incurred by the barrier, is therefore calculated based only on the device-barrier distance. If this distance is 0, the cutoff is set at 20 Hz, effectively blocking all sound in the direction of the barrier. The cutoff is gradually increased logarithmically with distance, set to reach 20 kHz at 10 metres, effectively neutralising the effect and approximating the negligible impact of real sound barriers when the path length difference is very small. The complete effect of this is to give a reasonable impression of the attenuation of high-frequency sound in the direction of the virtual barrier as the user moves through the scene and re-orientates the device.

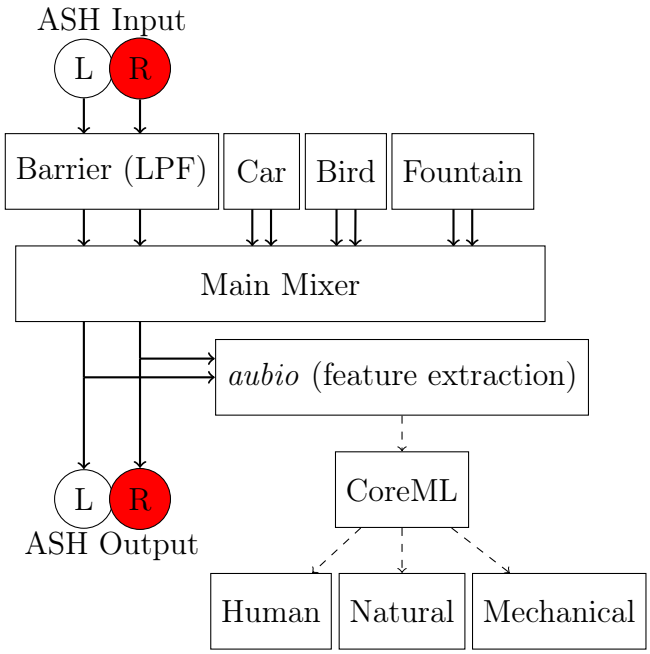


Figure 8.5: Diagram showing the complete audio flow in Soundscape AR from real-time ASH input, via AR audio objects, to binaural ASH output. Dotted lines indicate numeric (non-audio) data.

8.2.5 Audio Flow

The structure of the app's audio engine is shown in Figure 8.5. The main mixer combines audio from any active virtual sources with the binaural audio feed from the ASH. If the barrier object is active, the live audio is filtered through the dual LPF before reaching the mixer. The mix is then passed to the ASH earpieces with minimal latency between input and output. The MFCC features are extracted by *aubio* from a tap on the main mixer output. In this way, the features reflect the complete audio mix including virtual sounds, real sounds, and barrier filtering. The Core ML object receives the MFCCs and returns probabilities for the presence of human, natural, and mechanical audio content in near real-time. By disabling all virtual objects the user can monitor only the real sound scene. The scene can subsequently be re-analysed with added virtual objects to observe the effect on the ratings that the added objects may have.

8.3 Testing

8.3.1 Methodology

The app was tested with two objectives in mind:

1. Assess the effectiveness of the Core ML model in returning meaningful and representative sound metrics.

2. Determine the effect of each virtual object on said metrics (or lack thereof).

The app was loaded onto an Apple iPhone 7 and taken to six locations around the city of York as shown in Figure 8.6. These were chosen to represent a good variety of typical urban acoustic environments, including busy streets (Bishopthorpe Road, Exhibition Square), pedestrianised areas (Shambles Market), green space (Rowntree Park) and locations that combined these features (York Piccadilly, Tower Gardens).

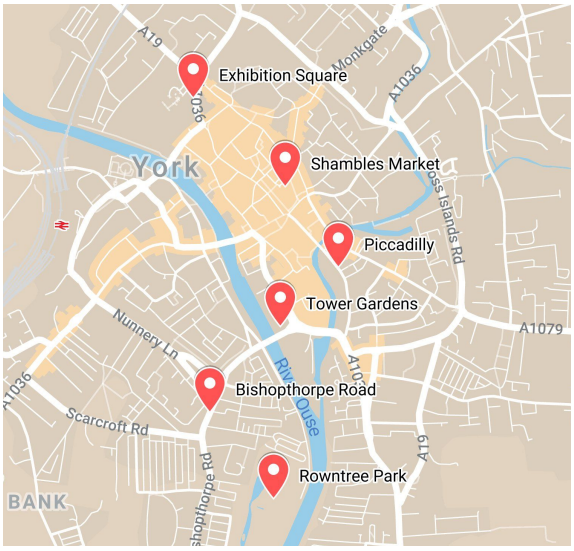


Figure 8.6: Map showing the locations in the city of York where the app was tested.

The one-minute audio analysis feature was used to record ratings at each location, firstly as a clean reading with no objects added, then with the activation of various virtual objects as follows:

- Barrier
- Bird
- Car
- Fountain
- Barrier/Bird/Fountain

All of these virtual objects were placed a distance between 2 and 4 metres in front of the listener location. In the final multi-object condition, the barrier and the fountain were placed to the left and right of the listening position, respectively, with the bird placed roughly 3 metres above the listener.

8.3.2 NDSI/Pleasantness Rating

Section 3.4.2 introduced the normalised difference soundscape index (NDSI), which provides a single number showing the relative contributions of biophonic and anthrophonic sound. Since the current frequency band power method for estimating the contributions of each sound category has been shown to be unreliable in urban environments [50], it was noted that a more reliable method for calculation of this metric would be a desirable research aim. Therefore, as well as analysing the raw ratings returned by the app for each sound category, the natural and mechanical ratings were used to calculate NDSI according to Equation 3.2, as substitutes for β and α, respectively, as it was reasoned that these most closely aligned with their original definitions as the *biophony* and *anthrophony* estimates. As noted in Section 3.4, however, the *natural* and *mechanical* categories are perceptually motivated and do not exactly align with *biophony* and *anthrophony*, so this new version of the NDSI could be considered in perceptual terms to provide ratings indicating a soundscape's *pleasantness* (i.e. its position along the horizontal axis in Figure 3.3).

8.4 Results and Discussion

8.4.1 Core ML Model Performance

The NDSI/pleasantness values calculated for each location are shown in Figure 8.7. These box plots represent the spread of values returned for all the test conditions, with and without virtual objects active. Rowntree Park has the highest pleasantness score, followed by Shambles Market and Tower Gardens. The single outlier value shown for each of these locations is the measurement recorded with the virtual car object active. These values show the effectiveness of the classifier, as the park and market are the locations with the lowest proportion of mechanical sounds, although there was some low-level machinery evident in the background at the market. Tower Gardens is very near the River Ouse, but also borders two of York's main roads, and

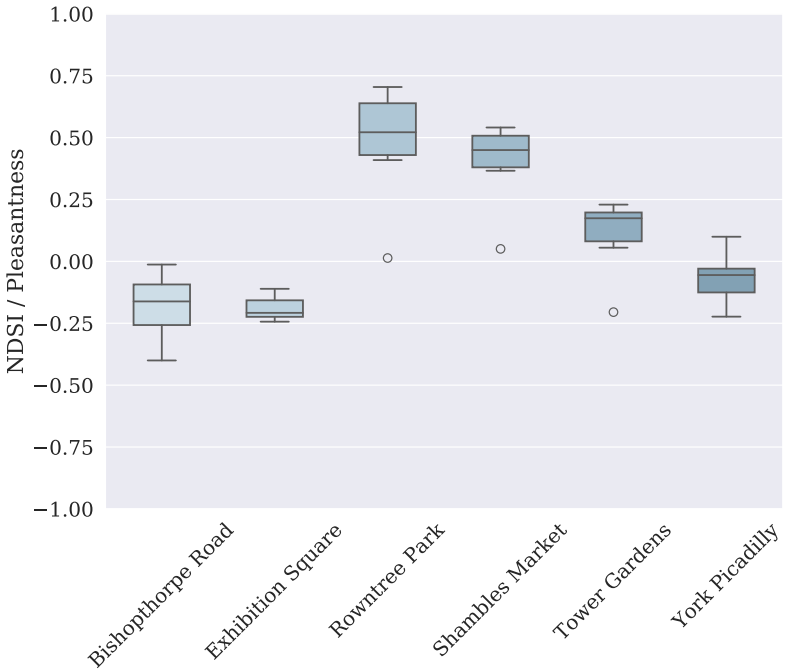
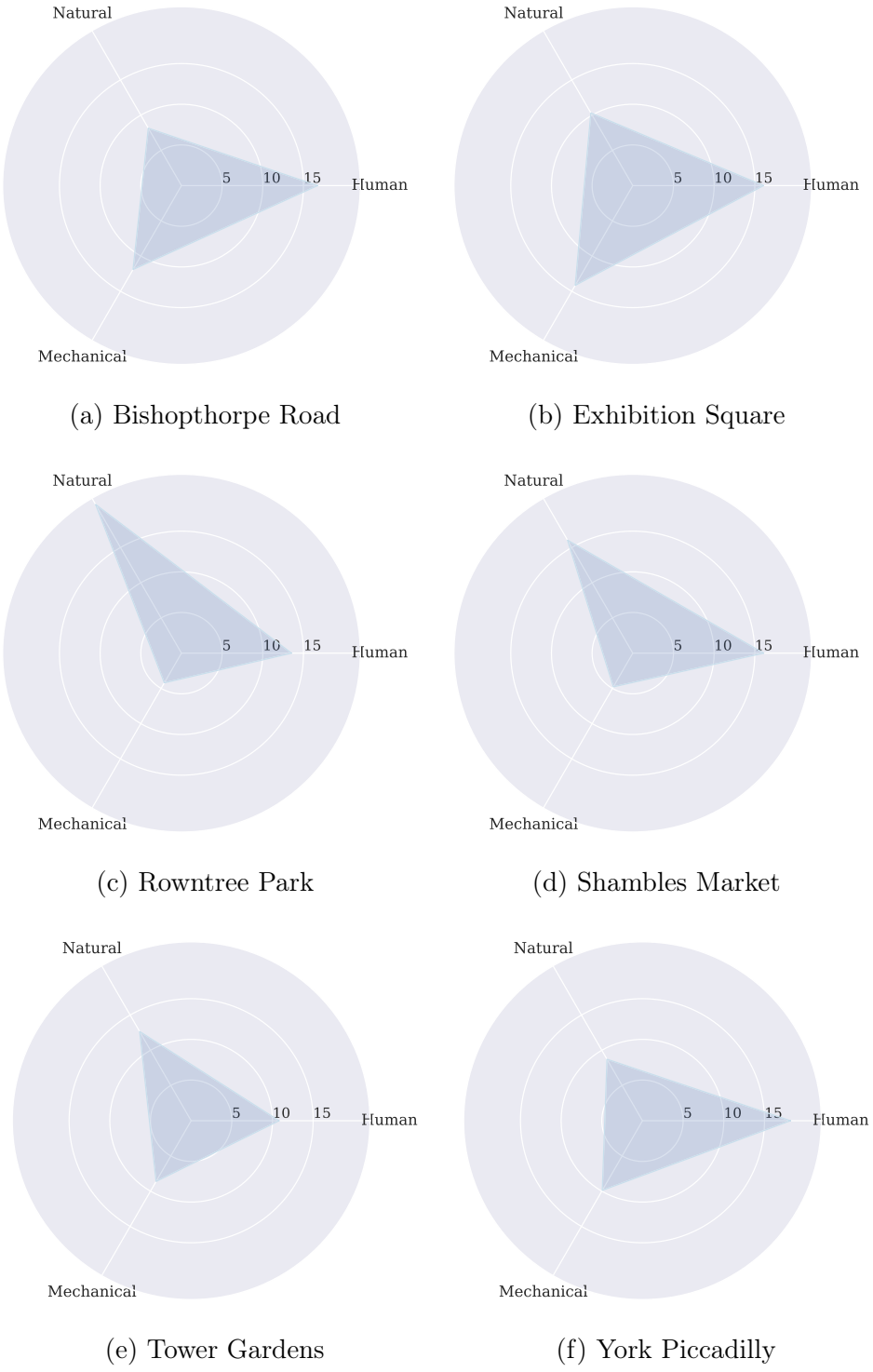


Figure 8.7: NDSI/pleasantness values for each recording location.

as such contains a mix of natural and mechanical sounds. Its lower value relative to the park and market locations reflects this. The lowest values returned were for Bishopthorpe Road and Exhibition Square, which are both locations with heavy traffic. York Piccadilly has somewhat lighter traffic than these locations, and this is reflected in its slightly higher pleasantness rating.

These results go some way to confirming the effectiveness of the Core ML model, though there seems to be a bias towards the upper end of the scale, with locations where mechanical sound dominates rated more towards the middle than might be expected. Exploring the individual probability readings (included in Appendix C) in detail reveals that the mean mechanical score overall was 8.93, whereas the mean natural rating was 12.87. Given the mixture of locations chosen, one would expect these to be more similar, which suggests the Core ML model output requires some calibration.

Figure 8.8 shows the three one-minute average human/natural/mechanical ratings given by the Core ML model for each tested location, with no virtual objects



(a) Bishopthorpe Road

(b) Exhibition Square

(c) Rowntree Park

(d) Shambles Market

(e) Tower Gardens

(f) York Piccadilly

Figure 8.8: Radar plots showing the ratings (percentage probability) returned for each location with no active virtual objects.

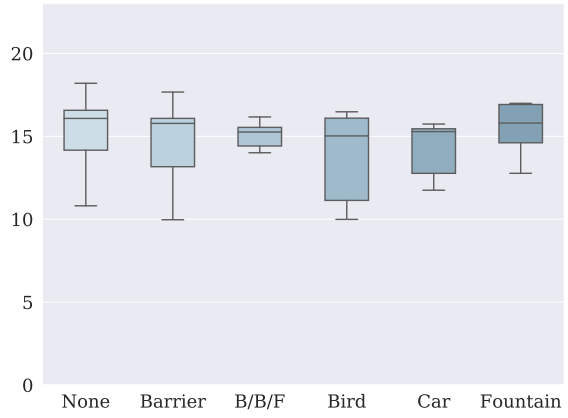
added, as a radar plot. Though smaller than the variations in natural or mechanical ratings, there are visible variations in the human ratings which show that it does give additional information beyond the two poles of the NDSI/pleasantness metric. For instance, Exhibition Square (Figure 8.8(b)) has a very similar rating for human sound to Shambles Market (Figure 8.8(d)), whereas their other ratings are very different. The variance in human ratings is markedly smaller, at 6.99, than that of mechanical ratings (15.35) or natural ratings (24.79). It is not clear whether this reflects a real phenomenon whereby the variation in levels of human sound are actually small at the locations tested, or whether this is a flaw in the classifier that requires alternative training. It should be noted that the classifier used to provide human ratings represented more of a compromise than those used for natural or mechanical sounds, as detailed in Section 8.2.1.

These results show that generating NDSI/pleasantness values using ML-based estimates for the relative contributions of natural and mechanical sounds can produce plausible results. Although in need of calibration, the trend of figures produced by the model matches the characteristics of the locations. This suggests that a ML approach to the problem of calculating meaningful soundscape indices could be effective and this app shows this can be incorporated into a handheld device that is as simple to use as an SPL meter. It should be noted that results in this regard will likely be somewhat contingent on the hardware/software platform used, as reflected in the study of SPL metering apps presented in [231].

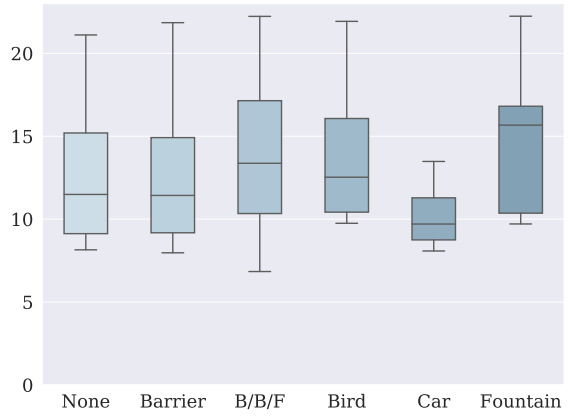
These results also show that the BusyStreet and Woodland classifiers trained on EigenScape data are generalisable to audio not contained within the EigenScape dataset. The classifiers in this study are working with audio recorded using the ASH microphones rather than an Eigenmike, a mismatched input condition.

8.4.2 Effect of Virtual Objects

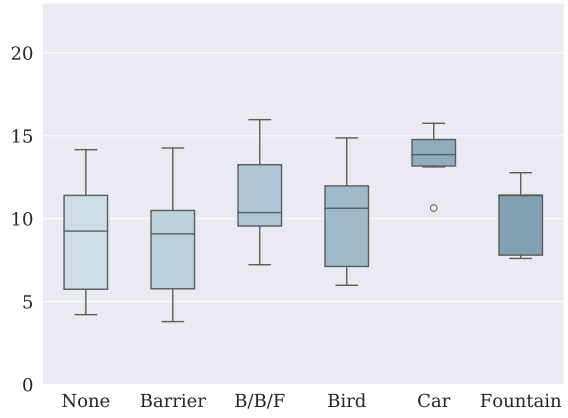
The distributions of the three sound category ratings for all locations under each virtual object activation condition are shown in Figure 8.9. A D'Agostino's K^2 test



(a) Human ratings



(b) Natural ratings



(c) Mechanical ratings

Figure 8.9: Distributions of ratings (percentage probability) across all locations for each virtual object condition.

[232] indicates normal distributions for all three sets of ratings.

It can be seen from Figure 8.9(a) that the median human ratings are all around 15% and the addition of any virtual objects has very little effect. Repeated measures analysis of variance (ANOVA) [233] confirms no significant effect of adding any object $F(4, 20) = 1.49, p = 0.24$. Figure 8.9(b) shows that the natural ratings are more spread than the human ratings generally, with the addition of the car object resulting in a reduction both in IQR and of median rating. Adding the fountain object increases the median natural rating, but there is no reduction in IQR, indicating that the effect of adding the fountain is not uniform, having more of an effect in some scenes than others. Repeated measures ANOVA indicates that adding objects has a significant effect on the natural ratings $F(4, 20) = 5.06, p < 0.05$. The effect of each individual condition was tested post-hoc using bonferroni-corrected paired t-tests [234], but these showed no individually significant contributors.

The largest effect by an individual object was that of the car on the mechanical ratings, as shown in Figure 8.9(c). The median rating increases from 8.93 with no objects added to 13.69 when the virtual car object is active. Repeated measures ANOVA shows this effect is significant $F(4, 20) = 10.79, p < 0.05$, and post-hoc testing shows that the car object individually has a significant effect on the ratings $t(5) = 4.24, p < 0.0125$, but there are no significant effects from any other object.

Despite the limited amount of data obtained, there is a strong indication that, whilst adding a virtual car causes an increase in mechanical ratings with a corresponding decrease in natural ratings, the effect of the fountain object is much more modest, and neither the bird nor barrier objects seem to have much effect at all. The pronounced effect of the car reflects the findings presented by Stevens [80], where subjects in a listening test rated a sound scene recorded by a lake as much more mechanical given the added presence of only a single car. This gives some evidence that the Core ML object is returning ratings that are to a degree aligned with human perception, though more study would be needed to corroborate this.

It is possible the lack of effect from the bird object is a result of the fact that

birdsong is intermittent as opposed to the consistent noise from the car and water fountain objects (see Figure 8.4). The lack of impact from the barrier object suggests that the barrier, either in principle, or in the rudimentary implementation used here, is ineffective. It is possible that this is due to the fact that the barrier attenuates mainly higher frequencies, but since MFCCs have reduced resolution at higher frequencies (see Figure 4.3) it is likely that these are less discriminative to the models than low frequencies, which are largely unaffected by the barrier. The effect of the barrier might therefore be more apparent using an alternative classifier, or in human listening tests.

The lack of any effect on the human ratings by any virtual object is possibly due to the fact that none of the virtual objects could be categorised as human sound sources. A virtual 'conversation' object could have been a good addition to the app in this regard.

8.5 Summary

This chapter has presented a practical real-time implementation of an acoustic environment monitoring system, drawing together ideas and techniques explored in the previous chapters of this thesis. It was shown that the app produces plausible estimates for the relative prevalence of mechanical and natural sounds, which can be used to calculate a *pleasantness* rating based on the NDSI soundecology metric. NDSI/pleasantness values calculated for the six locations tested matched well with the relative contributions of each sound category at those locations, though there was some bias towards the upper end of the scale, indicating that further calibration of the model is required. The classifier used to return ratings for human sound provided much less useful insight. To reliably calculate an *eventfulness* metric to complement *pleasantness* (based on the chart in Figure 3.3), a better model would be required. It is possible that better training of classifiers, perhaps by inclusion of more diverse training data, would help a great deal in this regard.

With the exception of the car object, the effect of adding virtual sound sources appeared negligible. There is some evidence that adding the fountain object improved natural ratings at some locations, but this was not a significant effect. The car object did significantly increase the mechanical ratings at most locations. In general it is thought that these objects might have more of an effect on ratings given more sophisticated sound spatialisation than the present implementation.

The next chapter will conclude this thesis, drawing together and summarising the work from all previous chapters and the contributions of this work to the wider field.

9 | Conclusion

This thesis has presented a body of work with the overarching goal of deriving novel methods of environmental sound monitoring inspired by the soundscape approach. The EigenScape database of spatial acoustic scene recordings was created as a basis for this research, and approaches to working with this data have combined aspects from various machine listening and spatial audio technologies. Results from these experiments have produced insights not only into the methods themselves, but also the nature of the recorded sound fields used and the patterns of similarity between these. This final chapter will summarise the work within the previous chapters, before restating the initial hypothesis given in Chapter 1, and reflecting on the results obtained in light of that hypothesis. This will be followed by some ideas for the possible next stages of research for each of the sub-projects contained in this work, and finally some concluding thoughts.

9.1 Summary

This thesis began in Chapter 2 with an exploration of the fundamentals of acoustics. There was a particular emphasis on the propagation of sounds in space. This subject is very relevant to the concept of sound fields central to the work in this thesis, both in terms of acoustic environments, and also the spherical harmonic-based Ambisonic audio format used to record them. This was followed in Chapter 3 by detail on the soundscape approach to environmental sound monitoring, and how ideas developed in this field regarding the perceptual shortcomings of the prevailing environmental

noise approach provided the primary motivation for this thesis. Chapter 4 began by covering in detail the algorithms behind the machine learning technologies that were used in this work, and concluded with descriptions of some of the previous machine listening approaches that inspired those developed as part of the research in this thesis.

Chapter 5 marked the beginning of the original work, describing the methods used for collecting the EigenScape database that provides the foundation for the research that follows. This included detail on the recording equipment and a description of the scene classes chosen, with reference to the already-available databases that provided inspiration in this regard. Specific recording locations were detailed, with further information, maps, and imagery included in Appendix A.

The EigenScape data was first used for an investigation into acoustic scene classification using spatial features, described in Chapter 6. Features capturing spatial properties of the EigenScape recordings were extracted from first-order Ambisonic channels using DirAC, and from the higher-order channels using spherical harmonic beamforming and COMEDIE diffuseness estimation. These features were used to train separate GMM classifiers. Accuracy obtained using various subsets of spatial features compared favourably to those obtained using traditional MFCC features. This was the first confirmation that acoustic scene recordings can be characterised by the spatial distribution of the sounds within them as well as their spectral content. Differences in specific misclassifications between the spectral and spatial classifiers also provided some indication that the two sets of features could in fact be complementary, a finding reinforced by further work using these features with CNN classifiers, conducted in collaboration with colleagues at Tampere University.

Given these positive initial results, the next step of the work, presented in Chapter 7, investigated using spatial features to estimate the directions and movement trajectories of individual sound sources within an acoustic scene. Where the work in Chapter 6 used features extracted from only the first and fourth-order Ambisonic channels, the source-tracking work in Chapter 7 used features extracted at all four

available orders. Results from this work indicated that second-order Ambisonics is a good level of compromise between sophistication, cost of equipment, and system performance. Whilst third and fourth-order Ambisonics did yield further improvements in performance, these were small relative to that seen between first and second-order.

The final research, presented in Chapter 8, represented a more applied approach than the preceding chapters. Whereas Chapter 7 was concerned with development of a method whereby the minutiae of individual sounds in an acoustic environment could be monitored, the work in Chapter 8 returned to the principles of soundscape perception and soundecology outlined in Chapter 3. A mobile application was developed utilising machine learning to generate high-level sound scene descriptors based on the soundscape approach. The application also incorporated AR technology to allow users to place virtual sound objects into their environments. It was shown that the model incorporated into this app was capable of producing plausible metrics based on estimates for the relative proportions of natural and mechanical sounds making up a scene. Estimates produced for the proportion of human sound were somewhat less reliable, and, aside from the AR car object, the effects of added virtual sound sources were small.

As a whole, the work in this thesis presents a coherent portfolio of methodologies for reliably monitoring acoustic environments, following ideas from the soundscape approach. Chapters 6 and 7 showed how spatial audio techniques may be used to enhance our understanding of acoustic environments, and point to techniques that might be developed for future environmental sound monitoring practice. Whilst the results of Chapter 7 indicated that second-order Ambisonics might provide a sufficient spatial resolution to use for these purposes, this would still require highly specialised and potentially costly equipment relative to that used for traditional noise level monitoring. To this end, Chapter 8 presented an example of a practical system for deriving perceptual soundscape metrics that runs on standard smartphone hardware.

9.2 Contributions to the Field

The unique contributions to the fields of machine listening and soundscape research made by this thesis are as follows:

- The EigenScape database, presented in Chapter 5, which is the largest publicly available collection of spatial acoustic scene recordings. Previous databases of this size were limited to mono or stereo audio only. The available spatial databases were all small in scope, with some using non-standard microphone array formats, limiting their utility beyond very specific applications. The format of eight examples of eight classes, each of exactly the same recording length, was designed specifically to make balanced partitioning and segmentation of the data straightforward. Use of the Ambisonic format for EigenScape provides a standard basis, so algorithms developed using this audio can be used on any future recordings encoded to Ambisonic format, which does not require a specific microphone array layout. The fourth-order Ambisonic spatial resolution also far exceeds that of any previously-published dataset.

- The results from the work in Chapter 6 were the first to prove that it is possible to characterise acoustic environments based on their spatial properties in addition to their spectral properties. Most interestingly, it was indicated that spatial similarity and spectral similarity do not always coincide, which suggests that both spatial and spectral information are needed for a full understanding of acoustic environments. No previous ASC systems had been developed using Ambisonic features, and no previous systems used spatial features exclusively.

- Chapter 7 presented the first system to combine peak-finding in SRP maps with unsupervised clustering using DBSCAN. The approach of testing via optimisation using the tree parzen estimator, as a way to assess the robustness of the system to changing environmental conditions, was also novel.

- Results from the approach taken in Chapter 7, indicating that second-order Ambisonics represents a good point of compromise when it comes to the performance of a sound source tracking system, is an important finding in the investigation of spatial audio for acoustic environment monitoring.

- The method used in the Soundscape AR app to generate estimates for the contribution of human, natural and mechanical sounds based on ASC techniques is unique to this project, as is the use of these estimates to generate NDSI scores. The idea of re-contextualising the NDSI soundecology metric for the soundscape approach as *pleasantness* is also unique to the work presented in this thesis.

9.3 Restatement of Hypothesis

The hypothesis guiding this work, as originally stated in Chapter 1, was:

> *The monitoring of acoustic environments, with a view to deriving information useful to the soundscape approach, can be assisted using spatial audio analysis.*

The research presented in Chapters 6, 7 and 8 supports this hypothesis, as these chapters detail unique approaches to acoustic environment analysis that could not have been achieved without the use of spatial audio. Results from the ASC work in Chapter 6 show that the spatial properties of acoustic scenes are complementary to their spectral properties, and that a combination of both yields accurate classification of acoustic scenes even when using a simple GMM classifier. This chapter also gave some evidence that higher spatial resolution can yield better results, an observation confirmed and expanded upon by Chapter 7.

These confirmations of the original hypothesis demonstrate the value of this research project to the wider field, and the contributions outlined previously highlight the many areas in which this research has also been valuable beyond the specific

focus of this hypothesis. The findings of the various sub-projects making up this work have indicated potential ways in which this research may be continued in the future, and these ideas will be outlined in the following section.

9.4 Future Work

Perhaps the simplest way in which this project could be continued is in the expansion of the EigenScape database. Whilst containing a great deal of data, it is still small relative to the databases available in the field of image recognition. One of the strengths of the database is its use of the standardised Ambisonic format. This could help to enable a crowd-sourcing approach to its expansion, with researchers from around the world contributing using whatever microphone arrays are at their disposal, subject to a minimum spatial resolution. Given the results of Chapter 7, it is proposed that second-order Ambisonics could be the minimum recommendation for audio to be used in SELD research. An expanded database could lead to improved spatial ASC results.

The work in Chapter 6 focused on only first and fourth-order features as it was speculated that using all of the available higher-order information would make the point most clearly as to whether or not use of HOA features was beneficial. It was reasoned that any increase in performance using features derived from, for instance, the second-order channels, might be more modest than that achievable using all available channels, thus making any conclusions less definitive. It was also thought that discussing results across all four orders would make the work more confusing in presentation. In light of the sound event localisation results of Chapter 7, however, it would seem a prudent next step to investigate ASC performance for second and third-order Ambisonic features.

Another clear avenue for investigation would be to investigate the use of classifiers explicitly taking into account the temporal ordering of the spatial and spectral features, thus effectively completing the picture when it comes to the dimensions

along which sound fields vary. There have already been a number of systems successfully developed using recurrent neural networks [235, 236] and hidden Markov models [237, 238] for this purpose, and it would be interesting to observe how well these methods might integrate with the spatial audio features used in this thesis. One of the key points of failure for the ASC system described here is the very poor performance in classification of Beach scenes. As previously mentioned, this could be due to their spectral and spatial similarity to street scenes. Taking into account the temporality of scenes could help to disambiguate between scenes such as these.

Although it has now been shown that spatial features can be very useful in ASC, it remains the case that microphone arrays capable of recording in full 3D Ambisonic format are more cumbersome than those typically available on a handheld device. It would therefore be a worthwhile area of research to determine methods by which features similar to those used here could be derived from, for instance, binaural audio.

The obvious next step for the sound event localisation work in Chapter 7 would be the addition of a source labelling component, enabling the system to perform fully-fledged SELD according to the DCASE definition. It would, perhaps, be more interesting to test the system further in its current state and modify it so as to enhance its trajectory prediction functions before adding such a labelling component. Firstly, the system should be tested using more complex audio, as the synthesised anechoic scenes used to test it so far are not representative of a real-world sound field. As a start, new test audio could incorporate some reverberation or add more overlapping audio sources. Further to this, some limited movement of sound sources could be introduced and gradually increased to assess tolerance to increasing scene complexity. As detailed in Section 7.3.3, a potential point of failure in the system as it currently stands is in clustering when the movement of two sound sources causes them to overlap spatially. In this situation, it is likely that peaks from the two sources would be assigned to the same cluster, and the fact that the DBSCAN algorithm is not constrained to clusters of any particular shape could result in large

sprawling clusters where several sources overlap. This could perhaps be mitigated somewhat by retaining the frequency-dependence of the SRP maps (i.e. omitting the summation across k in Equation 2.26), and performing peak-finding separately in each frequency band. In this way, sound sources that overlapped spatially but not spectrally would remain separable.

There are several ways in which the *Soundscape AR* application introduced in Chapter 8 could be improved. First among these is the creation of a more bespoke model for estimation of human, natural, and mechanical sound prevalence than the repurposed ASC models used so far. Such a model could continue to use MFCC features or perhaps spatial features derived from binaural audio, but be trained using human ratings of the sound scenes obtained from listening tests as targets. This would also broaden the scope for including a much larger set of data in a way that would not require so many assumptions as were necessitated for this initial study. For instance, there may be many classes of acoustic scene that are rated as very highly natural by humans that are nevertheless quite sonically distinct. An example of this are the beach scenes, which sound very different to woodland scenes, but might be thought of as mostly natural. Human ratings for training data would properly validate or exclude the use of these recordings as combined training data for the 'natural' class.

Since calculating *pleasantness* based on natural/mechanical ratings was shown to work well, a new version of this system could aim to augment this with an *eventfulness* metric based on reliable ratings for human sounds, after the two-dimensional scheme identified by Axelsson *et al.* [76]. Rather than showing the raw ratings, a future version of the app could use a graphic visualisation similar to the circular *pleasantness/eventfulness* chart shown in Figure 3.3 so that the user could see the perceptual placement of a scene in this space at a glance. Given the limited results when considering the effects of virtual objects on these ratings, it is felt that the analysis side of the app might be best served in the future by decoupling it from the AR component. Separately, however, the AR component presents an exciting

opportunity for future work. Recent updates to ARKit [92] enable the placement of AR objects to be made persistent between sessions, and 'location anchors' allow the placement of objects at specific geographic locations to be shared between users of different devices. These features combine to offer the opportunity for a researcher to design and conduct an AR soundwalk featuring a number of AR objects, consistently displayed to multiple subjects over an extended area. This could be done simultaneously with a control group experiencing the area as it is in reality with no AR components. Such a study could produce valuable insights into human responses to proposed environmental alterations in-situ.

9.5 Closing Comments

This thesis has presented a number of novel methods using modern machine learning techniques and spatial audio to derive information useful for the soundscape approach. At the very least, this work has made clear the great wealth of information present in acoustic environments that does not survive the bottleneck that is L_{Aeq} measurement.

Most of the methods presented are reliant on highly specialised spatial recording equipment, the expense of which makes widespread adoption of these specific algorithms unlikely as they stand presently. It is hoped, however, that this research points to practical ways in which spatial audio and the soundscape approach might be incorporated into wider practise moving forward, and that ideas from this thesis might be incorporated into the practical sound monitoring systems of the future.

Further study along these lines could lead to a complete methodology incorporating the spectral, spatial and temporal dimensions of a sound scene into a thorough holistic understanding. Such deep knowledge of acoustic environments could lead initially to highly targeted acoustic interventions. Instead of building large unsightly noise barriers or imposing blanket noise reduction policies, specific annoying noise sources could be addressed alongside programmes to encourage positive sounds.

Rather than viewing all human sound as essentially negative, as has too often been the case, it could be acknowledged that the sounds of certain human activities can do a great deal to enhance wellbeing, with a certain amount of music-making and lively conversation positively encouraged. Methods such as the AR sound walk could be the last stage in an urban planning chain incorporating acoustic environment monitoring, modelling, and interpretation according to soundscape approach principles, at every step of the process. This could remake our urban environments over time into more pleasant places to be, and would be much better than the present approach that often amounts to little more than damage limitation.

Indiscriminate suppression of that which we disapprove is an attitude that permeates far beyond the realm of environmental sound. It is hoped that the idea of sound as a resource, rather than a waste, might help to create a change in attitude that could eventually filter down into the psyche of the general population. If this could result in an increased awareness of the pulse of life, perhaps people might be encouraged to slow down and listen more, a practice of incalculable value.

A | EigenScape Metadata

A.1 File Details

The EigenScape database is available online at `https://zenodo.org/record/1284156#.XxBZcJO2n0c` and is organised as follows:

Filename	Size
Beach.zip	14.3 GB
BusyStreet.zip	14.9 GB
Park.zip	15.0 GB
PedestrianZone.zip	14.3 GB
QuietStreet.zip	13.8 GB
ShoppingCentre.zip	14.7 GB
TrainStation.zip	14.7 GB
Woodland.zip	13.2 GB
Lite-EigenScape.zip	12.6 GB
Metadata-EigenScape.csv	4.9 kB

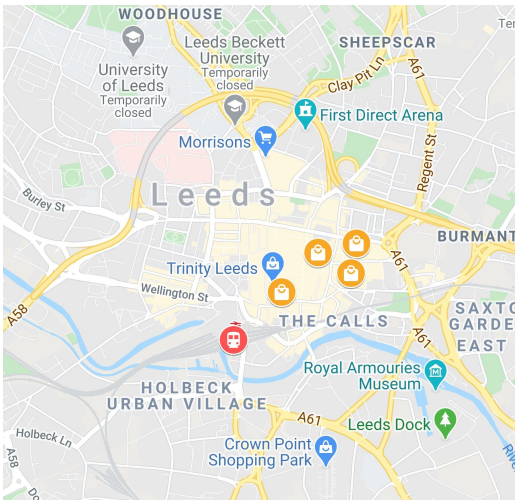
The first 8 ZIP files contain the full fourth-order Ambisonic WAV audio for the classes as named, whilst Lite-EigenScape.zip contains the complete dataset in first-order FLAC format. The metadata csv file contains a full list of recordings, and is here reproduced in full:

Filename	Location	Time	Date	Additional Information
Beach-01.wav	Bridlington Beach	10:42 am	9 May 2017	
Beach-02.wav	Filey Beach	11:48 am	9 May 2017	
Beach-03.wav	Cayton Bay	12:46 pm	9 May 2017	
Beach-04.wav	Redcar Beach	11:23 am	10 May 2017	
Beach-05.wav	Saltburn Beach	12:19 pm	10 May 2017	
Beach-06.wav	Sandsend	2:00 pm	10 May 2017	
Beach-07.wav	Whitby West Cliff	3:59 pm	10 May 2017	
Beach-08.wav	Robin Hood's Bay	5:08 pm	10 May 2017	
BusyStreet-01.wav	Wilbraham Road Manchester	9:28 am	5 May 2017	
BusyStreet-02.wav	Oxford Road Manchester	11:09 am	5 May 2017	
BusyStreet-03.wav	London Road Manchester	4:24 pm	5 May 2017	
BusyStreet-04.wav	Bishopthorpe Road York	1:39 pm	11 May 2017	Slight clipping from very loud car.
BusyStreet-05.wav	Micklegate Bar York	2:03 pm	16 May 2017	
BusyStreet-06.wav	Southgate Huddersfield	12:48 pm	18 May 2017	
BusyStreet-07.wav	Trinity Street Huddersfield	2:17 pm	18 May 2017	
BusyStreet-08.wav	St. Leonard's Place York	11:44 am	23 May 2017	
Park-01.wav	Rowntree Park York	2:00 pm	11 May 2017	
Park-02.wav	Greenhead Park Huddersfield	1:55 pm	18 May 2017	
Park-03.wav	Yorkshire Museum Gardens	3:52 pm	18 May 2017	
Park-04.wav	Scarcroft Road Park York	4:39 pm	18 May 2017	
Park-05.wav	West Bank Park York	9:44 am	23 May 2017	
Park-06.wav	Homestead Park York	11:05 am	23 May 2017	
Park-07.wav	Hull Road Park York	12:46 pm	23 May 2017	
Park-08.wav	Heslington Church Green York	3:13 pm	23 May 2017	
PedestrianZone-01.wav	Clayton Square Liverpool	11:40 am	1 May 2017	Original DIY windjammer.
PedestrianZone-02.wav	Church Street Liverpool	3:04 pm	1 May 2017	Original DIY windjammer.
PedestrianZone-03.wav	Shambles Square Manchester	1:23 pm	5 May 2017	
PedestrianZone-04.wav	Market Street Manchester	1:44 pm	5 May 2017	
PedestrianZone-05.wav	Church Street Whitby	3:17 pm	10 May 2017	
PedestrianZone-06.wav	St. Helen's Square York	3:28 pm	11 May 2017	
PedestrianZone-07.wav	Minster Yard York	4:05 pm	11 May 2017	Heavy wind.
PedestrianZone-08.wav	Stonegate York	4:40 pm	11 May 2017	
QuietStreet-01.wav	Chatfield Road Manchester	9:06 am	5 May 2017	
QuietStreet-02.wav	Thomas Street Manchester	3:54 pm	5 May 2017	
QuietStreet-03.wav	Matley Lane Hyde	2:24 pm	7 May 2017	
QuietStreet-04.wav	Church Lane York	11:07 am	11 May 2017	
QuietStreet-05.wav	Main Street York	11:28 am	11 May 2017	
QuietStreet-06.wav	St. Benedict Road York	2:26 pm	16 May 2017	
QuietStreet-07.wav	Windmill Rise Corner York	10:09 am	23 May 2017	
QuietStreet-08.wav	Holmefield Lane York	2:42 pm	23 May 2017	
ShoppingCentre-01.wav	St. John's Market Liverpool	11:15 am	1 May 2017	No windjammer (indoor).
ShoppingCentre-02.wav	Arndale Centre Manchester	12:56 pm	5 May 2017	
ShoppingCentre-03.wav	York Designer Outlet	4:26 pm	8 May 2017	
ShoppingCentre-04.wav	Victoria Gate Leeds	12:04 pm	15 May 2017	Microphone stand held.
ShoppingCentre-05.wav	Victoria Quarter Leeds	12:25 pm	15 May 2017	Microphone stand held.
ShoppingCentre-06.wav	Leeds Kirkgate Market	12:56 pm	15 May 2017	
ShoppingCentre-07.wav	Trinity Leeds	2:09 pm	15 May 2017	
ShoppingCentre-08.wav	Packhorse Centre Huddersfield	1:11 pm	18 May 2017	
TrainStation-01.wav	Liverpool Lime Street Station	10:34 am	1 May 2017	Original DIY windjammer.
TrainStation-02.wav	Manchester Oxford Road Station	10:46 am	5 May 2017	Microphone stand held.

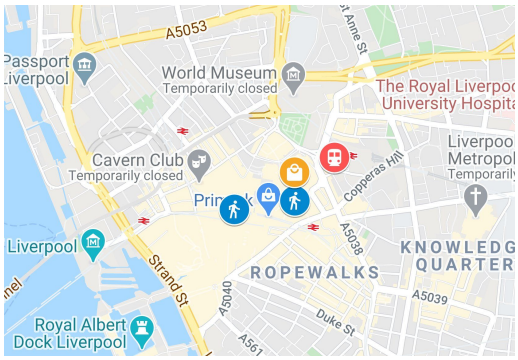
TrainStation-03.wav	Manchester Victoria Station	2:15 pm	5 May 2017	
TrainStation-04.wav	Leeds Station	10:50 am	15 May 2017	
TrainStation-05.wav	Manchester Piccadilly Station	10:31 am	16 May 2017	
TrainStation-06.wav	York Station	10:54 am	18 May 2017	
TrainStation-07.wav	Huddersfield Station	12:13 pm	18 May 2017	
TrainStation-08.wav	Scarborough Station	9:50 am	19 May 2017	+5 dB gain.
Woodland-01.wav	Knavesmire Wood York	11:41 am	8 May 2017	
Woodland-02.wav	Acomb Wood York	12:26 pm	8 May 2017	
Woodland-03.wav	Westfield Wood York	1:30 pm	8 May 2017	
Woodland-04.wav	Hagg Wood York	2:59 pm	8 May 2017	
Woodland-05.wav	Dalby Forest	2:36 pm	9 May 2017	
Woodland-06.wav	Dalby Forest Lake	3:09 pm	9 May 2017	
Woodland-07.wav	Pickering Castle Woods	4:10 pm	9 May 2017	Steam train whistle.
Woodland-08.wav	Rowntree Park Woods	2:22 pm	11 May 2017	

A.2 Additional Maps and Images

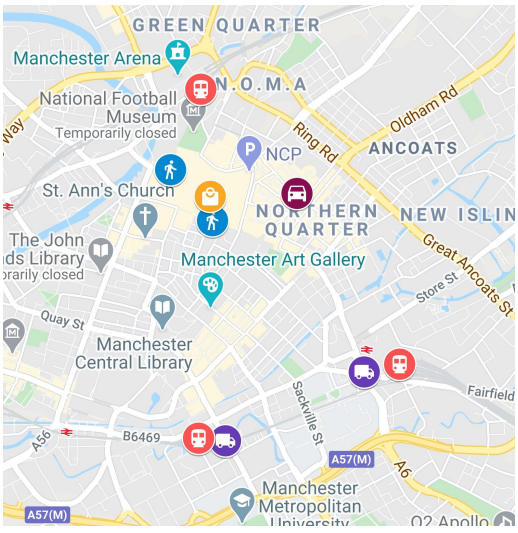
(a) Huddersfield



(b) Leeds



(c) Liverpool



(d) Manchester

Figure A.1: Detail of EigenScape recording locations in various city centres.

(a) BusyStreet-04

(b) QuietStreet-07

(c) ShoppingCentre-02

(d) Woodland-01

Figure A.2: EigenScape recording setup in various location classes.

A.3 Documentation

This appendix contains documentation relevant to the planning of, and permissions obtained for, the recording of the EigenScape database. These are:

- Letter of support, including method statement.
- Risk assessment.
- Permission for Filming in Scarborough.

THE UNIVERSITY of York

DEPARTMENT OF
ELECTRONICS
Heslington, York YO10 5DD
Telephone +44 (0) 1904 322320
Telex 57933 YORKUL
Fax +44 (0) 1904 322335

Dr Damian Murphy
Direct Line 01904 323221
Email damian.murphy@york.ac.uk

4th May 2017

To Whom It May Concern,

EigenScape Field Recording Research Project- Method Statement;

Researcher: Marc C. Green

The EigenScape field recording equipment list is as follows:

- EigenMike microphone
- Samsung Gear 360 camera
- Apple MacBook laptop
- EigenMike interface & battery
- 1x Microphone stand
- Hard case

The recording will take place as follows:

First, the equipment will be set up (5 mins). The microphone stand will be set up to head-height and the microphone and camera will both be mounted to it. The recording equipment will be placed nearby next to this. This consists of a recording interface, laptop and small battery. Everything will be connected together using short cables. Care has been taken to design this setup to be as compact as possible and there will be no trailing cables or mains power required.

Recording will then take place (15 minutes). The microphone and camera will be activated and the researcher will state location and clap for synchronisation. The researcher will then stand nearby, allowing the equipment to get a good recording without interference. The equipment will record continuously for around 11 minutes to ensure at least 10 minutes of clean ambient audio is recorded. The researcher will confirm that audio has recorded properly before continuing.

Then the equipment will be packed away (5 minutes). This is a very quick process.

The whole process should take no longer than 30 minutes although recording may be interrupted by passers-by or an equipment error. There have been no equipment errors so far in testing and it is reasoned that a certain degree of conversation from passers-by would not be a completely unusual component of an urban soundscape, so this is acceptable.

As Marc's PhD supervisor I am happy to confirm that this work has been discussed and refined internally and approved by our Departmental Ethics Committee.

Yours sincerely

Dr Damian Murphy
University Research Theme Champion for Creativity
Reader in Audio and Music Technology

Department of Electronics
University of York

Risk Assessment for UK Fieldwork Activities

The degree of risk associated with most UK fieldtrips is normally considerably less than those overseas, where the environment is often less predictable and our knowledge of that environment poor. However, it is still important that an assessment is completed for all departmental fieldwork trips. By recording the assessment, you will be able to demonstrate that you have considered the safety of your field work activity in a serious and systematic manner.

The headings provided in the template below are provided for guidance, although their relevance may vary from activity to activity.

Proposed destination:	Liverpool city centre (1/5/17), Manchester city centre (4/5/17 – 5/5/17), Scarborough (2/5/17), Scarbourough borough (8/5/17 – 9/5/17), York city centre, Huddersfield, Leeds city centre, misc. other locations across North of England.	
Description of fieldwork activity:	Field recording of soundscapes at various types of urban and natural locations around the north of England. Equipment consists of a microphone and camera mounted to a single microphone stand with a laptop positioned on a box next to the base of the stand for recording.	
	Specific Assessment for Single Fieldwork Trip **No** **Generic Assessment Covering Similar Fieldwork Trips** **Yes**	
Risk Statement: Tick as appropriate	**LOW RISK:** The fieldwork activity presents a low risk, adequately controlled by following the **code of good fieldwork practice**. The assessment table (below) will be used to highlight: • the most significant hazards associated with the work • measures needed to reduce risks to an acceptable level	**Yes**
	HIGHER RISK: The fieldwork activity presents higher risks (e.g. fieldwork expeditions in remote places) requiring specific, more detailed planning & assessment. The assessment table (below) will be used to a) highlight key hazards associated with the work &, b) measures needed to reduce risks to an acceptable level.	**No**
Assessor:	**Marc C. Green**	
Group / Course Leader:	**Dr. Damian Murphy**	

Fieldwork Risk Assessment

The nature and complexity of the risk assessment will vary with the type of activity, and should therefore, be commensurate with the actual risk that the identified hazards pose in the particular circumstances. The table below should be used to highlight the key hazards associated with the fieldwork and measures needed to reduce risk to an acceptable level. Many UK based fieldwork activities present are low risk, adequately controlled by following the code of good fieldwork practices. However, even for low risk activities, it is still useful to highlight the most significant hazards and summarise how these will be managed. Remember, a risk assessment that is too cluttered with minor concerns will be discarded in the field as a bureaucrat's folly, and will devalue the whole process!

Hazards	**Comment / Detail** Identify & describe the nature of all **significant hazards** associated with the fieldwork & how harm could occur / harmful effect of the hazard identified	**Measures to prevent or minimize risk** Identify the **key control measures**, proportionate to level of risk (i.e. likelihood + seriousness of harm occurring), needed to reduce risks to a low & acceptable level	**Residual Risk Level** (See Appendix 1)		
			S	**L**	**RR**
Personal safety: e.g. driving to & from site, off-road driving, lone working, security of accommodation	• Train travel to/from sites. • Driving to/from sites. • Lone working during recording process. • Recording work occurring near roads.	• Care will be taken in busy stations. • Frequent breaks will be taken, with all routes rigorously planned in advance. • Detailed travel plans will be provided to supervisor so researcher's location will be known at all times. Mobile phone will be kept well-charged. • Extra care will be taken in busy street areas.	1	1	2
Physical: e.g. extreme weather, exposure to heat & sun, mountains, cliffs, caves, mines and quarries, forests / woods, freshwater, sea & seashore, marshes & quicksands, roadside, work at height	• Potential wet weather conditions. • Potential water hazards at coastal / riverside locations. • Hazards at railway and road locations.	• Little risk to researcher. Waterproof sheet will be carried to protect equipment. • Extra care will be taken to position equipment well away from spray and avoid slippery surfaces. • Work at coastal areas will be postponed in the case of bad weather. • Local authorities have been contacted and informed and high-vis jacket will be worn in urban locations. • Train station managements have been contacted and will supervise work on stations.	1	1	2

Hazards	Comment / Detail Identify & describe the nature of all **significant hazards** associated with the fieldwork & how harm could occur / harmful effect of the hazard identified	Measures to prevent or minimize risk Identify the **key control measures**, proportionate to level of risk (i.e. likelihood + seriousness of harm occurring), needed to reduce risks to a low & acceptable level	Residual Risk Level (See Appendix 1)		
			S	L	RR
Biological: e.g. dangerous animals, plants, pathogenic microorganisms (e.g. cause of Lyme disease (transmitted by ticks), Tetanus, Leptospirosis)					
Chemical: e.g. chemicals associated with the fieldwork activity, agrochemicals, dusts, chemicals on site					
Mechanical & Electrical: e.g. equipment, tools, machinery & vehicles	• Potential electrical hazards from equipment.	• No on-site mains power will be used. All batteries are relatively low-power.	1	1	2
Other Hazards:	• Trip hazard to members of the public from equipment.	• Equipment setup has been designed to avoid any trailing cables whatsoever. Sites will be inspected to find an appropriate place to set up well away from heavy footfall.	1	1	2

Supervised student fieldwork:
Describe arrangements ('safe system of work') for supervision of students during fieldwork activities:

Lone Working / Unsupervised fieldwork:

- Describe arrangements for maintaining contact between worker(s) and Group Leader / Supervisor
- Describe other arrangements that will be applied to safeguard fieldworkers (consider what means of communication will be used and who will be made aware of fieldwork itinerary in case of emergency (**see Appendix 3 'Field *Safety Emergency Contact Details***')

Describe the First Aid Arrangements:

Training / Instruction / Information:
All individuals involved in fieldwork trips must receive appropriate instruction / information on significant hazards and appropriate precautions necessary to reduce risk to a low and acceptable level. Appendix 2 should be used to record that appropriate information & instruction has been provided to all fieldworkers.

Group / Course Leader's Declaration:

- I will provide full safety instruction and information (**including written safety protocols where required**) for all those involved in the fieldwork activity (see Appendix 2 for record)
- I will provide appropriate supervision to enable work to be conducted within acceptable safety standards

Name	Signature	Date
Damian Murphy		**20/04/2017**

Assessment Review
Review and update the assessment when either a significant change to the work activity occurs, or when there is evidence that a review is necessary e.g. following an incident or accident

Appendix 2: Fieldwork Training Record

Sign off sheet for all participants (include emergency contact information?)

All those people taking part in the field work must sign below to indicate that they:

- have read the risk assessment and relevant sections of the departmental **code of good fieldwork practice**
- have been given adequate safety information & instruction relevant to their fieldwork
- have been provided with written safety instructions/documentation where relevant
- agree to abide by any restrictions identified
- will report any concerns they have as to the safety provision, training to the Group Leader or directly to the Departmental Safety Advisor

Name	Signature	Date
Marc Green		20/04/2017

Appendix 3: Field Safety Emergency Contact Details

To the fieldworker:
Please complete this form and hand the details to a responsible person of your choice. This person must be available to carry out the emergency contact procedure in the event that you do not return at the time stipulated below. ***NB: Please remember to telephone if you are likely to be delayed to prevent the emergency contact procedure from being initiated.***

To the holder of the contact details:
If the fieldworker has not returned after 2 hours from the estimated time of arrival then the following procedure must be followed:

1. Call the emergency services giving the fieldworker's itinerary, type of habitat to be visited and the University contact telephone numbers.
2. Call the University contacts to alert that there may be a problem.
3. If there has been an accident, call the fieldworker's next of kin.

Name/s of fieldworker/s: Marc C. Green

Itinerary: Travel to various locations to conduct field recording.

Estimated time of return:

Contact Details: Mr Marc C. Green / 5A Bishopthorpe Road, York. YO23 1NA
Fieldworker's mobile phone number: 07849 559662
University contact (e.g. Group Leader / HoD): Dr. Damian Murphy (supervisor).
Next of Kin (name / address / phone number): Mr. Stuart Green / 184 Victoria Street, Hyde. SK14 4DH / 07719 585840

Permission for Filming in Scarborough

Print this permission form and keep it with you whilst filming as evidence should you be asked to show proof of filming permission.

I confirm that the production company named below has applied for permission to film in the Borough of Scarborough.

Public liability received

Rowena Marsden

Rowena Marsden
Scarborough Borough Council/Film Co-ordinator. 01723 383 615 / 07967 465 327

Contact Name: Marc Green Position: PhD Student	Tel No: 01904 324227 Mobile No: 07849 559662
Email: marc.c.green@york.ac.uk	Application Date: 07/04/2017
Office Address: Genesis 6, Innovation Way, Heslington, York, YO10 5DQ	Production Company: University of York Audio Lab
Name of Production: EigenScape Dataset	Number of Crew: 1

Type of Operation

(Please tick as appropriate)

☐ **Feature** ☐ **Documentary** ☐ **Reality / Observation**

☐ **TV Drama** ☐ **TV Other** ☐ **Commercial** ☐ **Music Video** ☐ **Photoshoot**

☐ **Student film** ☐ **Short Film** ☐ **Corporate** **Other (please specify e.g. Wildlife…….)** Research Field Recording

Locations To Be Used	**Date**
Italian Gardens Valley Park Valley Road (near roundabout) Westborough pedestrian zone South Bay Beach Peasholme Park North Bay Beach	02/05/17
Filey Beach Cayton Bay Cayton Woods Oliver's Mount Woodlands	08/05/17

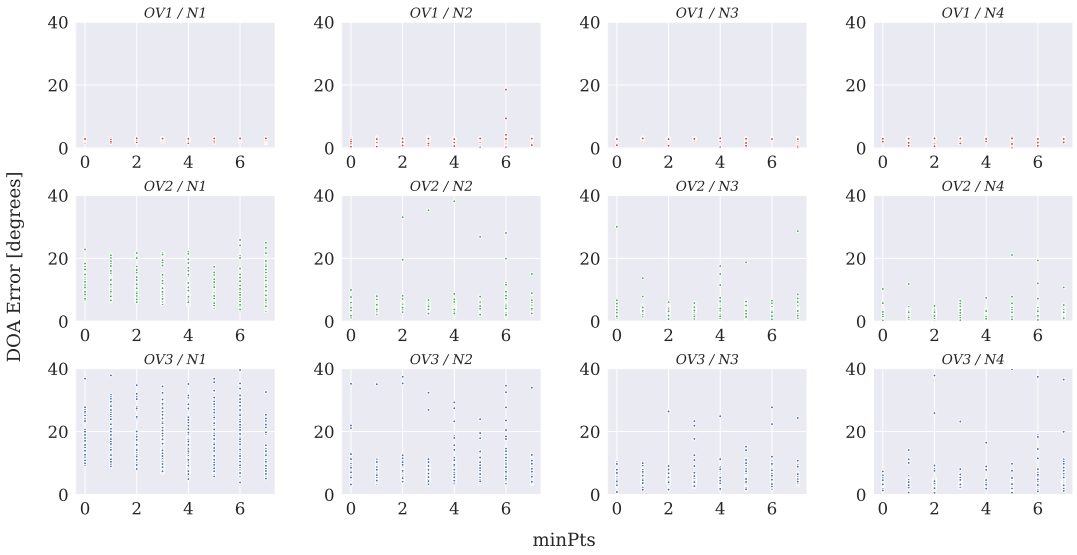
Robin Hood's Bay	09/05/14
Pannett Park, Whitby	
Bridge Street, Whitby	
Whitby Beach	
Runswick Bay	

Town Hall, St Nicholas Street, Scarborough, North Yorkshire YO11 2HG

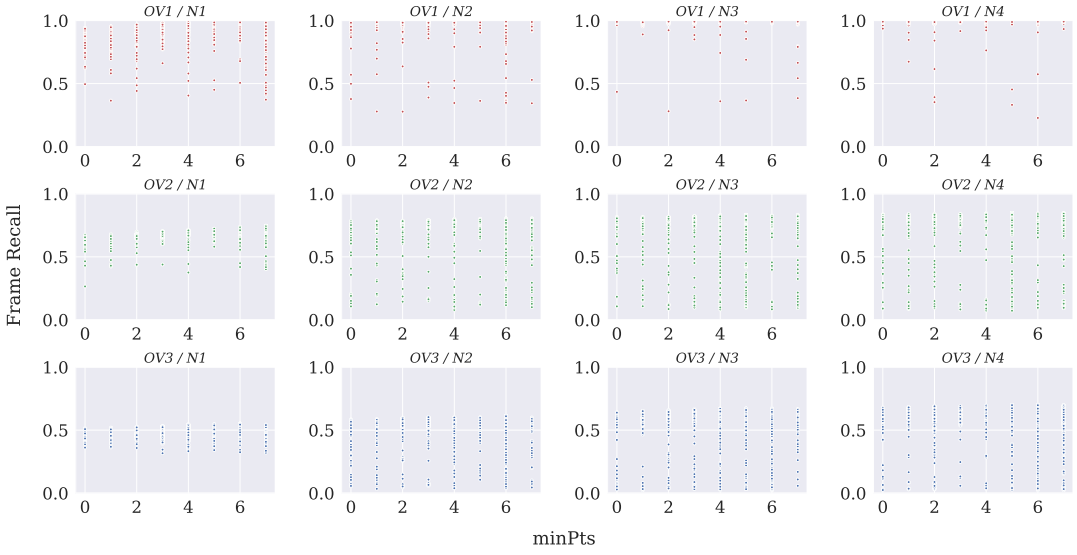
B | Source Tracking Parameter Charts

This appendix contains charts relating to Chapter 7, showing DOA Error and Frame Recall results for iterations of the new source tracking approach, varying with various input hyperparameters.

B.1 PWD

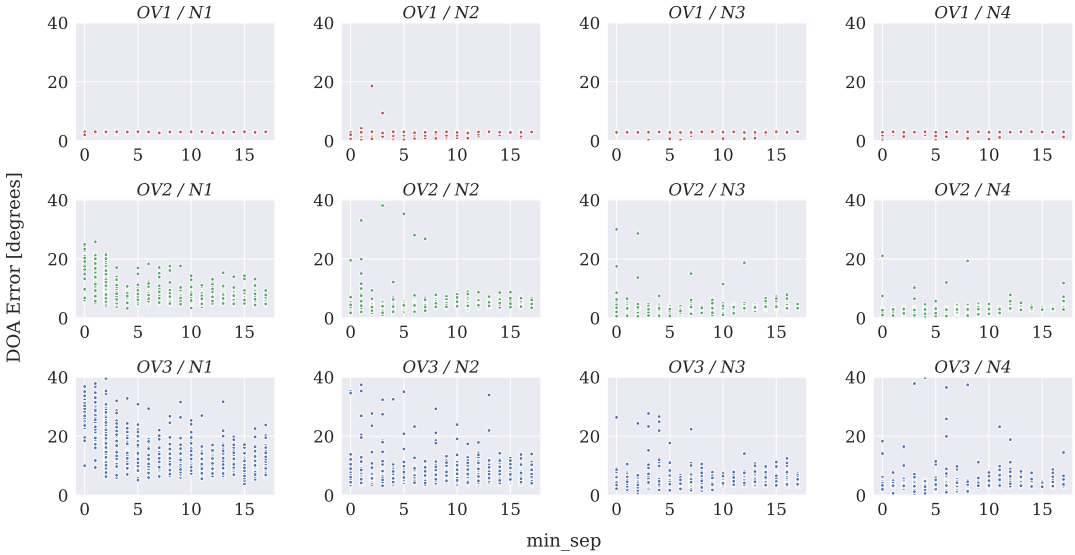


(a) DOA error values for all PWD iterations, varying with *MinPts*.

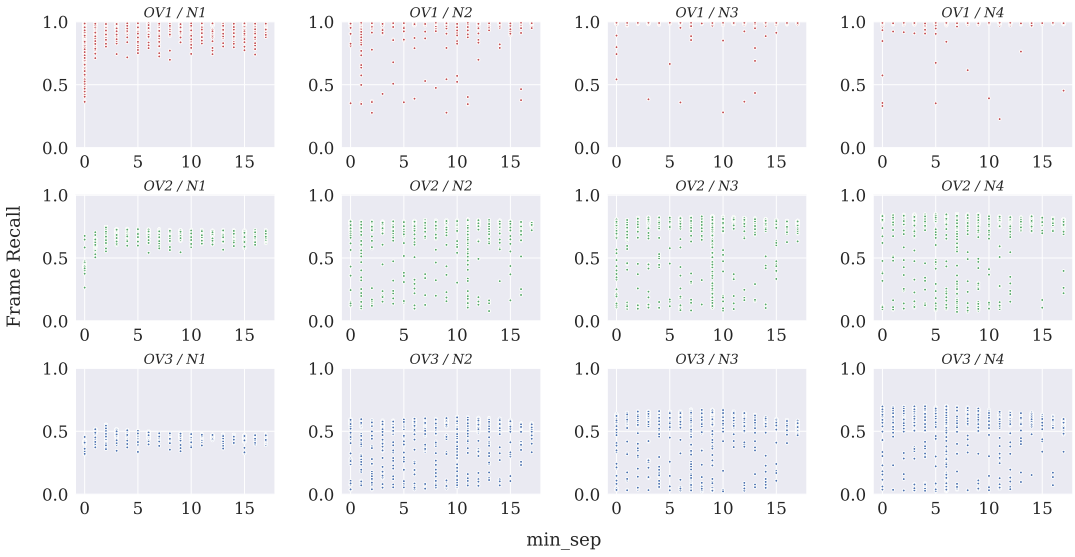


(b) Frame Recall for all PWD iterations, varying with *MinPts*.

Figure B.1: Performance metrics for iterations using PWD beams, varying with *MinPts*.



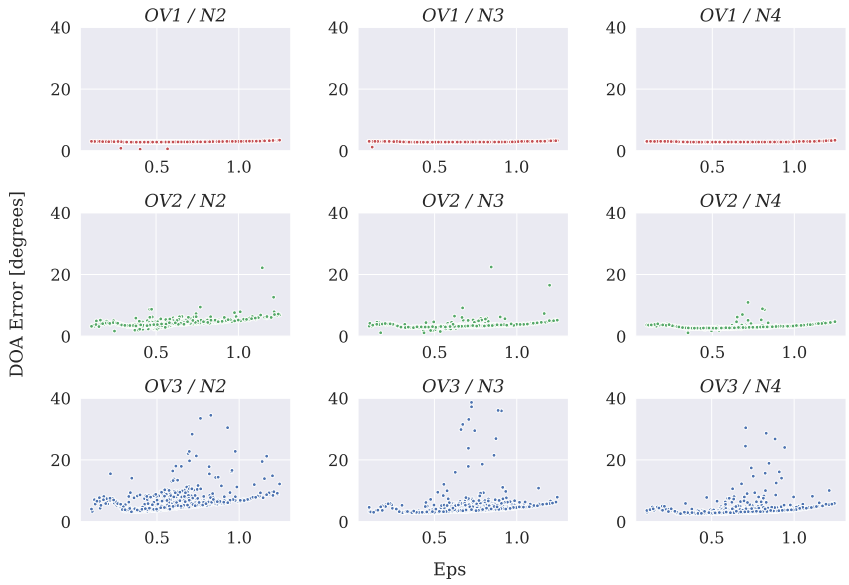
(a) DOA error values for all PWD iterations, varying with *min_sep*.



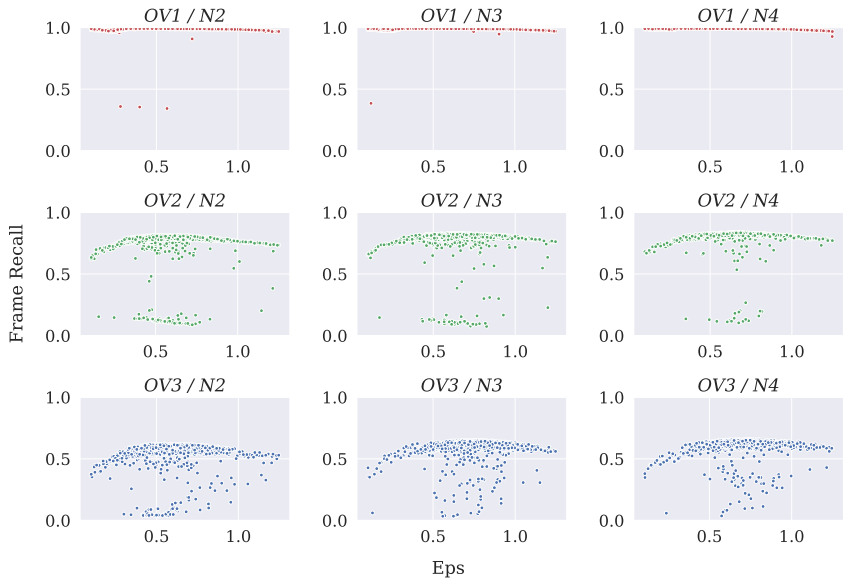
(b) Frame Recall for all PWD iterations, varying with *min_sep*.

Figure B.2: Performance metrics for iterations using PWD beams, varying with *min_sep*.

B.2 CroPaC

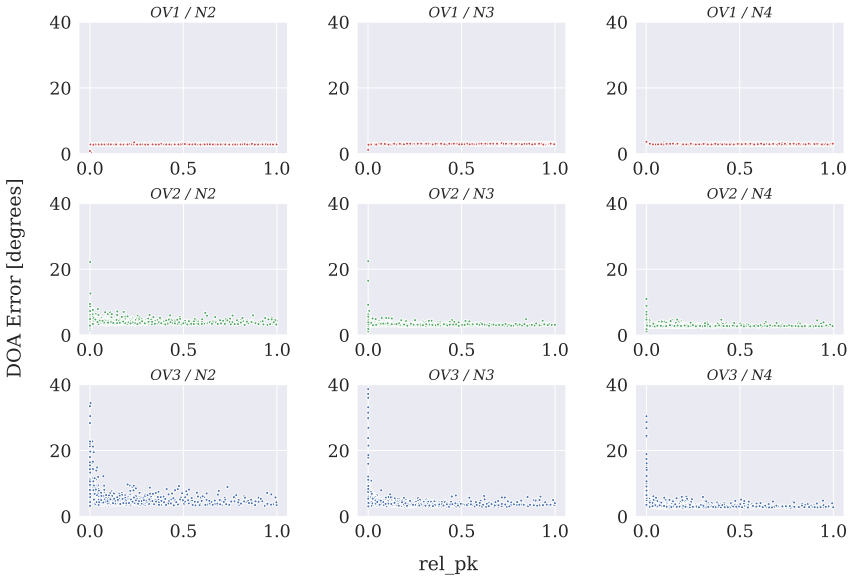


(a) DOA error values for all CroPaC iterations, varying with *Eps*.

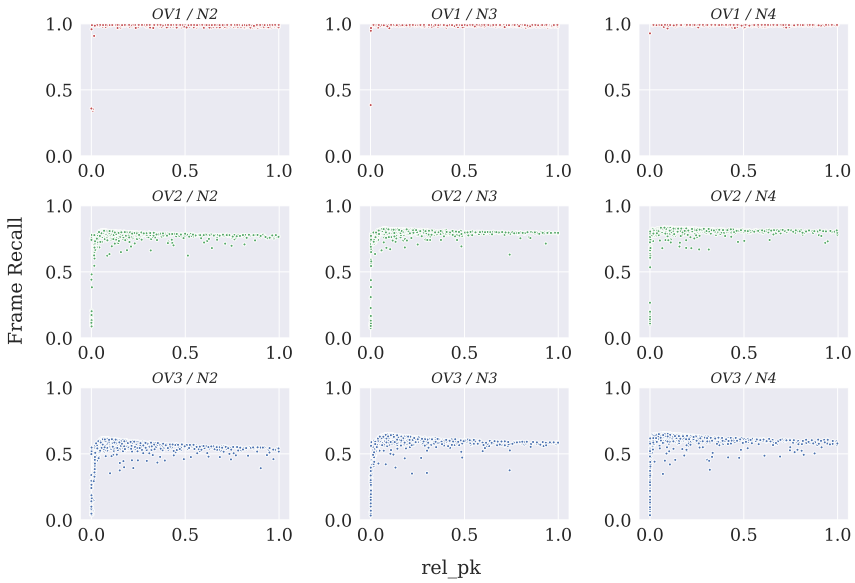


(b) Frame Recall for all CroPaC iterations, varying with *Eps*.

Figure B.3: Performance metrics for iterations using CroPaC beams, varying with *Eps*.

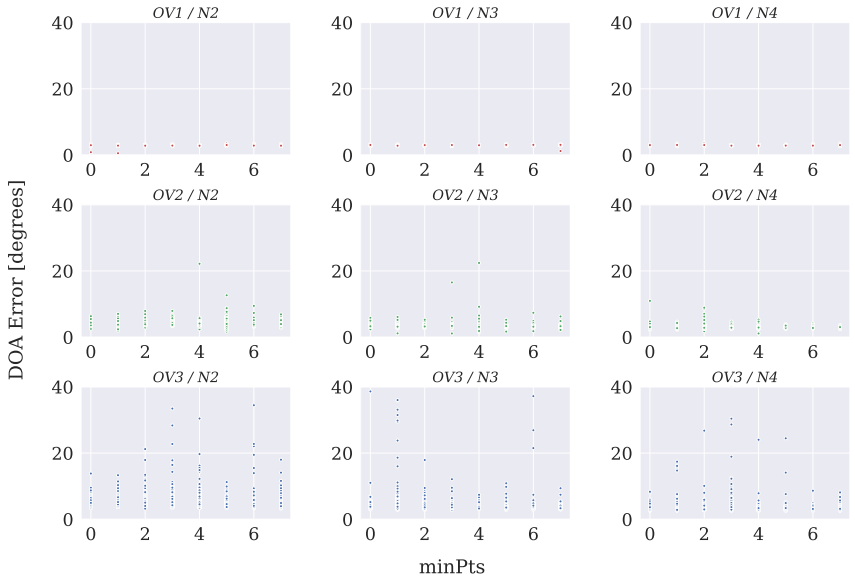


(a) DOA error values for all CroPaC iterations, varying with *rel_pk*.

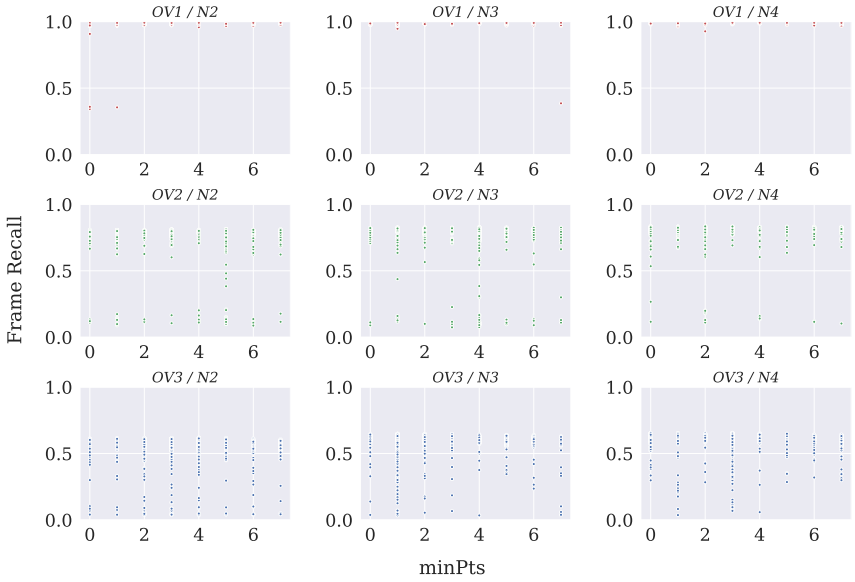


(b) Frame Recall for all CroPaC iterations, varying with *rel_pk*.

Figure B.4: Performance metrics for iterations using CroPaC beams, varying with *rel_pk*.

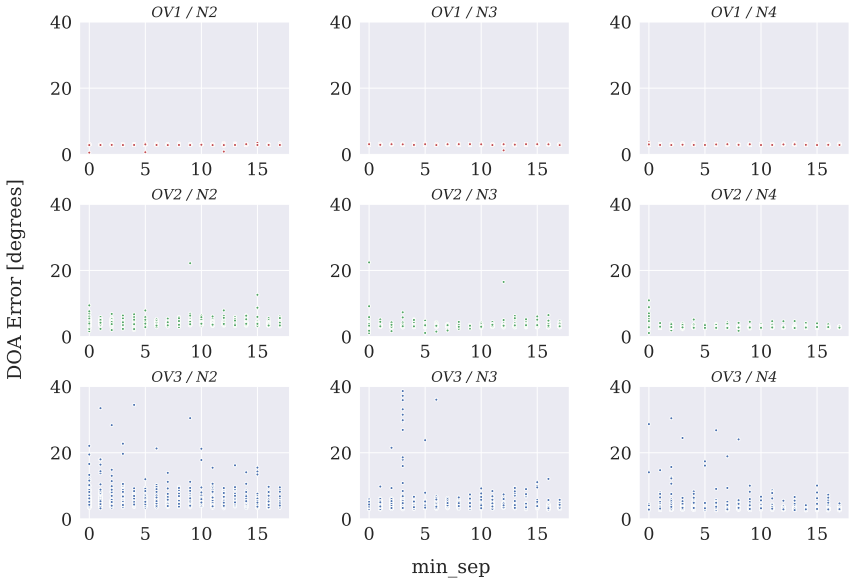


(a) DOA error values for all CroPaC iterations, varying with *MinPts*.

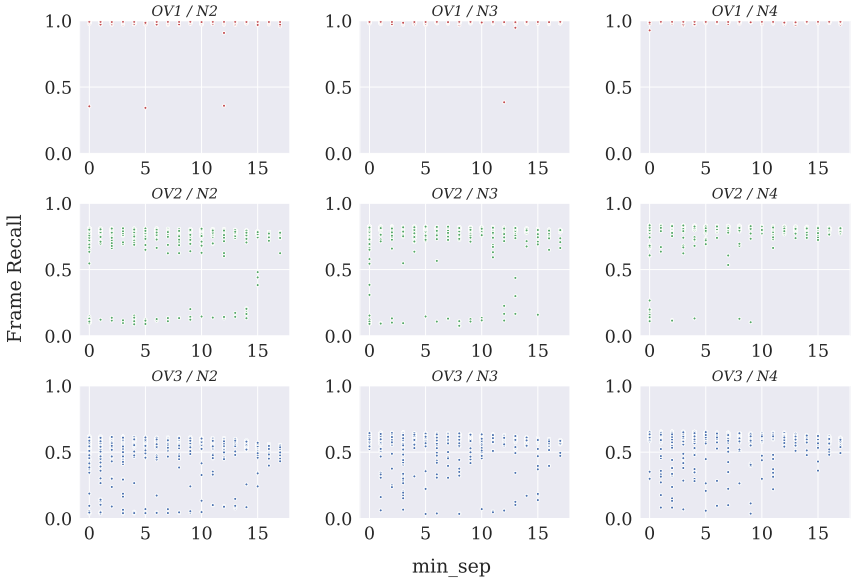


(b) Frame Recall for all CroPaC iterations, varying with *MinPts*.

Figure B.5: Performance metrics for iterations using CroPaC beams, varying with *MinPts*.



(a) DOA error values for all CroPaC iterations, varying with *min_sep*.



(b) Frame Recall for all CroPaC iterations, varying with *min_sep*.

Figure B.6: Performance metrics for iterations using CroPaC beams, varying with *min_sep*.

C | Soundscape AR Detailed Results

This appendix contains the complete one-minute average percentage probability ratings returned by the Soundscape AR app at each tested location under each virtual object condition.

Location	Condition	Human	Natural	Mechanical
Art Gallery	Clean	16.07	10.31	14.16
	Barrier	15.97	9.04	14.26
	Bird	16.48	9.75	14.87
	Fountain			
	Car	15.74	9.58	15.75
	Barrier/Bird/Fountain	15.45	11.35	14.18
Shambles Market	Clean	16.09	16.04	4.78
	Barrier	15.59	15.47	4.92
	Bird	16.14	16.94	5.98
	Fountain	15.80	16.81	7.80
	Car	15.19	11.77	10.64
	Barrier/Bird/Fountain	15.57	17.73	7.22
York Picadilly	Clean	18.20	8.73	9.88
	Barrier	17.67	7.97	10.71
	Bird	14.09	11.60	9.50
	Fountain	16.99	10.36	11.42
	Car	15.38	8.47	13.34
	Barrier/Bird/Fountain	15.07	10.00	10.47

Tower Gardens	Clean	10.81	12.67	8.62
	Barrier	9.97	13.27	8.32
	Bird	9.99	13.46	12.05
	Fountain	14.61	15.67	11.39
	Car	11.96	9.83	14.90
	Barrier/Bird/Fountain	14.20	15.39	10.26
Bishopthorpe Road	Clean	16.73	8.15	11.91
	Barrier	16.12	9.59	9.84
	Bird	15.96	10.03	11.75
	Fountain	16.92	9.71	12.77
	Car	15.48	8.08	14.38
	Barrier/Bird/Fountain	16.17	6.84	15.97
Rowntree Park	Clean	13.53	21.11	4.21
	Barrier	12.36	21.85	3.79
	Bird	10.15	21.93	6.32
	Fountain	12.77	22.24	7.60
	Car	11.75	13.48	13.12
	Barrier/Bird/Fountain	14.01	22.23	9.32

Bibliography

[1] A. Goudie, *The human impact on the natural environment : past, present, and future*. Malden, MA Oxford: Blackwell Pub, 2006.

[2] C. N. Waters, J. Zalasiewicz, C. Summerhayes, A. D. Barnosky, C. Poirier, A. Ga uszka, A. Cearreta, M. Edgeworth, E. C. Ellis, and M. Ellis, "The Anthropocene is functionally and stratigraphically distinct from the Holocene," *Science*, vol. 351, no. 6269, pp. aad2622–aad2622, Jan 2016.

[3] R. M. Schafer, *The Soundscape: Our Sonic Environment and the Tuning of the World*. Inner Traditions / Bear & Co, 1993.

[4] A. L. Brown, "Soundscapes and environmental noise management," *Noise Control Engineering Journal*, vol. 58, no. 5, pp. 493–500, 2010.

[5] M. Raimbault and D. Dubois, "Urban soundscapes: Experiences and knowledge," *Cities*, vol. 22, no. 5, pp. 339–350, Oct 2005.

[6] J. Kang and B. Schulte-Fortkamp, *Soundscape and the built environment*. CRC Press, 2016.

[7] J. Kang, F. Aletta, T. Oberman, M. Erfanian, M. Kachlicka, M. Lionello, and A. Mitchell, "Towards soundscape indices," in *23rd International Congress on Acoustics*, September 2019.

[8] D. Wang and G. J. Brown, *Computation Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley, 2006.

[9] T. Virtanen, *Computational analysis of sound scenes and events.* Springer, 2018.

[10] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, October 2015.

[11] D. Howard and J. Angus, *Acoustics and Psychoacoustics.* London: Focal, 2009.

[12] "Air - Composition and Molecular Weight," https://www.engineeringtoolbox.com/air-composition-d_212.html, Accessed May 12, 2020.

[13] E. W. Weisstein, "Harmonic addition theorem," https://mathworld.wolfram.com/HarmonicAdditionTheorem.html, Accessed May 12, 2020.

[14] L. Euler, *Introductio in analysin infinitorum*, 1748.

[15] "Decibel Table - SPL Loudness Comparison Chart," http://www.sengpielaudio.com/TableOfSoundPressureLevels.htm, Accessed May 13, 2020.

[16] Z. E. Hajj, "Analysis of Sound Localization and Microphone Arrays," *Ingenium Revista de la Facultad de Ingeniería*, vol. 15, no. 29, 2014.

[17] J. Merimaa, "Analysis, Synthesis, and Perception of Spatial Sound - Binaural Localization Modeling and Multichannel Loudspeaker Reproduction," Ph.D. dissertation, Helsinki University of Technology, 2006.

[18] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept- sine technique," in *108th Convention of the Audio Engineering Society*, February 2000.

[19] International Standards Organisation, "ISO 3382-1:2009 - Acoustics – Measurement of room acoustic parameters," 2009.

[20] F. K. M. Stevens, "Strategies for environmental sound measurement, modelling and evaluation," Ph.D. dissertation, University of York, August 2018.

[21] M. R. Schroeder, "A new method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 6, pp. 1187–1188, 1965.

[22] J. Piechowicz, "Sound wave diffraction at the edge of a sound barrier," *Acoustic and Biomedical Engineering*, vol. 119, no. 6A, pp. 1040 – 1045, 2011.

[23] G. D. Moore, "Lecture 8: Wavelength and diffraction," http://www.physics.mcgill.ca/~guymoore/ph224/, 2006, Accessed May 21, 2020.

[24] B. Rafaely, *Fundamentals of spherical array processing*. Heidelberg Germany: Springer, 2015.

[25] J. Eargle, *Handbook of Recording Engineering*. Springer US, 2006.

[26] D. G. Malham, "Ambisonics - a technique for low cost, high precision, three dimensional sound diffusion," in *International Computer Music Conference*, 1990, pp. 118–120.

[27] ——, "The early years of ambisonics at york," in *International Conference on Immersive and Interactive Audio*, March 2019.

[28] C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi, "Ambix - a suggested ambisonics format," in *Ambisonics Symposium*, 2011.

[29] G. Arfken, H. Weber, and F. Harris, *Mathematical Methods for Physicists: A Comprehensive Guide*. Elsevier Science, 2013.

[30] D. P. Jarrett, "Spherical Microphone Array Processing for Acoustic Parameter Estimation and Signal Enhancement," Ph.D. dissertation, Department of Electrical & Electronic Engineering, Imperial College London, 2013.

[31] L. T. McCormack, "Real-time microphone array processing for sound-field analysis and perceptually motivated reproduction," Master's thesis, Aalto University School of Electrical Engineering, 2017.

[32] mh Acoustics, *em32 Eigenmike® microphone array release notes*, mh acoustics, 25 Summit Ave, Summit, NJ 07901, April 2013.

[33] ——, *Eigenbeam Data: Specification for Eigenbeams*, 2016.

[34] ——, *Beamformer Data: Specification for Eigenmike Software Beamformer*, 2017.

[35] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2149–2157, Oct 2004.

[36] M. B. Coteli, O. Olgun, and H. Hacıhabiboğlu, "Multiple sound source localisation with steered response power density and hierarchical grid refinement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 26, no. 11, pp. 2215 – 2229, November 2018.

[37] J. Daniel, "Représentation de champs acoustiques application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Ph.D. dissertation, Université Paris, July 2001.

[38] S. Delikaris-Manias, D. Pavlidi, V. Pulkki, and A. Mouchtaris, "3D localization of multiple audio sources utilizing 2D DOA histograms," in *24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1473–1477.

[39] F. Zotter, *Ambisonics : a practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality.* Cham, Switzerland: SpringerOpen, 2019.

[40] L. McCormack, S. Delikaris-Manias, and V. Pulkki, "Parametric Acoustic Camera for Real-Time Sound Capture, Analysis and Tracking," in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*, Sep 2017.

[41] G. Andrews, *Special functions*. Cambridge, UK New York, NY, USA: Cambridge University Press, 1999.

[42] M. A. Blanco, M. Flórez, and M. Bermejo, "Evaluation of the rotation matrices in the basis of real spherical harmonics," in *Journal of Molecular Structure*, vol. 419, 1997, pp. 19–27.

[43] S. Harriet, "Application of Auralisation and Soundscape Methodologies to Environmental Noise," Ph.D. dissertation, University of York, 2013.

[44] W. M. Hartmann, "How we localize sound," *Physics Today*, vol. 52, no. 11, pp. 24–29, Nov 1999.

[45] F. Stevens, D. Murphy, and S. L. Smith, "Soundscape auralisation and visualisation: A cross-modal approach to soundscape evaluation," in *Proceedings of the 21st International Conference on Digital Audio Effects (DAFx-18)*, 2018.

[46] D. Botteldooren, B. De Coensel, and T. De Muer, "The temporal structure of urban soundscapes," *Journal of Sound and Vibration*, vol. 292, no. 1-2, pp. 105–123, Apr 2006.

[47] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in *ACM International Conference on Multimedia*, 2014, pp. 1041–1044.

[48] B. C. Pijanowski, A. Farina, S. H. Gage, S. L. Dumyahn, and B. L. Krause, "What is soundscape ecology? an introduction and overview of an emerging new science," *Landscape Ecology*, vol. 26, no. 9, pp. 1213–1232, 2011.

[49] E. P. Kasten, S. H. Gage, J. Fox, and W. Joo, "The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology," *Ecological Informatics*, vol. 12, pp. 50–67, 2012.

[50] P. Devos, "Soundecology indicators applied to urban soundscapes," in *Internoise 2016*, August 2016.

[51] International Standards Organisation, "ISO 12913-1:2014 - Acoustics – Soundscape – Part 1: Definition and conceptual framework," 2014.

[52] B. Brooks and B. Schulte-Fortkamp, "The soundscape standard," in *Internoise 2016*, 2016, pp. 2043 – 2047.

[53] "Lexico online dictionary," https://www.lexico.com/definition/auditory, 2020, Accessed July 17, 2020.

[54] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic Scene Classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, May 2015.

[55] W. Nogueira, G. Roma, and P. Herrera, "Sound Scene Identification Based on MFCC, Binaural Features and a Support Vector Machine Classifier," IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events, Tech. Rep., 2013.

[56] European Commission, "Noise - Environment - European Commission," 2016.

[57] Department for Environment, Food & Rural Affairs, "2010 to 2015 government policy: environmental quality," 2015.

[58] International Electrotechnical Commission, "IEC61672 - A Standard for Sound Level Meters," 2002.

[59] "Hearing protection," https://www.hse.gov.uk/noise/hearingprotection.htm, Accessed May 27, 2020.

[60] B. Berglund, T. Lindvall, and D. H. Schwela, "Guidelines for community noise," https://www.who.int/docstore/peh/noise/Comnoise-1.pdf, 1995, Accessed May 27, 2020.

[61] "Directive 2002/49/ec of the european parliament and of the council (environmental noise directive)." https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32002L0049\&from=EN, Accessed May 27, 2020.

[62] M. Zhang and J. Kang, "Towards the evaluation, description, and creation of soundscapes in urban open spaces," *Environment and Planning B: Planning and Design*, vol. 34, no. 1, pp. 68–86, 2007.

[63] J. Liu, J. Kang, T. Luo, and H. Behm, "Landscape effects on soundscape experience in city parks," *Science of the Total Environment*, vol. 454-455, pp. 474–481, 2013.

[64] D. Birchfield, N. Mattar, and H. Sundaram, "Design of a Generative Model for Soundscape Creation," in *International Computer Music Conference*, 2005.

[65] F. Perraudin, "For whom the bell tolls: York Minster to fall silent as ringers sacked," 2016.

[66] S. Gage, P. Ummadi, A. Shortridge, J. Qi, and P. K. Jella, "Using GIS to Develop a Network of Acoustic Environmental Sensors," in *ESRI international user conference*, 2004.

[67] J. Sueur, A. Farina, A. Gasc, N. Pieretti, and S. Pavoine, "Acoustic Indices for Biodiversity Assessment and Landscape Investigation," *Acta Acustica united with Acustica*, vol. 100, no. 4, pp. 772–781, Jul 2014.

[68] A. Brown, J. Kang, and T. Gjestland, "Towards standardization in soundscape preference assessment," *Applied Acoustics*, vol. 72, no. 6, pp. 387–392, May 2011.

[69] C. Karatsovis and S. J. C. Dyne, "Instrument for soundscape recognition, identification and evaluation: an overview and potential use in legislative applications," *Proceedings of the Institute of Acoustics*, vol. 30, no. 2, 2008.

[70] O. Bunting, J. Stammers, D. Chesmore, O. Bouzid, G. Y. Tian, C. Karatsovis, and S. Dyne, "Instrument for soundscape recognition, identification and evaluation (ISRIE): technology and practical uses," in *Euronoise 2009*, October 2009.

[71] F. Aletta, J. Kang, and Ö. Axelsson, "Soundscape descriptors and a conceptual framework for developing predictive soundscape models," *Landscape and Urban Planning*, vol. 149, pp. 65–74, 2016.

[72] S. Harriet and D. T. Murphy, "Auralisation of an Urban Soundscape," *Acta Acustica united with Acustica*, vol. 101, no. 4, pp. 798–810, Jul 2015.

[73] A. S. Sudarsono, Y. W. Lam, and W. J. Davies, "Soundscape Perception Analysis Using Soundscape Simulator," in *Internoise 2016*, August 2016, pp. 6868–6875.

[74] Ö. Axelsson, "How to measure soundscape quality," in *Euronoise 2015*, May 2015.

[75] W. J. Davies, N. S. Bruce, and J. E. Murphy, "Soundscape Reproduction and Synthesis," *Acta Acustica united with Acustica*, vol. 100, no. 2, pp. 285–292, Mar 2014.

[76] Ö. Axelsson, M. E. Nilsson, and B. Berglund, "A principal components model of soundscape perception," *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 2836–2846, Nov 2010.

[77] R. Cain, P. Jennings, and J. Poxon, "The development and application of the emotional dimensions of a soundscape," *Applied Acoustics*, vol. 74, no. 2, pp. 232–239, Feb 2013.

[78] P. Lundén, Ö. Axelsson, and M. Hurtig, "On urban soundscape mapping: A computer can predict the outcome of soundscape assessments," in *Internoise 2016*, August 2016, pp. 4725–4732.

[79] G. Watts, A. Miah, and R. Pheasant, "Tranquillity and soundscapes in urban green spaces; predicted and actual assessments from a questionnaire survey," *Environment and Planning B: Planning and Design*, vol. 40, no. 1, pp. 170–181, 2013.

[80] F. Stevens, D. Murphy, and S. L. Smith, "Soundscape categorisation and the self-assessment manikin," in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*, 2017.

[81] C. Kroos, O. Bones, Y. Cao, L. Harris, P. J. B. Jackson, W. J. Davies, W. Wang, T. J. Cox, and M. D. Plumbley, "Generalisation in environmental sound classification: The 'making sense of sounds' dataset and challenge," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[82] "If you never heard a vulcan howl, you simply have to," https://worldwarwings.com/if-you-never-heard-a-vulcan-howl-you-simply-have-to/, 2020, Accessed April 8th, 2020.

[83] B. Boren, M. Musick, J. Grossman, and A. Roginska, "I hear ny4d: Hybrid acoustic and augmented auditory display for urban soundscapes," in *20th International Conference on Auditory Display (ICAD-2014)*, June 2014.

[84] J. Kang, *Urban sound environment*. Taylor & Francis, 2007.

[85] N. S. Bruce and W. J. Davies, "The effects of expectation on the perception of soundscapes," *Applied Acoustics*, vol. 85, pp. 1–11, 2014.

[86] J. Carles, F. Bernáldez, and J. de Lucio, "Audio-visual interactions and soundscape preferences," *Landscape Research*, vol. 17, no. 2, pp. 52–56, 1992.

[87] A. Tamura, "Effects of landscaping on the feeling of annoyance of a space," *Contributions to psychological acoustics : Results of the seventh Oldenburg symposium on psychological acoustic*, pp. 135–161, 1997.

[88] F. Ruotolo, L. Maffei, M. D. Gabriele, T. Iachini, M. Masullo, G. Ruggiero, and V. P. Senese, "Immersive virtual reality and environmental noise assessment: An innovative audio–visual approach," *Environmental Impact Assessment Review*, vol. 41, pp. 10–20, 2013.

[89] L. Maffei, M. Masullo, A. Pascale, G. Ruggiero, and V. P. Romero, "On the validity of immersive virtual reality as tool for multisensory evaluation of urban spaces," *Energy Procedia*, vol. 78, pp. 471–476, Nov 2015.

[90] ——, "Immersive virtual reality in community planning: Acoustic and visual congruence of simulated vs real world," *Sustainable Cities and Society*, vol. 27, pp. 338–345, 2016.

[91] G. Kınayoğlu, "Using audio-augmented reality to assess the role of soundscape in environmental perception," in *International Conference on Education and Research in Computer Aided Architectural Design in Europe*, Instanbul, Turkey, 2009.

[92] "Apple ARKit," https://developer.apple.com/arkit/, Accessed January 7, 2019.

[93] "Ikea place," https://itunes.apple.com/app/ikea-place/id1279244498, Accessed July 1, 2019.

[94] "Sirvo llc, housecraft," https://itunes.apple.com/app/housecraft/id1261483849, Accessed July 1, 2019.

[95] "Sennheiser AMBEO Smart Headset," https://en-uk.sennheiser.com/finalstop, Accessed January 7, 2019.

[96] "Apple AirPods Pro," https://www.apple.com/uk/airpods-pro/specs/, Accessed April 14, 2020.

[97] Z. Raś, *Advances in music information retrieval.* Berlin Heidelberg: Springer-Verlag, 2010.

[98] A. Pasick, "The magic that makes spotify's discover weekly playlists so damn good." https://qz.com/571007/the-magic-that-makes-spotifys-discover-weekly-playlists-so-damn-good/, Accessed July 23, 2020.

[99] C. Cherry, *On human communication : a review, a survey, and a criticism.* Cambridge, Mass: MIT Press, 1978.

[100] D. Deutsch, "Music Perception," *Frontiers of Bioscience*, vol. 12, pp. 4473–4482, May 2007.

[101] "DCASE2018 Challenge," http://dcase.community/challenge2018/, Accessed April 16, 2020.

[102] J. S. Downie, X. Hu, J. H. Lee, K. Choi, S. J. Cunningham, and Y. Hao, "Ten years of mirex: Reflections challenges and opportunities," in *15th International Society for Music Information Retrieval Conference*, 2014.

[103] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, April 2008.

[104] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR Evaluation of Acoustic Event Detection and Classification Systems," in *Multimodal Technologies for Perception of Humans*, vol. 4122, 2007, pp. 311–322.

[105] T. Heittola, E. Çakır, and T. Virtanen, *Computational Analysis of Sound Scenes and Events.* Springer International Publishing, Sep 2017.

[106] E. Stevens and L. Antiga, *Deep Learning with PyTorch.* Shelter Island, NY: Manning Publications, 2019.

[107] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features," in *Detection and Classification of Acoustic Scenes and Events*, 2017.

[108] H. Phan, P. Koch, F. Katzberg, M. Maass, R. Mazur, I. McLoughlin, and A. Mertins, "What makes audio event detection harder than classification?" in *25th European Signal Processing Conference*, 2017, pp. 2739 – 2743.

[109] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, 2013.

[110] A. Mesaros, T. Heittola, and A. P. Klapuri, "Latent semantic analysis in sound event detection," in *EUSIPCO*, 2011.

[111] J. Lyons, "Mel Frequency Cepstral Coefficient (MFCC) tutorial," http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/, Accessed December 16, 2016.

[112] D. O'Shaughnessy, *Speech communication : human and machine.* Reading, Mass: Addison-Wesley Pub. Co, 1987.

[113] S. B. David and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, August 1980.

[114] A. Meyer-Baese and V. Shmid, *Pattern recognition and signal analysis in medical imaging.* Oxford, UK: Academic Press, 2014.

[115] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-Based Context Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 321–329, January 2006.

[116] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," in *Proc. of the 14th Python in Science Conference (SciPy 2015)*, 2015.

[117] M. J. B. Garrido, A. J. Torija, M. D. Fernandez, and J. A. Ballesteros, "Differences Between Road Traffic and Leisure Noise in Urban Areas. Developing a Model for Automatic Identification," *Acta Acustica united with Acustica*, vol. 102, no. 1, pp. 35–44, Jan 2016.

[118] D. A. Reynolds, "Gaussian mixture models." *Encyclopedia of biometrics*, vol. 741, 2009.

[119] J. Lyons, "Gaussian Mixture Model Tutorial," http://practicalcryptography.com/miscellaneous/machine-learning/gaussian-mixture-model-tutorial/, Accessed December 16, 2016.

[120] S. Marsland, *Machine Learning: An Algorithmic Perspective*. CRC Press, 2015.

[121] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, ser. Springer Series in Statistics. Springer New York, 2013.

[122] R. Grosse and N. Srivastava, "Lecture 16: Mixture Models," http://www.cs.toronto.edu/~rgrosse/csc321/mixture_models.pdf, Accessed April 22, 2020.

[123] Y. Zhu, "Tutorial and Illustration: Gaussian Mixture Model," http://yulearning.blogspot.com/2014/11/einsteins-most-famous-equation-is-emc2.html, 2014, Accessed September 7, 2020.

[124] D. Stowell, "smacpy - simple-minded audio classifier in python," https://github.com/danstowell/smacpy, 2012, Accessed July 13, 2017.

[125] A. Mesaros, T. Heittola, and T. Virtanen, "TUT Database for Acoustic Scene Classification and Sound Event Detection," in *24th European Signal Processing Conference (EUSIPCO)*, August 2016.

[126] A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.

[127] "Why is the svm margin equal to $\frac{2}{\|\mathbf{w}\|}$?" https://math.stackexchange.com/q/1306417, 2015, Accessed September 7, 2020.

[128] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273 – 297, 1995.

[129] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer New York, 2013.

[130] "Support Vector Machines," https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html, Accessed April 27, 2020.

[131] "RBF SVM parameters," https://scikit-learn.org/stable/modules/svm.html\#svm-mathematical-formulation, Accessed April 27, 2020.

[132] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[133] N. Mehta and S. Dang, "A review of clustering techniques in various applications for effective data mining," *International Journal of Research in IT & Management*, vol. 1, no. 2, pp. 50 – 66, June 2011.

[134] "DCASE2019 Challenge," http://dcase.community/challenge2019/, Accessed April 16, 2020.

[135] J. Keller, D. Liu, and D. Fogel, *Fundamentals of Computational Intelligence: Neural Networks, Fuzzy Systems, and Evolutionary Computation*, ser. IEEE Press Series on Computational Intelligence. Wiley, 2016.

[136] C. E. Kim and M. G. Strintzis, "High-speed multidimensional convolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-2, no. 3, pp. 269–273, 1980.

[137] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[138] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[139] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[140] K. Koutini, H. Eghbal-Zadeh, M. Dorfer, and G. Widmer, "The receptive field as a regularizer in deep convolutional neural networks for acoustic scene classification," in *EUSIPCO*, 2019.

[141] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The Bag-of-frames Approach to Audio Pattern Recognition: A Sufficient Model for Urban Soundscapes But Not For Polyphonic Music," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, 2007.

[142] J.-J. Aucouturier and F. Pachet, "Improving Timbre Similarity : How high's the sky ?" *Journal of negative results in speech and audio sciences*, 2004.

[143] S. M. Wiseman and P. S. Wilson, "An argument for a standardized method to record, measure, characterize, and compare captive animal soundscapes," in *Internoise 2016*, August 2016, pp. 2074–2085.

[144] M. Lagrange, G. Lafay, B. Défréville, and J.-J. Aucouturier, "The bag-of-frames approach: A not so sufficient model for urban soundscapes," *Journal of the Acoustical Society of America*, vol. 128, no. 5, November 2015.

[145] Y. E. Kim, D. S. Williamson, and S. Pilli, "Towards quantifying the album-effect in artist classification," in *In Proceedings of the International Symposium on Music Information Retrieval*, 2006.

[146] O. Bunting and D. Chesmore, "Time frequency source separation and direction of arrival estimation in a 3D soundscape environment," *Applied Acoustics*, vol. 74, no. 2, pp. 264–268, Feb 2013.

[147] O. Bunting, "Sparse Seperation of Sources in 3D Soundscapes," Ph.D. dissertation, Department of Electronics, University of York, 2010.

[148] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU Submissions for DCASE-2016: A Hybrid Approach Using Binaural I-Vectors and Deep Convolutional Neural Networks," in *Detection and Classification of Acoustic Scenes and Events*, Sep 2016.

[149] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," in *Detection and Classification of Acoustic Scenes and Events*, 2018.

[150] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379 – 393, February 2018.

[151] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound Event Detection in Multisource Environments Using Spatial and Harmonic Features," in *Detection and Classification of Acoustic Scenes and Events*, 2016.

[152] C. H. Knapp, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, August 1976.

[153] S. Chakrabarty and E. A. P. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017.

[154] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, March 2018.

[155] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[156] S. Adavanne, A. Politis, and T. Virtanen, "Direction of Arrival Estimation for Multiple Sound Sources Using Convolutional Recurrent Neural Network," *arXiv:1710.10059*, 2018.

[157] S. Adavanne, P. Pertila, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[158] S. Chu, S. Narayanan, C.-c. Kuo, and M. Mataric, "Where am i? scene recognition for mobile robots using audio features," *2006 IEEE International Conference on Multimedia and Expo*, Jul 2006.

[159] R. Duraiswami, D. N. Zotkin, E. Grassi, N. A. Gumerov, and L. S. Davis, "High order spatial audio capture and its binaural head-tracked playback over headphones with hrtf cues," in *119th Convention of the Audio Engineering Society*, October 2005.

[160] A. J. Torija, D. P. Ruiz, and Á. F. Ramos-Ridao, "A tool for urban soundscape evaluation applying Support Vector Machines for developing a soundscape classification model," *Science of the Total Environment*, vol. 482-483, pp. 440–451, 2014.

[161] A. Mesaros, T. Heittola, and T. Virtanen, "Tut acoustic scenes 2016, development dataset," Feb. 2016. [Online]. Available: https://doi.org/10.5281/zenodo.45739

[162] ——, "Tut acoustic scenes 2017, development dataset," Mar. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.400515

[163] T. Heittola, A. Mesaros, and T. Virtanen, "TUT Urban Acoustic Scenes 2018, Development dataset," Apr. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1228142

[164] ——, "TAU Urban Acoustic Scenes 2019, Development dataset," Mar. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2589280

[165] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Detection and Classification of Acoustic Scenes and Events*, 2019.

[166] ——, "TAU Spatial Sound Events 2019 - Ambisonic and Microphone Array, Development Datasets," Feb. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2599196

[167] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Detection and Classification of Acoustic Scenes and Events*, 2020.

[168] G. Lafay, E. Benetos, and M. Lagrange, "IEEE DCASE 2016 Challenge - Task 2 - Train/Development Datasets," 2016, Accessed June 23, 2020. [Online]. Available: https://archive.org/details/dcase2016_task2_train_dev

[169] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The nigens general sound events database," 2020.

[170] N. I. Joachim Thiemann and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, vol. 19, June 2013.

[171] European Language Resources Association, "AURORA Project Database 2.0 - Evaluation Package," 2008. [Online]. Available: http://catalog.elra.info/en-us/repository/browse/ELRA-AURORA-CD0002/

[172] Speech at CMU, "NOISEX-92 Database," 1992. [Online]. Available: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

[173] "ST450 MKII SoundField Portable," https://www.soundfield.com/#/products/st450mk2, 2020, Accessed Jun 1, 2020.

[174] E. Bates, M. Gorzel, L. Ferguson, H. O'Dwyer, and F. M. Boland, "Comparing Ambisonic Microphones – Part 1," in *Audio Engineering Society Conference: 2016 AES International Conference on Sound Field Control*, Jul 2016.

[175] E. Bates, S. Dooney, M. Gorzel, H. O'Dwyer, L. Ferguson, and F. M. Boland, "Comparing Ambisonic Microphones – Part 2," in *142nd Convention of the Audio Engineering Society*, May 2017.

[176] mh Acoustics, "Eigenstudio user manual," https://mhacoustics.com/sites/default/files/EigenStudio%20User%20Manual%20R02C.pdf, 2019, Accessed September 7 , 2020.

[177] "Macbook pro (15-inch, mid 2012) - technical specifications," https://support.apple.com/kb/sp694?locale=en_GB, Accessed Jun 1, 2020.

[178] M. Chapman, W. Ritsch, T. Musil, J. Zmolnig, H. Pomberger, F. Zotter, and A. Sontacchi, "A standard for interchange of ambisonic signal sets," in *Ambisonics Symposium*, 2009.

[179] "Rycote modular windshield," https://rycote.com/microphone-windshield-shock-mount/modular-windshield-kit/, 2020, Accessed Jun 3, 2020.

[180] M. W. W. van Grootel, T. C. Andringa, and J. D. Krijnders, "DARES-G1: Database of Annotated Real-world Everyday Sounds," in *Proceedings of the NAG/DAGA Meeting*, January 2009.

[181] "Samsung gear 360 camera," https://www.samsung.com/us/support/owners/product/gear-360-2016, Accessed Jun 1, 2020.

[182] "To film or not to film," https://www.theiac.org.uk/resourcesnew/filming-in-public/filming-in-public.html, 2016, Accessed June 2, 2020.

[183] M. C. Green and D. Murphy, "Eigenscape," Oct. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1284156

[184] "Creative commons attribution 4.0 international," https://creativecommons.org/licenses/by/4.0/legalcode, 2020, Accessed Jun 3, 2020.

[185] "Uk data service - recommended formats." https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats, 2020.

[186] A. Choudhury, "10 open datasets you can use for computer vision projects," https://analyticsindiamag.com/10-open-datasets-you-can-use-for-computer-vision-projects/, 2019, Accessed Jun 3, 2020.

[187] marc1701, "marc1701/EigenScape: Initial Dataset Tools and Analysis," Oct. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1039661

[188] M. C. Green, S. Adavanne, D. Murphy, and T. Virtanen, "Acoustic scene classification using higher-order ambisonic features," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2019.

[189] V. Pulkki, "Directional audio coding in spatial sound reproduction and stereo upmixing," in *AES 28th International Conference*, 2006.

[190] ——, "Spatial Sound Reproduction with Directional Audio Coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, June 2007.

[191] V. Pulkki, M.-V. Laitinen, J. Vilkamo, J. Ahonen, T. Lokki, and T. Pihlajamäki, "Directional audio coding - perception-based reproduction of spatial sound," in *International Workshop on the Principle and Applications of Spatial Hearing*, 2009.

[192] J. Merimaa and V. Pulkki, "Spatial impulse response rendering: Listening tests and applications to continuous sound," in *Audio Engineering Society Convention 118*, May 2005.

[193] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *The Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3075–3089, Nov 2004.

[194] A. Politis, J. Vilkamo, and V. Pulkki, "Sector-Based Parametric Sound Field Reproduction in the Spherical Harmonic Domain," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 852–866, Aug 2015.

[195] N. Epain and C. T. Jin, "Spherical harmonic signal covariance and sound field diffuseness," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1796–1807, October 2016.

[196] T. McKenzie, D. Murphy, and G. Kearney, "Diffuse-field equalisation of binaural ambisonic rendering," *Applied Sciences*, vol. 8, no. 10, p. 1956, Oct 2018.

[197] M. Atiyah and P. Sutcliffe, "Polyhedra in physics, chemistry and geometry," pp. 33–58, 2003.

[198] V. I. Lebedev, "Spherical quadrature formulas exact to orders 25–29," *Siberian Mathematical Journal*, vol. 18, no. 1, pp. 99–107, 1977.

[199] E. B. Saff and A. B. J. Kuijlaars, "Distributing many points on a sphere," *The Mathematical Intelligencer*, vol. 19, no. 1, pp. 5 – 11, 1997.

[200] H. Vogel, "A better way to construct the sunflower head," *Mathematical Biosciences*, vol. 44, pp. 179–189, 1979.

[201] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929 – 1958, June 2014.

[202] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, Vancouver, Canada, May 2015.

[203] "Acoustic scene classification - DCASE," http://dcase.community/challenge2018/task-acoustic-scene-classification-results-a\#machine-learning-characteristics, Accessed April 11, 2019.

[204] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, "Integrating the data augmentation scheme with various classifiers for acoustic scene modelling," in *Detection and Classification of Acoustic Scenes and Events*, October 2019.

[205] "The zoom h2n," https://www.zoom-na.com/products/field-video-recording/field-recording/zoom-h2n-handy-recorder, Accessed June 15, 2020.

[206] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio Surveillance: A Systematic Review," *ACM Computing Surveys*, vol. 48, no. 4, Feb 2016.

[207] M. R. Azimi-Sadjadi and N. R. A. Pezeshki, "Wideband DOA Estimation Algorithms for Multiple Moving Sources using Unattended Acoustic Sensors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 44, no. 4, pp. 1585–1598, October 2008.

[208] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor Auditory Scene Analysis Using a Moving Microphone Array Embedded in a Quadrocopter," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[209] T. N. T. Nguyen, D. L. Jones, and W. S. Gan, "Dcase 2020 task 3: Ensemble of sequence matching networks for dynamic sound event localization, detection, and tracking," in *Detection and Classification of Acoustic Scenes and Events*, 2020.

[210] E. Garyfallidis, M. Brett, B. Amirbekian, A. Rokem, S. van der Walt, M. Descoteaux, and I. Nimmo-Smith, "Dipy, a library for the analysis of diffusion mri data," *Frontiers in Neuroinformatics*, vol. 8, Feb 2014.

[211] M. Celik, F. Dadaser-Celik, and A. S. Dokuz, "Anomaly detection in temperature data using DBSCAN algorithm," in *2011 International Symposium on Innovations in Intelligent Systems and Applications*. IEEE, jun 2011.

[212] S. Kiware, "Detection of outliers in time series data," Master's thesis, Marquette University, 2010.

[213] M. Cartwright, A. Seals, J. Salamon, A. Williams, S. Mikloska, D. Macconnell, E. Law, J. P. Bello, and O. Nov, "Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations," in *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. 2, 2017.

[214] S. Adavanne, A. Politis, and T. Virtanen, "TUT Sound Events 2018 - Ambisonic, Anechoic and Synthetic Impulse Response Dataset," Apr. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1237703

[215] ——, "TUT Sound Events 2018 - Ambisonic, Reverberant and Synthetic Impulse Response Dataset," Apr. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1237707

[216] "Sound event localization and detection challenge results," https://www.zoom-na.com/products/field-video-recording/field-recording/zoom-h2n-handy-recorder, Accessed June 30, 2020.

[217] "IEEE DCASE 2019 task 3," http://dcase.community/challenge2019/task-sound-event-localization-and-detection, Accessed June 23, 2020.

[218] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, 1955.

[219] R. Lang and K. Warwick, "The Plastic Self Organising Map," in *Proc. of the 2002 International Joing Conference on Neural Networks*, 2002.

[220] J. Bergstra, D. Yamins, and D. D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, Atlanta, Georgia, 2013.

[221] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kegl, "Algorithms for hyper-parameter optimization," in *Neural Information Processing Systems*, Granada, Spain, Dec 2011.

[222] "Core ml," https://developer.apple.com/documentation/coreml, Accessed July 1, 2019.

[223] "Apple ios," https://www.apple.com/uk/ios/ios-13/, Accessed August 24, 2020.

[224] P. Brossier, Tintamar, E. Müller, N. Philippsen, T. Seaver, H. Fritz, cyclopsian, S. Alexander, J. Williams, J. Cowgill, and A. Cruz, "aubio/aubio: 0.4.8," https://aubio.org/, Nov. 2018, Accessed September 7, 2020.

[225] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances In Large Margin Classifiers*. MIT Press, 1999, pp. 61–74.

[226] "Scikit-learn svc," https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html, Accessed August 24, 2020.

[227] "Scenekit apple developer documentation," https://free3d.com/, Accessed: August 24, 2020.

[228] "Scenekit apple developer documentation," https://freesound.org/, Accessed: August 24, 2020.

[229] "Scenekit apple developer documentation," https://developer.apple.com/documentation/scenekit, Accessed: July 2, 2020.

[230] "Creating an immersive ar experience with audio," https://developer.apple.com/documentation/arkit/creating_an_immersive_ar_experience_with_audio, Accessed July 1, 2019.

[231] E. Murphy and E. A. King, "Testing the accuracy of smartphones and sound level meter applications for measuring environmental noise," *Applied Acoustics*, vol. 106, 16-22 2016.

[232] R. B. D'Agostino, A. Belanger, and R. B. D. Jr., "A suggestion for using powerful and informative tests of normality," *The American Statistician*, vol. 44, no. 4, pp. 316 – 321, Nov 1990.

[233] E. Girden and i. Sage Publications, *ANOVA: Repeated Measures*, ser. ANOVA: Repeated Measures. SAGE Publications, 1992, no. no. 84.

[234] G. Norman and D. Streiner, *Biostatistics: The Bare Essentials*. B.C. Decker, 2008.

[235] H. Phan, P. Koch, F. Katzberg, M. Maass, R. Mazur, and A. Mertins, "Audio scene classification with deep recurrent neural networks," in *Interspeech*, August 2017.

[236] J. Abeßer, "A review of deep learning based methods for acoustic scene classification," *Applied Sciences*, vol. 10, no. 6, p. 2020, Mar 2020.

[237] I. Bocharov, T. Tjalkens, and B. De Vries, "Acoustic scene classification from few examples," *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep 2018.

[238] A. Vafeiadis, D. Kalatzis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Acoustic scene classification: From a hybrid classifier to deep learning," in *Detection and Classification of Acoustic Scenes and Events*, November 2017.