# Auralisation of Traffic Flow using Procedural Audio Methods

**Yang Fu**

Doctor of Philosophy

University of York

Electronic Engineering

**January 2021**

# *Abstract*

This thesis investigates approaches for the auralisation of traffic noise in an outdoor environment. A novel auralisation framework for multiple vehicle pass-bys using procedural audio methods is proposed. This includes sound source modelling of single vehicle pass-bys and traffic flow, sound propagation modelling, and HRTF processing for spatial audio reproduction. Compared to prior auralisation studies in which sound source recordings have been used, no pre-recorded sounds are used with a procedural audio approach. Instead, synthetic sounds created by programmatic rules form the basis of the auralisation framework proposed in this thesis.

Such an auralisation based on procedural audio gives greater freedom and range in the implementation and integration of vehicle pass-by sounds, with the advantage of high flexibility and variable computational cost for the algorithms defining the properties of any given audio objects. However, such synthetic sounds might not be perceived as being plausible when compared to their recorded counterparts, especially for the case of traffic noise where it is difficult to imitate the intrinsic rich and varied sound source content by artificial means. Therefore, two subjective listening tests are implemented to evaluate the plausibility of the proposed auralisation framework by comparing procedurally generated vehicle sounds to their counterparts created using a recording-based granular synthesis method.

Engine sounds, engine plus tyre sounds, and single vehicle pass-by sounds, all generated using a procedural audio approach, are compared with their counterparts created using a granular synthesis method, and evaluated in an ABX listening test. It is found that a similar level of plausibility is achieved by using either method for the auralisation of single vehicle pass-bys. Based on this validation, the plausibility of multiple vehicle pass-by sounds with engines synthesised using a procedural, a mix of procedural and granular, and granular approaches is evaluated in a MUSHRA test under various traffic flow conditions regarding different vehicle types, speeds, driving directions, and flow rates. It is found that a similar level of plausibility is achieved by using either method under most traffic flow conditions.

These results verify that the auralisation of traffic flow using procedural audio methods is comparable to recording-based approaches when considering the plausibility of the results obtained. Such an approach provides a solution for implementing the auralisation of environmental sounds that is both flexible and plausible, which is useful for communicating and demonstrating the important changes in our soundscape to the wider population, leading to a more holistic understanding of environmental sound.

# *Acknowledgements*

I would like to thank everyone who has supported me and contributed to this work. First of all, I need to express my deepest thanks to my PhD supervisor Prof. Damian Murphy who has supported me throughout the whole journey of my PhD study. I am extremely grateful for his constant patient, enthusiastic and positive attitude to me, as well as the great expertise and guidance for my work. Without the support from my supervisor it is impossible for me to complete such a PhD project.

I would also like to thank Dr. Dave Chesmore and Mr. Tony Tew for their respective contributions as my thesis advisor and progression chair, and to Dr. Helena Daffern who acted as my thesis advisor in the later stage of my research. Thanks to all the past and present members at the AudioLab, whom it is a great pleasure for me to work with and learn from.

This work is funded by China Scholarships Council joint research scholarships. Thanks for the funding for this PhD project.

Last but not least, my heartfelt thanks to my parents and my dearest girlfriend Sophia, for their greatest support and encouragement throughout the journey of my PhD work.

# *Declaration of Authorship*

I declare that this thesis titled ‘Auralisation of Traffic Flow using Procedural Audio Methods’ is an original work written by myself as the sole author. All contributions from outside sources, through direct contact or publications, have been explicitly stated and referenced. I also declare that some parts of this program of research have been presented previously at conferences and workshops, listed as follows:

1. Fu, Y., & Murphy, D. (2017). Spectral Modelling Synthesis of Vehicle Pass-by Noise. Paper and oral presentation in INTER-NOISE and NOISE-CON Congress and Conference Proceedings (Vol. 255, No. 1, pp. 5997-6006). Institute of Noise Control Engineering.

2. Fu, Y., & Murphy, D. (2018). A Comparative Overview of Traffic Flow Modelling Approaches for Auralistion. pp. 273-278. Paper and oral presentation at EuroNoise 2018, Crete, Greece.

3. FU, Y., MURPHY, D. , & SOUTHERN, A. (2019). Traffic Flow Auralisation based on Single Vehicle Pass-by Noise Synthesis. pp. 1705-1712. Paper and oral presentation at ICA 2019, the International Congress on Acoustics 2019, Aachen, Germany.

4. Auralisation of Traffic Flow using Procedural Audio Methods, Invited presentation at the Institute of Acoustics (IOA) Yorkshire and North East Branch Evening Meeting, November 2019, Leeds, UK.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

We live in a world full of sound, a world in which these sounds may have both positive and negative impacts on our daily lives. Our vision dominates our sensory perception of the world, and we can close our eyes if we feel that visual scenes are not pleasant or comfortable. But we cannot close our ears to the sounds of our world. Therefore, sounds unconsciously influence our behaviours, attitudes, and emotions, at all times, no matter what is going on in our conscious minds. Hearing is one of the most important ways by which we perceive and recognise the environments we inhabit, and develop impressions and memories of places, people, and events. Under most circumstances, we pay peripheral attention to sounds that surround us. For some specific sound-related events, we may focus all of our attention on these particular events [24], such as when listening to symphony music in a concert hall. Sometimes, our thinking or main attention may be interrupted by some other specific sounds, causing abrupt transitions between peripheral attention and central attention, which can make us feel disturbed, irritated, or anxious.

In recent decades, there has been increased attention paid to environmental noise and in particular due to the soaring number of different means of modern transport in our daily lives such as cars, aircraft, and trains. Here the term *noise* can be defined as ‘unwanted sound’, or sound ‘out of place’ [25]. Among noise emitted from various means of transport, the influence and treatment of road traffic noise has been widely investigated as it is one of the most common noise sources in our daily lives. In fact, the noise emitted by road traffic noise has been acknowledged as a form of environmental pollution [26].

However, it is still not sufficient to develop a complete and comprehensive understanding of environmental noise using only traditional sound evaluation methods such as noise level measurements. Likewise, the most effective way to treat sources of noise might not be through simply reducing sound levels. This is because the subjective experience of the content of a sound is significantly neglected in traditional methods based on noise levels.

A prospective solution to this deficiency is by considering the *auralisation* of traffic noise as determined by different traffic flow conditions in an outdoor acoustic environment. Whereas most existing auralisation techniques are typically used for the reproduction of room acoustics in a virtual environment, their application to outdoor environments is less widely practised. Part of the reason for this may be the incompatibility of recording-based sound source modelling techniques as used in room acoustics, with the requirement for more complex, dynamic, and wide-scale outdoor sound scenes. For example, a 'dry' music or speech signal recorded in an anechoic chamber (an environment designed to completely absorb sound) is usually used for sound source modelling in room acoustic applications. It is hard, or even impossible, to record vehicle pass-by noise in an anechoic environment, particularly for a traffic flow consisting of multiple pass-by vehicles. Therefore, it is vital to obtain the required source signal for traffic flow sounds via other methods, or in addition to, taking anechoic recordings.

One potential way is through the use of *procedural audio*, which is a sound synthesis technique based purely on codes and algorithms without having an extensive database of recordings. Procedural audio has been mainly used by sound designers in video games and animations for creating some ambient sounds with algorithms. This thesis considers the value of exploring the suitability and potential of procedural audio for sound source modelling in an outdoor environment for auralisation purposes.

## 1.1 Motivation

The most direct motivation behind this study is to develop and apply auralisation techniques for outdoor sound environments. Nowadays, the role of most acousticians covers both architectural acoustics and environmental sound. When planning a new building or designing a new room, acousticians can be involved in the project at very early stages. Auralisation for room acoustics has been used at different stages throughout a whole project to support the acoustic design with audible experiences, being both useful and convenient for internal communication as part of the design process, as well as for demonstrating to, e.g. stakeholders and the wider public. However, this is in contrast to what typically happens in projects involving outdoor planning and development, when considering environmental sound. For most outdoor projects, the acoustic aspects are often ignored at early stages until acoustic-related problems become severe and need to be treated with remedial actions. This is partly due to the lack of tools for auralisation of environmental sounds for outdoor environments, which may lead to higher communication and dissemination costs between acousticians, urban designers, stakeholders, and the public. Therefore, it is worth seeking the opportunity to translate the workflow of architectural acoustics into outdoor projects, with the development and application of outdoor sound auralisation integrated at early stages of landscape design.

One motivation for this research relates to the technical challenges associated with the acoustic modelling and auralisation of sound scenes in an outdoor environment. As it is difficult, or not cost-effective, to make anechoic recordings for some environmental sounds in order to build up a source model, i.e. traffic flow noise emitted by multiple pass-by vehicles, it is necessary to seek other methods for sound source modelling for auralisation according to the nature of the required environmental sound – something that procedural audio may be able to assist with.

An additional motivation for this study concerns the diverse attitudes to environmental sounds in different disciplines. In traditional environmental noise studies, outdoor sounds are often considered as unwanted, and so need to be suppressed by reducing sound levels towards values that are statistically acceptable. From the perspective of soundscape studies, the relationship between human and environment is more connected, and environmental sounds are evaluated from more diverse perspectives, including both positive effects and negative impacts on people's lives. From the perspective of sound designers and artists working with sounds, outdoor sounds, including some 'unwanted sounds' (i.e. traffic noise) in other contexts, can be considered as useful resources for rendering an acoustic design for a virtual environment or the real world. It is worth exploring the potential for auralisation as a tool for bridging the gaps between different understandings, and treatment methods, for environmental sounds from different perspectives and form a more holistic understanding of how humans interact with outdoor sounds and their sound environment.

## 1.2 Statement of Hypothesis

The hypothesis guiding this thesis can be summarised as follows:

> *The auralisation of traffic flow sounds using a procedural audio approach is comparable to methods based on recorded or sampled audio when considering the plausibility of the results obtained.*

The key concepts of this hypothesis can be briefly explained as follows to form a general understanding of what will be included in this thesis and how they relate to the motivation and justification of this study:

- **Auralisation**. The general idea of auralisation can be considered as the analogue of visualisation but for auditory perception. A widely used definition of auralisation as proposed by Vörlander, states that: '*Auralisation is the technique for creating audible sound files from numerical (simulated, measured, synthesised) data.*' [7]

- **Traffic Flow sounds**. This refers to road traffic noise emitted by multiple pass-by vehicles. In traditional traffic flow studies, the efficient movement of multiple vehicles is mainly concerned with reducing traffic congestion problems. In this study, traffic flow is considered as one of the most common outdoor sound sources in our daily lives and so needs to be evaluated and treated to perceptually improve our environment.
- **Procedural Audio**. In general, *procedural audio* can be considered as sound synthesis techniques based purely on algorithms and codes running in real-time without the use of extensive recorded or sampled audio database. A more detailed discussion on the concepts and characteristics of procedural audio can be found in Chapter 4.
- **Plausibility**. This refers to how an auditioned sound scene agree with expectations. In terms of auralisation, a 'plausible' auralised sound event or sound scene means that listeners should consider the sound event or sound scene credible for the real-world circumstance that it proposes to replicate.

In order to prove or disprove this hypothesis, two steps will be taken in the thesis. Firstly, a traffic flow auralisation model using procedural audio approaches will be explored, and secondly, the plausibility of the proposed auralisation model will be tested via a series of subjective listening tests.

## 1.3 Structure of Thesis

This thesis starts with an introduction to the fundamental principles of acoustics in Chapter 2. It then gives an overview of concepts and methodologies in three traditionally distinct areas of acoustics: auralisation, procedural audio, and environmental sound, in Chapter 3, 4, and 5. In each chapter, a literature review of some existing research in the relevant area is presented in addition to some discussion about how they are related to this study, so as to establish the context within which this thesis is presented. Throughout the literature review, it is considered reasonable and meaningful to develop an auralisation framework for some environmental sounds, such as traffic flow noise, using procedural audio, which has a rational compromise between plausibility, flexibility and interactivity. This leads to the content in Chapter 6, which includes details regarding the methodologies and key considerations for sound source modelling, sound propagation modelling, and binaural audio reproduction for building up an auralisation framework for traffic flow sounds using procedural audio approaches.

In order to evaluate the performance of the proposed auralisation framework in terms of plausibility, two listening tests are conducted in Chapter 7. These listening tests compare the plausibility of single vehicle pass-by and traffic flow auralisation implemented using procedural audio, to a granular synthesis, recording-based audio synthesis model. This helps to develop an understanding as to what level of plausibility procedural audio techniques might achieve and whether they might be considered comparable to recording-based sound synthesis techniques in an auralisation context.

The Chapter 8 provides a summary of the work presented in this thesis along with the main conclusions that have been drawn. The main contributions to the relevant fields are then presented, followed by some considerations on future research topics and directions related to this study.

# Chapter 2

# Fundamentals of Acoustics

Acoustics is an area that involves the study of all types of mechanical wave that are transmitted in solids, liquids and gases. For the purposes of auralisation that focuses on the sensation and perception of sound, the interest lies in airborne sound that is detectable by the human hearing system. When investigating acoustics problems via auralisation techniques, it is imperative to understand the properties of sound and sound evaluation methodologies to inform the design of acoustic models for auralisation. This chapter introduces the fundamental concepts, theories and applications of acoustics related to auralisation. The sound generation mechanism is first described, followed by a series of basic properties of sound propagation in air, including the concept and description of longitudinal waves, sound wave phenomena, and the Doppler effect. As the aim of auralisation lies in creating audible sound, the mechanism of the human hearing system and the background knowledge of sound evaluation methodologies is introduced. The purpose of this chapter is to set a knowledge base and context for the rest of this thesis. More detailed information about the fundamentals in this section can be found in a series of textbooks on acoustics, psychoacoustics, and auralisation, such as: [1, 7, 27–30].

## 2.1 Sound Generation Mechanisms

Sound is a wave phenomenon in fluid or solid mediums. Waves are caused by particles vibrating at a rate in a medium while physical force applied. As air is an elastic medium, when vibration occurs, the displacement of a particle from its stable equilibrium causes fluctuation to its neighbouring particles, then reciprocated by the other neighbouring particles in the medium, leading to a density fluctuation in the medium. This kind of density fluctuation causes time- and space-dependent compression and decompression of the medium, which forms a ripple effect. The energy will be dissipated by the medium when vibration is ceased. Compression occurs within the areas of high pressure where particles are more densely packed, while rarefaction decompression occurs within the areas of low pressure where the particles are more sparsely distributed, as demonstrated in Fig.2.1. The human hearing system is able to sense these fluctuations near the ears, and sound can be perceived if the frequencies of vibration fall into a specific range within which the human hearing system is sensitive (20–20kHz). A wave caused by vibration at these audible frequencies is called a *sound wave*.

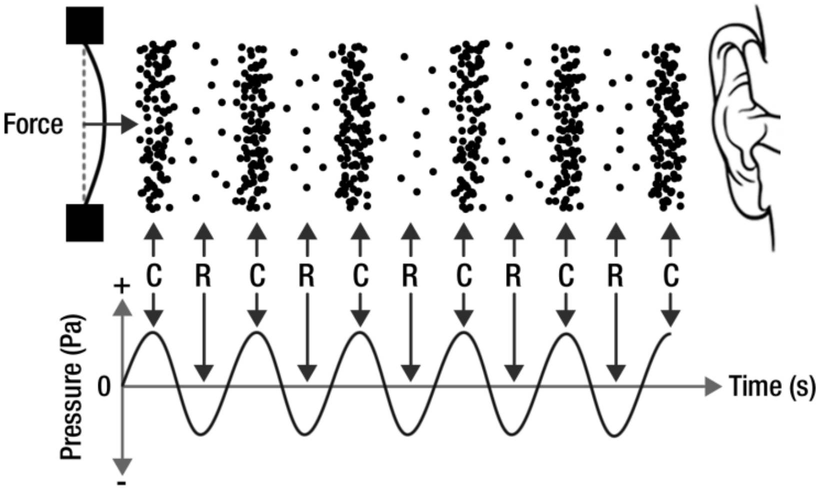


**Figure 2.1:** *Sound is a pressure wave consisting of regions of compression (C) and rarefaction (R), reproduced from [3].*

## 2.2 Sound Propagation in Air

### 2.2.1 Longitudinal Waves

Sound travels as *longitudinal waves* in air. Longitudinal waves are waves in which the displacement of the medium is in the same direction as, or the opposite direction to, the direction of propagation of the wave. This travelling pattern can be represented by a 1-D mass spring system, in which particles are represented by mass, and elastic properties are represented by springs linked between masses, as shown in Figure 2.2.

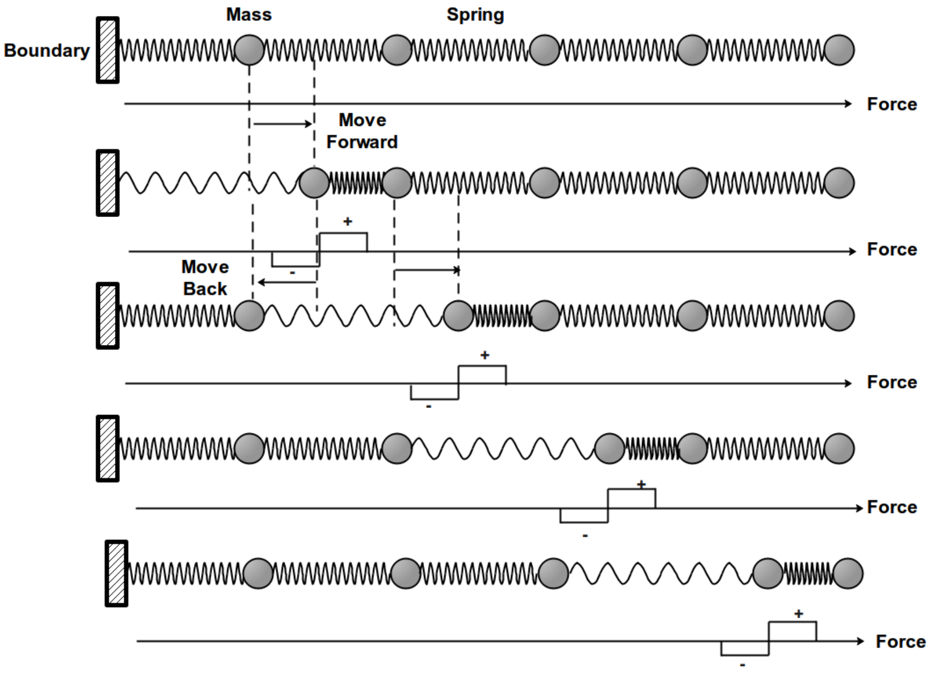


**Figure 2.2:** *Spring-Mass system representing sound propagation in air. The transportation of energy throughout the medium is represented by compression/rarefaction of the springs. Reproduced from [4].*

The system is static initially. When a force is added to the first mass on the left, the mass will be displaced towards the right, which causes the spring linked on the right to compress and the spring linked on the left to stretch. The compression will transfer kinetic energy to the next mass and the rarefaction imposed on the first mass restores the mass to its starting position. In other words, the movement of the first mass causes a rarefaction between the first two masses as the second mass is compressed on its right hand spring. This process will go through every mass-spring block of the system.

It is noted that although the energy in the system is transferred through the mass-spring blocks by a series of compressions and rarefactions, for each single mass, the motion pattern is just back and forth between the starting point and its peak displacement, rather than travelling with the energy flow. This reveals a key concept of sound waves travelling in air: it is not the travelling of molecules in the air that forms a sound wave, but the transportation of energy throughout the medium due to the changing inter-molecules forces causing the pressure fluctuation. The fluctuation is then perceived by the human hearing system. The distance between the starting point and the peak displacement is called *amplitude*. What should be also noticed is that the shape of the waveform remains the same throughout the propagation process, if the energy dissipation caused by the medium itself can be neglected. A wave with such a property is called a *travelling wave*.

**Speed of Sound**

The speed of a wave is usually taken as constant in the same medium and is determined by the density and stiffness of the medium.

Density (kg·m$^{-3}$) means the mass per unit volume of a substance, which is denoted as $\rho$ and defined as:

$$\rho = \frac{m}{V} \tag{2.1}$$

in which:

$m$ is unit mass of the medium (kg).

$V$ is the unit volume of the medium (m$^{-3}$).

Stiffness refers to the resistance of the medium to a uniform compression, which is defined as:

$$K = V\frac{\partial p}{\partial V} = \rho\frac{\partial p}{\partial \rho} \tag{2.2}$$

in which:

$K$ is bulk modulus (N·m$^{-2}$) which is used to describe the elasticity of a substance, i.e. its ability to oppose deformation, and to restore itself to equilibrium in response to a deformation.

$\frac{\partial p}{\partial V}$ is the derivative of pressure with respect to volume.

$\frac{\partial p}{\partial \rho}$ is the derivative of pressure with respect to density.

Both the density and the stiffness are properties of the medium. According to these two factors, the speed $v$ of a longitudinal wave can be calculated as:

$$v = \sqrt{\frac{K}{\rho}} \tag{2.3}$$

According to the theory of thermodynamics and acoustics, air can often be seen as an ideal gas and follows the ideal gas law [1]. For some extreme cases such as explosions and supersonic booms with sufficiently large amplitudes, the ideal gas model is not accurate enough to describe the phenomenon, and *non-linear* effects of acoustics must be considered, such as dispersion and shock waves [1]. However, these extreme circumstances are rarely encountered for auralisation problems. Therefore, the ideal gas model and linear acoustics is assumed throughout the rest of this thesis.

According to the ideal gas model, the stiffness denoted by $K_{air}$ and density of air denoted by $\rho$ are defined as:

$$K_{air} = \gamma p \tag{2.4}$$

$$\rho = \frac{pM}{RT} \tag{2.5}$$

in which:

$\gamma$ is the Adiabatic Gas Constant which describes the relationship between heat capacity at constant pressure and heat capacity at constant volume for a gas. For dry air, $\gamma$ can be seen as a constant value of 1.4 [1].

$p$ is the air pressure (Pa).

$M$ is the molecular mass of gas (kg·mole$^{-1}$).

$R$ is a constant number called the *gas constant*, representing the required energy per temperature increment per mole for the gas. This value is standardised by the International System of Units (SI Units) as approximately 8.314J·K$^{-1}$mole$^{-1}$ for the ideal gas model.

$T$ is the absolute temperature in Kelvin (K).

Considering Equations 2.3 - Equation 2.5, the speed of sound $c$(m·s$^{-1}$) can be calculated by:

$$c = \sqrt{\frac{\gamma RT}{M}} \tag{2.6}$$

As can be seen from Equation 2.6, the speed of sound in air depends primarily on the temperature of the air. For example, at 20 Celsius degrees the speed of sound can be calculated approximately as 343m/s.

### 2.2.2 Frequency, Wavelength, and Phase

The motion of a particle, moving back and forth, returns through its equilibrium point twice in one complete cycle. The time taken for a complete cycle is known as the *period*, denoted by $T$. According to the definition of frequency in Hertz (Hz) and period in seconds (s), these two factors are reciprocal with each other, as $f = 1/T$.

Since waves propagate at a specific rate in the same medium, we sometimes talk about space and time in a comparable way, which leads to the concept of *wavelength*. Wavelength is defined as how long the wave transports within one period, denoted by $\lambda$. Wavelength can be measured between the closest two points with the same displacement moving in the same direction. Combining the definition of speed of sound and period above, the relationship between distance and time for a wave function can be found as:

$$c = \frac{\lambda}{T} = f\lambda \tag{2.7}$$

Therefore, for air at a specific temperature, the wavelength of a sound is proportional to its period, and inversely to its frequency. For example, at 20 degrees Celsius the wavelength of a 100Hz sine wave can be calculated approximately as 3.43m.

Phase is used to describe the location of a point within a wave cycle in a repetitive waveform. It can be measured in distance, time, or more often degrees or radius. Usually, no useful information can be obtained with measured phase of only one single wave, but it is of high interest to compare the *phase difference* between two waves. The phase difference is defined as (donating $G$ and $F$ for the two waves):

$$\phi(t) = \phi_G(t) - \phi_F(t) \tag{2.8}$$

When the phase difference equals zero at the time $t$, the two waves are *in phase* which means they perfectly match in terms of peaks, valleys, and zeros. When the positive part of one wave coincides with the negative part of another wave, they are *out of phase.*

### Wave Function in Time and Space

With the definition of amplitude, frequency, wavelength, and phase above, a 1-D wave function can be expressed mathematically as:

$$y(x,t) = A\text{sin}(\omega t \pm kx) \tag{2.9}$$

in which:

$A$ is the amplitude of the wave.

$\omega$ is the angular frequency of the wave, defined as $\omega = 2\pi f$.

k is the wave number, defined as $k = 1/\lambda$.

The negative sign is used when a wave goes in the positive $x$ direction and the positive sign is used for a wave traveling in the negative $x$ direction.

### Wave Superposition

Wave superposition means to add together waves travelling through the same medium simultaneously. The net displacement of the medium at any point in space/time is the sum of the individual wave displacements for that point. Superposition of two waves with the same frequency and amplitude may lead to constructive interference if the waves are in phase, or destructive interference if the waves are out of phase. Mathemetically, the resulting wave function can be written as:

$$\begin{aligned} y(x,t) &= A\text{sin}(\omega t \pm kx) + A\text{sin}(\omega t \pm kx + \phi) \\ &= 2A\text{cos}(\phi/2)\text{sin}(\omega t \pm kx + \phi/2) \end{aligned} \tag{2.10}$$

There will be constructive interference if $\phi = 0$, while destructive interference will occur if $\phi = \pi$ and the two waves will be cancelled with each other. These effects are demonstrated in Fig 2.3.

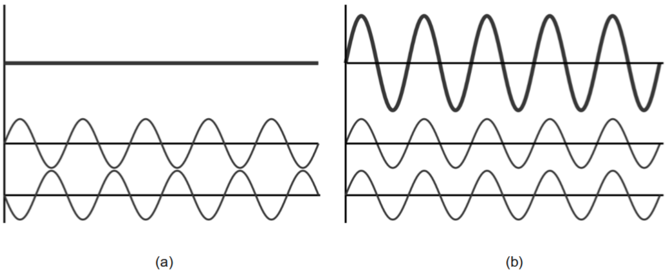


**Figure 2.3:** *Diagram of (a) destructive interference; (b) constructive interference.*

## 2.2.3 Sound Levels and Inverse Square Law

The most used sound level metric is *sound pressure level (SPL)*. Sound pressure $p$ represents a force per unit area in N/m$^2$ or Pa caused by sound waves. It is a scalar property that does not consider the direction of the wave. It is usually not convenient to directly use sound pressure in Pa as a metric to evaluate the energy of the sound, as it has a quite wide range of variation up to $10^6$Pa. Therefore, SPL which is a sound pressure ratio in decibles (dB) scale, is more often used. SPL can be calculated as:

$$SPL(\text{dB}) = 20\text{log}_{10}\frac{p}{p_{ref}} \tag{2.11}$$

in which $p_{ref} = 2\times\ 10^{-5}$Pa is the threshold of the human hearing system in terms of sound pressure.

Another sound level metric is *sound intensity level (SIL)*. Intensity is defined as the power per unit area carried by a wave. Power is the rate at which energy is transmitted. In equation 2.12, sound intensity denoted by **I** is defined as:

$$\mathbf{I} = \frac{P}{\mathbf{A}} \tag{2.12}$$

in which $P$ is the sound power in Watts, **A** is the unit area that the sound power goes through. Both **I** and **A** are vectors, meaning that they have both direction and magnitude. For sound waves in air, a more commonly used equation in terms of sound intensity, is the sound intensity in a plane wave, expressed as:

$$||\mathbf{I}|| = \frac{p_{rms}^2}{\rho c} \tag{2.13}$$

in which:

$\rho$ is the density of air.

$c$ is the speed of sound.

$p_{rms}^2$ is the *root mean square* of sound pressure $p$, defined as:

$$p_{rms} = \sqrt{\frac{1}{T}\int_0^T p^2(t)dt} \tag{2.14}$$

Sound intensity can be also denoted with a level using a decibel scale for convenience, which is SIL, defined as:

$$SIL(\mathrm{dB}) = 10log_{10}\frac{\|\mathbf{I}\|}{I_{ref}} \tag{2.15}$$

where $I_{ref} = 10\ ^{-12}$ W/m$^2$ is the threshold of the human hearing system in terms of sound intensity.

When considering sound propagation in three dimensions as a spherical wave in a free field (free propagation in all directions), the sound intensity in any radial direction from the source (which is considered at the centre of the sphere) can be expressed as a function of distance:

$$\|\mathbf{I}\|(r) = \frac{P}{\|\mathbf{A}(r)\|} = \frac{P}{4\pi r^2} \tag{2.16}$$

in which $r$ is the distance from the spherical centre along the radial direction. This equation indicates that for a spherical wave in a free field, sound intensity $\|\mathbf{I}\|$ is inversely proportional to the square of the distance from the source $r$. This corresponds to the idea of *inverse square law*, and can be used to explain why a sound source is perceived as being louder when close by and quieter when far away, while the sound power of the source keeps the same. For a spherical wave that propagates a long distance, the wavefronts are similar to plane waves, for which the relationship between sound intensity and sound pressure is similar to $I \sim p^2$, and the inverse square law can be expressed as $p \sim 1/r$.

### 2.2.4 Reflection and Absorption

When waves travel to a boundary between two mediums, only part of the energy can be transmitted from the original medium to the other, with the other part returning to the original medium. This phenomenon is called *reflection*. The portion of energy transmitted and reflected, and the pattern of the waves that return depend on the properties of the two mediums.

In terms of the direction of sound wave reflection, the basic law is Snell's law [1], which is the same as the law of light reflection, saying that, for specular

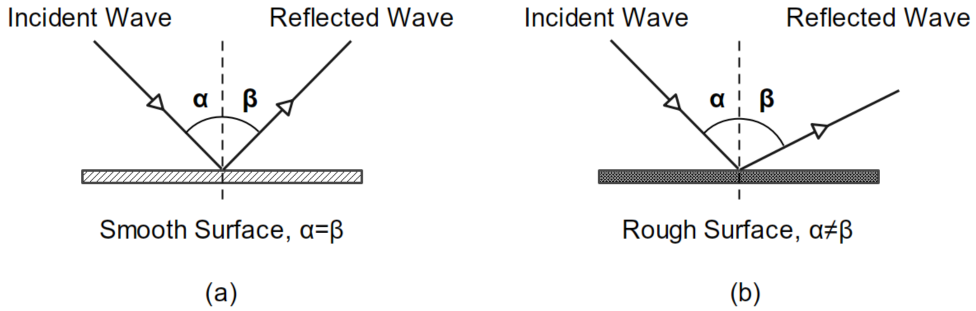


**Figure 2.4:** *Sound wave reflections: (a) Specular reflection happens on a smooth surface; (b) Non-specular reflection happens on a rough surface.*

reflection, the angle at which the wave is incident on the surface equals the angle at which it is reflected. When a sound wave in air travels to a boundary that is extremely smooth, the reflection can be considered as being specular. If the boundary is rough with minor irregularities in geometry, the resultant reflection will not be perfectly specular because of the diffusion caused by the roughness of the surface, as demonstrated in Figure 2.4.

When a sound wave in air is incident upon a new boundary, the portion of energy transmitted and reflected for the incident sound depends on the property of the new boundary and the incident angle of the sound wave. In acoustics, this property can be described by a metric called *acoustic impedance*, denoted by $Z$, which is defined as the ratio between the pressure $p$ and the particle velocity $u$ in a wave at normal incidence:

$$Z = \frac{p}{u} \tag{2.17}$$

Here both $p$ and $u$ are complex numbers including magnitude and phase values, therefore $Z$ is also a complex number. Acoustic impedance is an inherent property of the medium. For sound propagation in air as plane waves, the acoustic impedance can be calculated as [1]:

$$Z = \rho c \tag{2.18}$$

where $\rho$ is the density of air, and $c$ is the speed of sound in air.

When waves travel through two different mediums, the velocity and local pressure on the boundary must be continuous [1]. The proportion of reflected sound energy relative to the transmitted sound energy can be derived if the acoustic impedance of both mediums are known. Reflection occurs when there is a mismatch in the impedances between two mediums separated by the boundary. This can be further explained via a 1-D single frequency wave model, as shown in Fig 2.5. The pressure and the velocity of the 1-D incident wave can be expressed as:

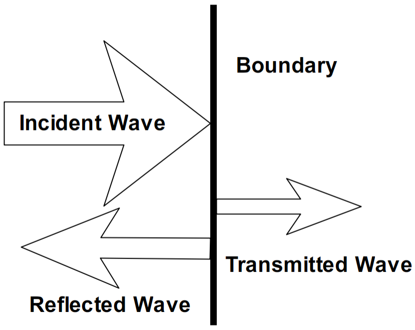


**Figure 2.5:** *A 1-D single frequency wave interacting with boundary with reflection energy and transmission energy.*

$$p(x,t) = A_0 e^{i(\omega t - kx)} \tag{2.19}$$

$$u(x,t) = \frac{A_0}{\rho c} e^{i(\omega t - kx)} \tag{2.20}$$

When the wave is incident upon the boundary, some of the energy will be transmitted into the new medium, while the rest of energy will be reflected. The amplitude of the reflected wave will be smaller than the incident wave. The reduction of amplitude is described by a factor $R$ called the *reflection coefficient*, $-1 < R < 1$. Under phase-preserving absorbing boundary conditions (the phase of the incident wave is identical to that of the reflected wave when reflection occurs), the reflected pressure $p_r$ and velocity $v_r$ can be expressed as:

$$p_r(x,t) = RA_0 e^{i(\omega t - kx)} \tag{2.21}$$

$$u_r(x,t) = -R\frac{A_0}{\rho c}e^{i(\omega t - kx)} \tag{2.22}$$

The minus symbol in Equation 2.22 indicates a reversal in propagation direction. According to the definitions above, the acoustic impedance $Z_b$ of the boundary can be expressed in terms of the reflection coefficient as:

$$Z_b = \rho c \frac{1+R}{1-R} \tag{2.23}$$

According to Equation 2.23, some extreme cases of acoustic impedance can be defined as follows:

- R = -1, $Z_b = 0$: Soft boundary from which an incident wave will be fully reflected. The phase of the outgoing wave will be inverted due to the negative reflection coefficient.

- R = 0, $Z_b = \rho$c: A completely absorbing boundary surface. The wave will fully transmit to the new medium as the acoustic impedance is identical to air.

- R = 1, $Z_b = \infty$: Rigid boundary which is fully reflective and preserving phase. The reflected wave will differ from the incident wave only respect to a reversed velocity component.

In fact, as $-1 < R < 1$, the wave will be reflected with some phase change when it transmits through two different mediums. The transmitted wave will not come back to the air. In other words, it is absorbed by the new medium. A more commonly used metric related to sound reflection is the *sound absorption coefficient*. The sound absorption coefficient is the measurement of the ratio between the sound intensity of the reflected waves and the normal incident waves. As the reflection coefficient is defined based on amplitude, recalling the definition of sound intensity in Equation 2.13, the incident/reflected pressures in Equation 2.19 and Equation 2.21, the relationship between sound absorption coefficient and reflection coefficient can be simplified as:

$$\alpha = 1 - R^2 \tag{2.24}$$

The sound absorption coefficient is an inherent property of the material and structure relative to the frequency of incident sound. In practice, the values of sound absorption coefficient are often presented in octave or one-third octave bands. It is widely used in room acoustics to control the reverberation time of a space. Some examples of sound absorption coefficients in octave bands for some typical materials are listed in Table 2.1.

Table 2.1: Absorption coefficients for some typical materials, reproduced from [1]

| **Material** | **Absorption coefficients in octave bands** | | | | |
|---|---|---|---|---|---|
| | **125Hz** | **250Hz** | **500Hz** | **1000Hz** | **2000Hz** |
| Wooden Floor | 0.15 | 0.11 | 0.10 | 0.07 | 0.06 |
| Marble | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 |
| Heavy curtain | 0.15 | 0.35 | 0.55 | 0.75 | 0.70 |
| Painted Concrete | 0.10 | 0.05 | 0.06 | 0.07 | 0.09 |
| Brick | 0.03 | 0.03 | 0.03 | 0.04 | 0.05 |
| Ordinary Window Glass | 0.3 | 0.2 | 0.2 | 0.1 | 0.07 |
| Acoustic Ceiling, suspended | 0.5 | 0.7 | 0.6 | 0.7 | 0.7 |

### 2.2.5 Scattering and Diffraction

#### Scattering

In practice, it is difficult or even impossible to find an ideal smooth surface. Thus, when an incident wave travels to a boundary, the reflected wave will be at an angle which is different from the angle of incidence, as shown in Figure 2.4(b). This phenomenon is called *scattering*. Scattering occurs because of the irregularities in the geometry of a surface. The degree of deviation depends on the dimension relationship between the wavelength and the degree of surface irregularity.

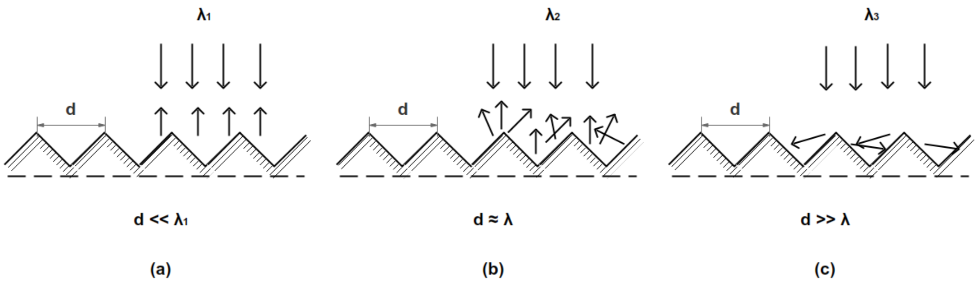


**Figure 2.6:** *An example of the scattering effect for different wavelengths for a surface with same degree of irregularity. (a) $\lambda_1 >> d$, specular reflection. (b) $\lambda_2 \simeq d$, diffusion. (c) $\lambda_3 << d$, specular reflections respect to the surface irregularities.*

Figure 2.6 shows an example of the scattering effect for different wavelengths in terms of the same surface irregularity. The wavelength is denoted by $\lambda$, and the irregularity dimension is denoted by $d$. If the wavelength is much larger than the irregularity dimension (e.g. low frequency waves), the irregularity can be ignored and the reflection will be specular respect to the surface. If the wavelength has a similar dimension to the irregularity, the reflection will be complex and potentially be distributed over a range of different directions, which is called *diffusion*. If the wavelength is much smaller than the irregularity dimension (e.g. high frequency waves), a series of specular reflections will occur but potentially based on the surface irregularities rather than from the surface.

## Diffraction

When a sound wave encounters an obstacle in the medium, a bending effect of waves around the corners of the obstacle can be found, which is called *diffraction*. The diffracting object will become a secondary source of the propagation wave. In practical situations, diffraction often make a listener perceive the sound coming from the direction of the corner itself rather than the original source direction. Low frequencies will bend more obviously than high frequencies for this situation.

Figure 2.7 shows a 2-D example of a car sound diffracted by a wall corner. In this example, the listener receives two sounds from the car: one transmitted

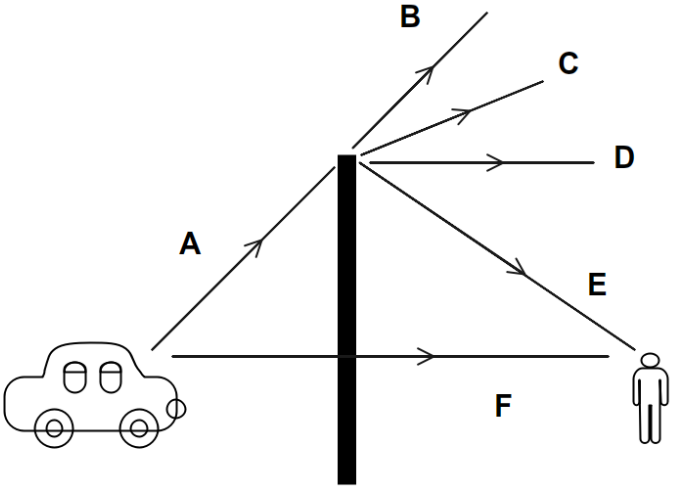


**Figure 2.7:** *An example of a car sound diffracted by a wall corner. Without diffraction, the path A would continue straight as A → B only. The paths A → C, A → D, and A → E represent the bending effect for high frequencies, mid frequencies, and low frequencies in the sound wave, respectively, due to diffraction. The path A → F represents sound transmission through the wall.*

through the wall denoted by path F, and one diffracted around the corner via path A → E. Without diffraction, the path A would continue straight as A → B only. The paths A → C, A → D, and A → E represent the bending effect for high frequencies, mid frequencies, and low frequencies in the sound wave, respectively, due to diffraction.

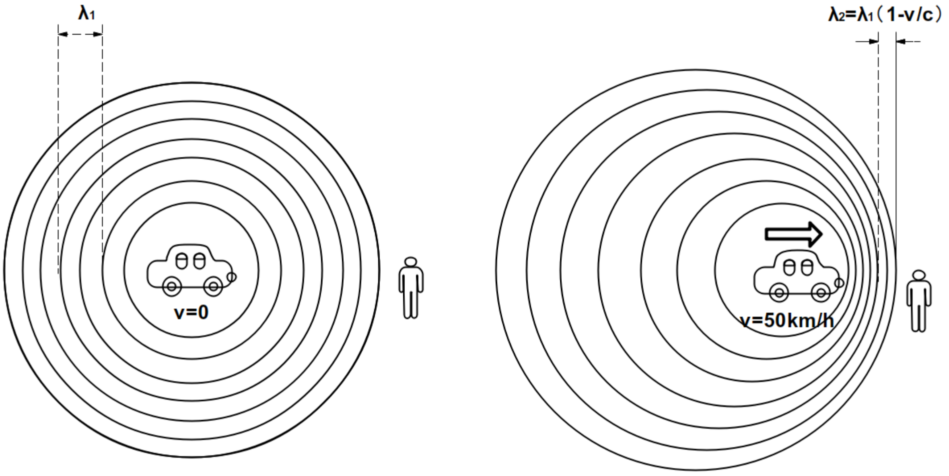


**Figure 2.8:** *Example of Doppler effect: a car is moving towards a static listener, leading to an increase in the frequency of the sound source as perceived by the listener.*

### 2.2.6 Doppler Effect

When a sound source is moving relative to a listener, the source frequency content as perceived by the listener will be shifted up or down compared to the emitted frequencies, according to the source velocity relative to the listener. This phenomenon is called *Doppler effect.* The relationship between observed frequency $f$ and emitted frequency $f_0$ is given by:

$$f = \left( \frac{c \pm \mathbf{v}_{\mathrm{r}}}{c \pm \mathbf{v}_{\mathrm{s}}} \right) f_0 \tag{2.25}$$

in which:

$c$ is the speed of sound in air.

$\mathbf{v}_{\mathrm{r}}$ is the speed of the receiver in vector.

$\mathbf{v}_{\mathrm{s}}$ is the speed vector of the source in vector.

The symbol plus is used when the source is moving towards the listener, while the symbol minus is used when the source is moving away from the listener. Figure 2.8 shows an example of a car moving towards a static listener. As can be seen from this figure, the wavelength is ‘squashed’ along the moving direction when the car travels towards the listener, leading to a positive frequency shift as perceived by the listener.

## 2.3 Sound Perception and Evaluation

### 2.3.1 The Human Hearing System

The human hearing system is responsible for detecting vibrations and the local fluctuations in the air pressure near the ears, and transducing these signals into nerve impulses that can be perceived by the brain. It is usually the case for a normal person that two ears can work together and simultaneously to receive sounds, which is described by the term *binaural.* In order to better investigate how sounds are perceived and evaluated by a person, it is necessary to understand the human hearing system from an anatomical perspective.

The ear is formed of three sections, which are the outer ear, middle ear, and inner ear. Figure 2.9 demonstrates the structure of these sections.

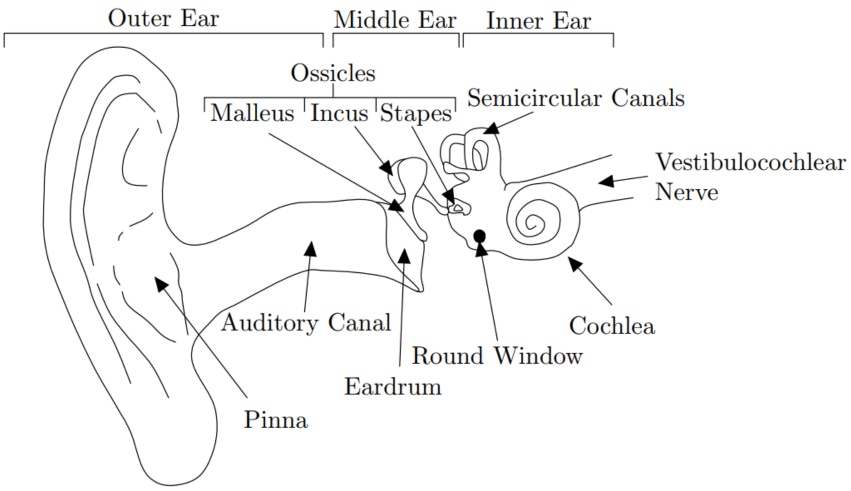


**Figure 2.9:** *The anatomy of the ear indicating the outer ear, middle ear, and inner ear, reproduced from [5].*

The outer ear consists of the external pinna and the outer ear canal. Their main function is to funnel sound towards the middle ear. There are some convex and concave shapes around pinna, which can enhance particular frequencies and aid with sound localisation [31].

The middle ear consists of the eardrum (also called tympanic membrane) and three small bones: the malleus, incus, and stapes, which are known as the ossicles as a whole. The eardrum is the interface between the outer ear and the middle ear acting to transform the airborne incoming vibrations. The ossicles receive the vibrations from the eardrum, and then transmit the vibrations mechanically to the cochlea via the *oval window*.

The cochlea is a coiled structure comprised of three chambers (called *scala tympani*, *scala mediums* or *cochlear duct*, and *scala tympani*, respectively), situated in the inner ear, as shown in Figure 2.10. Two of the fluid-filled chambers (scala tympani filled with *perilymph*, and scala mediums filled with *endolymph*) are separated by a stiff structural element called the *basilar membrane* that runs along the coil of the cochlea. The basilar membrane is tapered in shape, so that it can resonate at different frequencies at different points along its length,

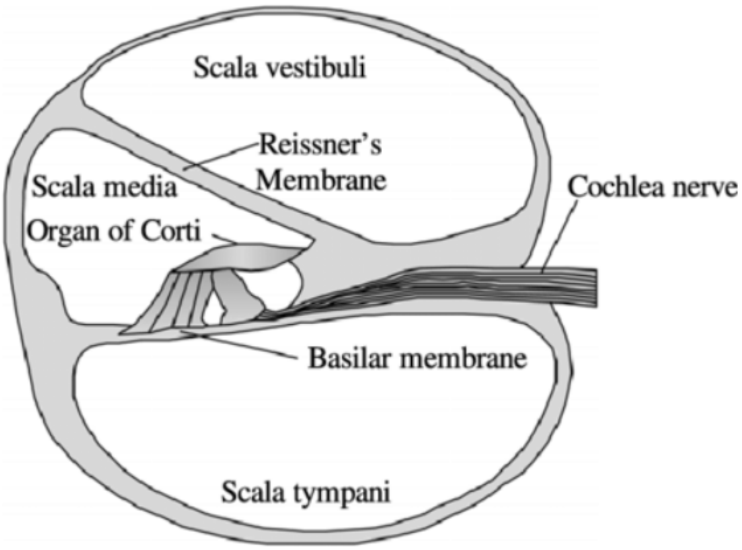


**Figure 2.10:** *The cross section diagram of cochlea, reproduced from [6].*

ranging from 20Hz at the apex to 20kHz at the base. These resonances activate the *Organ of Corti* which a set of hair-like cells called *stereocilia*, situated along the length of basilar membrane. These cells can transduce the resonances of the basilar membrane into neural impulses, so that the *vestibulocochlear nerve* will be triggered by these impulses and can send the information of sound to the brain.

### 2.3.2 Sound Evaluation in Psychoacosutics

#### Critical Bands

As the basilar membrane can vibrate at different points along its length when activated by input sounds with different frequencies, it is possible to distinguish two components that are of similar amplitude and similar frequencies depending on the extent to which the two displacements on the basilar membrane can be clearly separated. *Critical bands* are used to describe the ability to discriminate two simultaneous pure tones with similar frequencies. When the frequency difference is sufficiently small such that they are within the corresponding critical band, the sensation will be of one tone with 'beats' or a 'rough' effect rather than two tones [29]. If the frequency difference is large and beyond the corresponding critical band, the sensation will be two separate tones. This ability varies according to the frequencies of the two sounds considered, even though the frequency difference between them is identical. For example, it is possible to separate a 440Hz pure tone and a 410Hz pure tone simultaneously as two tones, but a 1230Hz pure tone and a 1200Hz pure tone heard simultaneously will be perceived as one tone with a sense of 'roughness'.

Glasberg and Moore [32] proposed an equation that defines a filter with an ideal rectangular frequency response passing the same power as corresponding to auditory filter, which is a direct measurement of the critical bandwidth in quantity, known as *equivalent rectangualr bandwidth (ERB)*, expressed as:

$$ERB = 24.4 \times [(4.37 \times f_c) + 1] \tag{2.26}$$

in which:

$f_c$ is the cutoff frequency of the filter in kHz (0.1kHz< $f_c$ <10kHz).

ERB is the equivalent rectangular bandwidth in Hz.

Another division of critical bands is proposed by Zwicker [29] known as the *Bark scale*, in which human hearing range in terms of frequency is divided into 24 bands, and each band is referred to as a Bark. In the Bark scale, the bandwidths of the critical bands are small (around 100Hz) for frequencies below 500Hz, and then rise up at an approximate rate for higher frequencies.

### Loudness

In Section 2.2, the concepts of sound pressure level and sound intensity level using a dB scale have been introduced. These metrics can be used to describe the amplitude and energy of a sound wave, but they can not depict exactly the loudness of sound perceived by the human hearing system. In fact, an ear is a pressure sensitive organ that performs as a 'filter', dividing the sound signal into a set of overlapping frequency bands. The sensitivity of our hearing system varies for different frequency bands. This sensitivity can be demonstrated by considering two similar frequencies, and the sensation of how loud these frequencies are. For example, a sound wave with a lower SPL(dB) within a given frequency band might be perceived as being louder than a sound wave with a higher SPL(dB) within another frequency band.

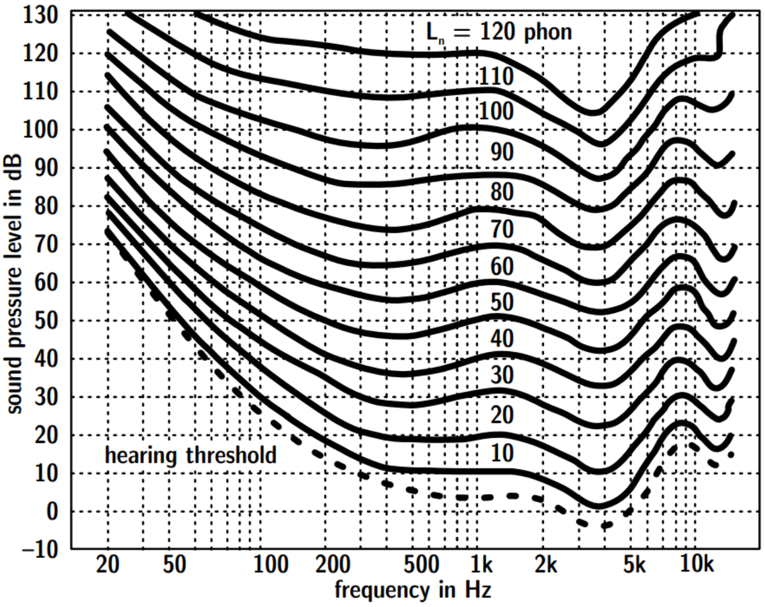


**Figure 2.11:** *Equal loudness contours, reproduced from [7].*

Figure 2.11 shows the *equal loudness contours* for human ears. In this figure, the contours represent the relationship between the measured SPLs and the perceived loudness. The unit of loudness is denoted as *phon*, which is a subjective scale matching the SPL of a given sound with its perceived loudness. A pure tone at 1kHz is used as a benchmark for these contours. That is, for a pure tone at 1kHz the SPL values is identical to the phon values. The contour 0 phon (the dashed line in Figure 2.11) represents the SPL(dB) hearing threshold for human ears at each frequency.

Another unit of subjective impression of loudness is *sone* which creates a linear scale rather than a dB scale. One sone is defined as the loudness of a 1 kHz pure tone at 40 dB sound pressure level. The loudness in phon ($L_N$) and sone ($N$) can be approximated by the following equations [29]:

For loudness above 40 phone or 1 sone:

$$N = 2^{0.1L_N-4} \tag{2.27}$$

$$L_N = 40 + 10log_2 N \tag{2.28}$$

For loudness below 40 phon or 1 sone:

$$N = (\frac{L_N}{40})^{2.86} - 0.005 \tag{2.29}$$

$$L_N = 40(N + 0.005)^{0.35} \tag{2.30}$$

As the objective values of SPL and SIL do not correspond to the perceived loudness at some frequencies, a series of filters based on the equal loudness contours have been developed in order to approximately account for the ears' sensitivity to different frequencies. Some widely used filters including A-filter, B-filter, and C-filter, etc. SPL weighted with the A-filter is denoted as dB(A), the B-filter as dB(B), and C-filter as dB(C), respectively. These filters have been standardised by International Electrotechnical Commission (IEC) in IEC 61672: 2013 [33], as shown in Figure 2.12 (in practice the B-filter has been removed from the standard because it is no longer in common use currently). Because of the nonlinear behaviour of the hearing system, it is not feasible to use a single filter to describe the subjective loudness appropriately at all sound levels. For example, dB(A) is the most reliable approximation for normal sound levels in our daily lives, while dB(C) can be more accurate when the sound is very loud, e.g. over 100dB [1].

### Auditory Masking

In our daily lives, it is usually the case that we receive a variety of sounds coming from multiple directions at different times, and hear them together. Sometimes it can be very difficult, or impossible, to distinguish a specific sound from

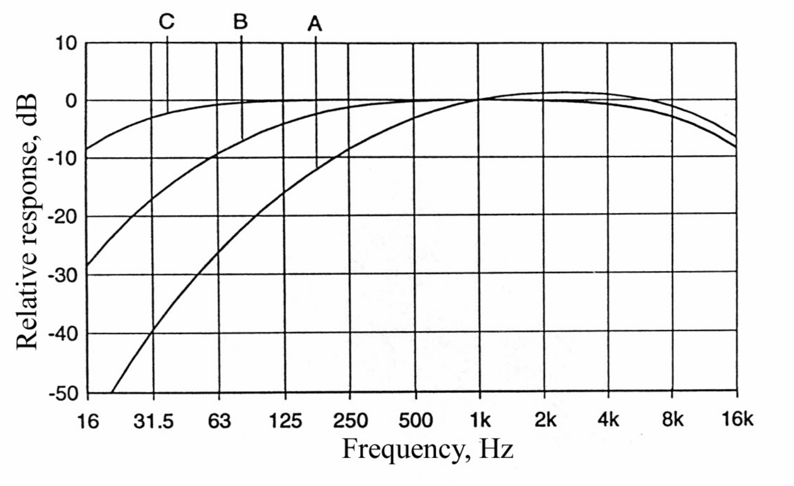


**Figure 2.12:** *A-, B-, and C-weighted sound pressure level filters, reproduced from [1].*

the whole. This phenomenon is called *auditory masking*. Auditory masking can happen either in the frequency domain or in the time domain.

Auditory masking in the frequency domain is called *simultaneous masking*, and refers to when a sound is made partly or completely inaudible by another sound during the same time period. The sound that masks another sound is called a *masker*. For example, a 950Hz pure tone with 60dB in SIL will be fully masked by a 1000Hz pure tone with 80dB in SIL, while a 1050Hz pure tone with 70dB in SIL can be partially masked by the same masker [34]. It is found that low frequency sounds are better maskers for higher frequency sounds, while higher frequencies are much poorer maskers of low frequencies [35]. One of the recent overviews of the characteristics of auditory masking can be found in the paper written by Moore et al. [36].

Auditory masking in the time domain is called *nonsimultaneous masking*, which refers to when a sound is masked by a masker presented just before or after the masker. There are two types of non-simultaneous masking: 1) pre-masking (also called *backward masking*) is when a quiet sound immediately preceding a loud sound is not heard; and 2) post-masking (also called *forward masking*) is when a quiet sound following the masker sound is not detectable. According to the study in [30], the characteristics of pre-masking and the reasons for this phenomenon are still poorly understood. Some properties of the post-masking phenomenon regarding the relationship between the sound levels, frequencies, and masking duration have been discovered, which are detailed in [30]. In general, the effective time duration for post-masking (at least 50ms) is longer than that of pre-masking (less than 50ms), and the effectiveness of the post-masking relies on the relationship between the maskers and the masked sounds in terms of the loudness, frequency, and time interval.

#### Sound Quality

So far, there has been no formal definition of the concept of *sound quality*. Generally, it refers to the concept of the audible experience of a product for a use when compared with users' expectations. As expectations may vary a lot from person to person, the ultimate goal of sound quality studies is to change the perceived sound of a product to improve customer satisfaction and thereby make the product more competitive on the market. According to studies in marketing, sound quality has a strong relationship with some non-auditory concepts, such as luxury, sporty, safety, etc.[37]. In some industries, such as automobile and consumer electronics, sound quality has been integrated as part of the design and evaluation process for product suitability. For example, some car manufacturers have realised that the engine sound is an important part of the overall impression of a car [37]. Sound quality for an engine does not always mean quieter. Some customers are fond of a quiet driving environment, while some others prefer a tone of engine roar which makes them feel 'sportier' or more 'luxurious'. Consequently, some automobile companies have used loudspeakers in the cabin with audio signal processing techniques to create a 'fake engine sound' for some vehicle models to enhance the user experience in terms of overall sound quality and the interaction with pedestrians [38]. These artificial engine noises can also be used to identify a vehicle model. In the case of electric cars, these sound are mandatory for the safety of pedestrians [39].

There are a variety of methods for sound quality testing. These methods can be divided into two categories, which are subjective listening tests and objective metrics tests. The aim of these tests is to provide guidance for the product design. For a subjective listening test, it is essential to design the test process centred around a given context carefully in order to prevent results being biased. As it is impossible to remove all of the contextual impacts, defining and controlling the uncertainties appropriately for the research questions proposed is crucial. Some well-documented guidance for designing and conducting listening tests can be found in [28, 29].

Although subjective listening test is a reliable method to get user opinion/comment data directly, it is often time-consuming to carry out and can be challenging to assemble a statistically significant sample of users from across the whole population. Therefore, some researchers aim to discover objective metrics that directly relate to subjective evaluation. It is much more convenient to evaluate sound quality based on calculating objective metrics, if there are appropriate ones available. One of the most used metrics is A/B/C-weighted sound pressure levels as introduced previously in this section. Some other metrics that are in use for sound quality evaluation (mainly in the automobile industry, and partly in the domestic appliance industry, soundscape studies, etc.) include loudness (which has also been discussed in this section), sharpness, roughness, and fluctuation strength, etc. [29]. What should be noted is that it is not always possible to find appropriate metrics that correlate with the results of a subjective listening test. Even for some metrics that have been proved to correlate well with subjective responses for a given context, they might not be such useful in a different context.

### 2.3.3 Spatial Hearing

So far, all sound perception and evaluation fundamentals that have been introduced are based on 'single ear'. Albeit most experiments behind these findings are conducted binaurally, the stimuli in the two ears are considered as being the same, so the perceived differences between the two ears are discarded. In this subsection, the perceived difference between two ears is discussed, which is important for our spatial experience of sound perception.

Spatial hearing largely depends on the time and level differences between sound arriving at two ears. These differences are mainly due to the difference in the relative positions of the left/right ear to the source, and also the spectral variations caused by the shape of outer ear (pinna), head, and torso [40].

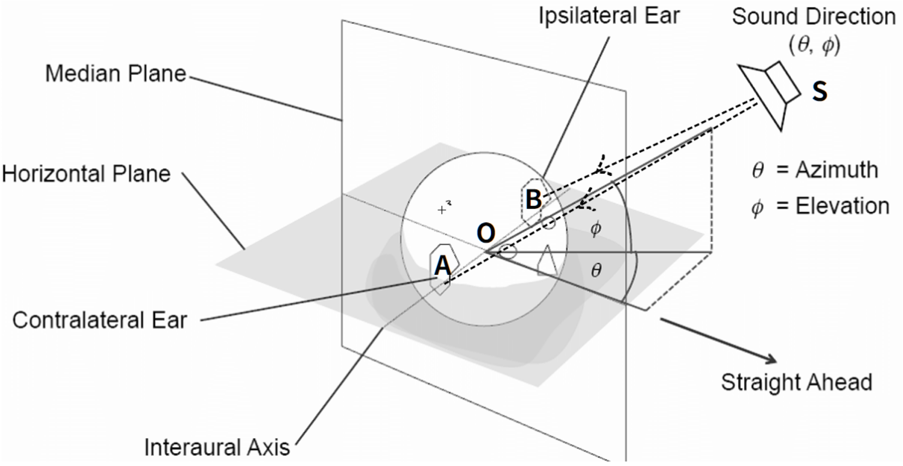


**Figure 2.13:** *A spatial hearing diagram showing the positions of head, ears, and the sound source, reproduced from [8]. The sound source is denoted by $S$. The centre point between the two ears in the head is situated at point $O$. The source direction is expressed by angle $\theta$ for the azimuth, and $\phi$ for the elevation, respectively. $B$ is the ear on the side closer to the source. $A$ is the ear farther away from the source.*

A diagram of a human head with two ears and a sound source some distance away from the subject is shown in Figure 2.13. The sound source is denoted by $S$. The centre point between the two ears in the head is situated at point $O$. The head is facing towards straight ahead, in parallel with the horizontal plane and the medium plane. The source direction is expressed by angle $\theta$ for the azimuth, and $\phi$ for the elevation, respectively. The ipsilateral ear, denoted as $B$, is the ear on the side closer to the source. The contralateral ear, denoted as $A$, is the ear farther away from the source. As can be seen from the figure, a sound travels a longer distance to the contralateral ear than to the ipsilateral ear ($SA > SB$), and the head occlusion is not symmetrical to the two ears. Both of these phenomena bring subtle differences in terms of arriving time and sound levels at the two ears. The two most used metrics to describe these differences are Interaural Time Difference (ITD) and Interaural Level Difference (ILD). Both ITD and ILD are imoportant for sound localisation, but their effectiveness varies according to the frequency of the sound. It has been found that ILD is more reliable for sound localisation for higher frequencies (above 1.5kHz), while ITD plays a better role for lower frequencies (below 1.5kHz) [29]. Apart from ITD and ILD, sound localisation is also influenced by the outer ear, head, and torso, because the spectral information of a sound is changed in a directional dependent manner by these structures to some extent.

Sound localisation on the vertical plane is much weaker compared with that on the horizontal plane. According to some relevant studies [31, 41], it has been concluded that the ability for vertical localisation primarily relies on pinna cues which will bring more ambiguity compared to horizontal localisation cues. In the real environment, vertical localisation is usually realised by turning and tilting head, in order to obtain horizontal cues that help the brain to make more precise localisation in the three dimensional space [31, 41].

Apart from sound localisation, our spatial experience of sound perception may also support the analysis of some other auditory cues. One important phenomenon is the *binaural masking level difference (BMLD)*, which can be stated as: 'whenever the phase or level differences of a signal at the two ears are not the same as those of a masker, our ability to detect the signal is improved relative to the case where the signal and masker have the same phase and level relationships at the two ears [30]'. In other words, if a sound signal and a masker are located at different directions, it is easier for us to detect the signal in binaural listening conditions. There have been several studies on the measurement of the improvement of ability of detection for the BMLD [36, 42–44], and it is found that the threshold of signal detection from noise can be decreased by 3–15dB for the binaural listening when compared to mono hearing conditions [30]. Although the BMLD appears to be related to the well-known 'cocktail party effect' [45], this phenomenon has not been fully understood, and needs to be further investigated. It is claimed that such binaural processing is not always effective in terms of the improvement of detecting signal from maskers, and in some situations it appears to play little role [30].

## 2.4 Summary

This section has discussed the fundamentals of acoustics for auralisation and has provided a context for the further chapters of this thesis. This context includes the basic knowledge of sound generation mechanisms, the basic properties of sound waves, and the concepts of sound perception and sound evaluation methods. It is helpful to be familiar with this content to form a suitable knowledge level for this thesis. The acoustic concepts reviewed and discussed in this chapter will be utilised in the following chapters as background knowledge for the further discussion on the relevant fields, including auralisation, procedural audio, and environmental sounds. The next chapter will focus on the fundamentals of auralisation, including a series of sound source and propagation modelling methods, and some spatial audio reproduction techniques.

# Chapter 3

# Auralisation

Before discussing *auralisation* techniques, it is important to define the meaning of this term. Auralisation can be considered as the analogue of visualisation but for auditory perception. The first definition of the term 'auralisation' was proposed by M. Kleiner et al. [46] in 1993 as follows:

> *Auralisation is the process of rendering audible, by physical or mathematical modelling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modelled space.*

The authors pointed out the listening experience of the target-oriented tasks and three key aspects when doing auralisation. Firstly, audio signals must be generated throughout the process using physical or mathematical methods; secondly, the source must be represented in the context of a specific acoustic environment, providing audible clues as to the nature of the space it occupies; and thirdly, the binaural auditory perception at specific locations should be considered. Before this definition, most of the studies related to acoustic modelling focused on only parts of these three aspects, evaluating the results with objective metrics (e.g. SPLs) rather than binaural auditory perception. It is the concept of auralisation that links these aspects together from a holistic perspective with a clear goal for listening experience.

From a historical point of view, auralisation techniques have developed from ‘physical’ into ‘digital’. The first auralisation model was developed by Spandock et al. in the 1930s [47]. Though the specific term had not appeared during that time, the motivation was the same, which was finding a way to understand how a room would sound like without actually being inside it. They first built a 1:5 scale model of the studied room, and implemented a series of controlled acoustic experiments in both scale model and full-sized room. Both the scale model and the full-size room were sonified by speech signals, and binaural recordings were played back to compare speech intelligibility. A series of scale modelling work has been reported in the field of architectural acoustics during the following decades, with different scaling factors, materials, and impulse response processing methods utilised [47–49]. This ‘physical’ auralisation technique by scale model has not been widely applied in the industry, mainly because it is too expensive and time-consuming to build a scale model and conduct acoustic experiments (taking 11-32 weeks typically [50]) for auralisation applications. Moreover, it is difficult to simulate air absorption effects in a scale model, and/or impossible to find appropriate transducers in terms of size and directivity patterns suitable for a specific scale factor. These inherent deficiencies impede the development and implementation of auralisation using physical models in a wide context.

Compared to ‘physical’ auralisation by scale modelling, ‘digital’ auralisation has unique advantages in terms of saving time and costs. With the rapid development in the hardware and software for digital signal processing, it takes just hours or even minutes to build a computational model for auralisation and conduct acoustic analysis. The theoretical basis for digital auralisation can be traced back to the 1960–70s, when Schroeder proposed a methodology to simulate room acoustic properties for subjective evaluation before a space is constructed [51, 52]. One of the first digital auralisation systems was implemented by Pösselt et al. [53], in which an *image source* model based on purely specular reflection was applied for a rectangular room. There has been fast development of digital auralisation based on computational acoustics since the early 1990s, especially with the launch of commercial software for industrial applications in architectural acoustics, e.g. ODEON [54], CATT [55], and EASE [56], etc. These commercial tools utilised *geometrical acoustic methods (GAMs)* including *ray-tracing* and *image source* algorithms to simulate acoustic phenomena in rooms. The wide application of GAMs has illustrated that these models are computationally efficient, offering acceptable results of accuracy for commercial usage in practice.

With the rapid development of computational power and resources in recent years, there have been studies trying to integrate auralisation with other techniques for broader applicability, such as Virtual Reality (VR) and Augmented Reality (AR). Some real-time auralisation tools [57–60] have been proposed, from which some methodologies can be taken for reference when creating interactive audio cues in a virtual environment. In order to enhance the user experience of VR/AR, the auralisation should be dynamic and immersive, satisfying conditions for an unbroken and unconscious feeling of the surroundings. This is not an easy task, and will be further discussed in Section 3.2.

As the development of 'digital' auralisation is much faster than 'physical' auralisation, some recent definitions concentrate mainly on the digital part of auralisation. One of the most used definitions, proposed by Vörlander, states as follows [7]:

> *Auralisation is the technique for creating audible sound files from numerical (simulated, measured, synthesised) data.*

This definition will be used for the rest of this thesis. This chapter will now consider existing methods for building an auralisation framework and some relevant applications of these auralisation models, based on the fundamentals of acoustics discussed in Chapter 2. Although primarily used in architectural acoustics for producing the audible sensation of being inside a specific room, the application of auralisation is being extended into other fields, such as evaluation of road traffic noise, and VR/AR applications with specific considerations on spatial audio perception and real-time performance.

## 3.1 Auralisation Framework Elements

### 3.1.1 Introduction to acoustic systems

The principle of auralisation has been proposed by Vörlander [7], which includes three basic aspects within an auralisation framework: sound generation, sound transmission, and audio reproduction, as shown in Figure 3.1. In this figure, each element is represented by a block. For each block, some typical examples of key factors to be considered are listed. The arrows between blocks represent the *signal flow* within the framework. Here signal flow refers to a series of steps that the sound signals go through in order to be perceived with altered characteristics. For airborne sounds, signal flow can be modelled as a *Linear Time-Invariant (LTI) system*, and the back arrow between sound generation and transmission blocks can be ignored because the properties of sound wave propagation depend on the medium and boundary conditions for any kind of source.

#### Properties of Linear Time-Invariant System

For any LTI system for airborne sounds, there are two basic properties: *linearity* and *time-invariance*.

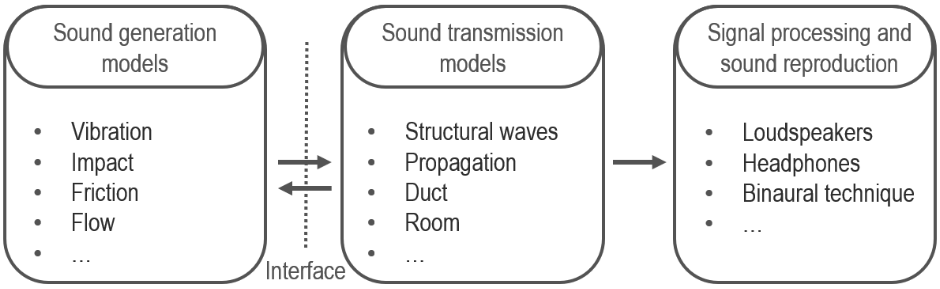


**Figure 3.1:** *Auralisation Framework reproduced from [7].*

Linearity means that the operations of scaling and summing to the input signal will lead to the corresponding linear combination for the output signal. For example, if the input signal $x_1(t)$ produces the output signal $y_1(t)$ and the input signal $x_2(t)$ produces the output signal $y_2(t)$, the input $x_1(t) + x_2(t)$

produces the output $y_1(t)+y_2(t)$, and the input $a_1x_1(t)+a_1x_2(t)$ produces the output $a_1y_1(t)+a_1y_2(t)$ for the scale factors $a_1$ and $a_2$. Mathematically, the linearity property can be expressed as:

$$T[\sum a_i x_i(t)] = a_i \sum x_i(t) = \sum a_i y_i(t) \tag{3.1}$$

in which:

$T$ represents the transformation of a signal fed into a system. $T[x_i(t)] = y_i(t)$, $x_i(t)$ is the input signal with the index number $i$, $y_i(t)$ is the output signal, and $a_i$ is the scale factor of the signal with the index number $i$.

Time-invariance implies that the system alters an input signal the same way no matter when the input signal is applied. Mathematically, for any time shift $t_0$, this property can be expressed as:

$$T[x(t-t_0)] = y(t-t_0) \tag{3.2}$$

In addition to linearity and time-invariance, LTI systems also typically have other properties such as stability, casuality, and invertibility etc.

### Impulse Response in the Time Domain

With the concepts of linearity and time-invariance defined, an LTI system can be described with respect to its reaction to an analytic input signal in the time domain, called the *system impulse response* and denoted by $h(t)$. The output signal of an LTI system can be obtained by convolving the input signal with the system impulse response. Mathematically, for a continuous-time system, the convolution integral of an LTI system can be expressed as:

$$y(t) = T[x(t)] = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau = x(t) * h(t) \tag{3.3}$$

in which $h(t)$ is the impulse response of the system, and the operation * represents the convolution integral.

The impulse response is defined as the output signal of the system when a Dirac delta function is fed in as a impulse signal. The Dirac delta function, denoted by $\delta$, is defined as zero for everywhere except for at the origin point where it is positive infinite, mathematically expressed as:

$$\delta(x) = \begin{cases} +\infty, & x = 0 \\ 0, & x \neq 0 \end{cases} \tag{3.4}$$

This is an idealised function with the property that integral over the entire range of $x$ equals to one:

$$\int_{-\infty}^{\infty} \delta(x)\, dx = 1 \tag{3.5}$$

Another important property of the Dirac delta function is that if a signal is convolved with it, the output signal is identical to the input signal:

$$x(t) * \delta(t) = \int_{-\infty}^{\infty} x(\tau)\delta(t-\tau)d\tau = x(t) \tag{3.6}$$

### Stationary Transfer Function in the Frequency Domain

Apart from the impulse response $h(t)$, the transformation property of an LTI system can be also represented in the frequency domain by a *stationary transfer function* denoted by $H(f)$, which is the Fourier transform of the impulse response, $h(t)$. For continuous-time systems, this is expressed as:

$$H(\omega) = \mathscr{F}[h(t)] = \int_{-\infty}^{\infty} h(t)e^{-j\omega t}dt \tag{3.7}$$

where $\omega = 2\pi f$.

According to the Fourier inversion theorem, it is also possible to reconstruct the impulse response in the time domain via inverse transform. For continuous-time systems, the inverse transform is defined as:

$$h(t) = \mathscr{F}^{-1}[H(\omega)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} H(\omega)e^{j\omega t}d\omega \tag{3.8}$$

As Fourier theory states that any complex periodic waveform can be expressed as the sum of an infinite number of individual frequency components, it is possible to break down the input signal $x(t)$ and the output signal $y(t)$ into their constituent sinusoids, $X(\omega)$ and $Y(\omega)$, respectively. For continuous-time systems:

$$X(\omega) = \mathscr{F}[h(t)] = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt \tag{3.9}$$

$$x(t) = \mathscr{F}^{-1}[X(\omega)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)e^{j\omega t}d\omega \tag{3.10}$$

$$Y(\omega) = \mathscr{F}[y(t)] = \int_{-\infty}^{\infty} y(t)e^{-j\omega t}dt \tag{3.11}$$

$$y(t) = \mathscr{F}^{-1}[Y(\omega)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} Y(\omega)e^{j\omega t}d\omega \tag{3.12}$$

For the Dirac delta function, the Fourier transform equals one for any frequency. In other words, the Dirac delta function has a flat frequency response:

$$X(\omega) = \mathscr{F}[X(t)] = \int_{-\infty}^{\infty} \delta(t)e^{-j\omega t}dt = 1 \tag{3.13}$$

If the input signal and the output signal are known or can be measured, the stationary transfer function $H(\omega)$ of an LTI system can be expressed as:

$$H(\omega) = \frac{Y(\omega)}{X(\omega)} \tag{3.14}$$

If the transfer function is known for an LTI system, the output signal can be calculated by:

$$Y(\omega) = X(\omega) \cdot H(\omega) \tag{3.15}$$

As can be seen from Equation 3.15, the relationship between the input signal and the output signal can be described by multiplication in the frequency domain, which is equivalent to the convolution in the time domain. In practice, convolution in the time domain is often implemented in the frequency domain as multiplication, in order to save computation time.

### Discrete-Time Representation

So far, the properties of an acoustic system and its description in the time and frequency domains have been discussed. All the discussion above is based on continuous-time systems. To process signals digitally, it is necessary to convert continuous analogue signals into discrete-time signals. This is achieved by digitisation of the amplitude and time of a continuous signal. The amplitude is quantised by a binary scale, which is called *bit depth*. The resolution of amplitude using $n$ bits means that the full amplitude scale can be represented by $2^n$ possible values. For example, 16-bit resolution has $2^{16}$=65536 possible values between -32768 and +32767 for the full amplitude scale. The discretisation of time is called the *sampling* process, which refers to measuring the values of the continuous-time signal regularly at specific time intervals. The time interval between two samples is called the *sampling period*, denoted by $T$. The sampling rate (in Hz), denoted by $f_s$, is the number of samples obtained per second, which is the reciprocal to the sampling period:

$$f_s = \frac{1}{T} \tag{3.16}$$

According to the *sampling theory*, the sampling rate should be at least twice the highest frequency of the continuous-time signal to enable accurate representation and reproduction of the information [61]. For sound signals covering the entire human hearing range between 20–20kHz, the typical sample rates are set at 44.1kHz for CD audio, or 48kHz for professional digital video equipment.

With the concepts of bit depth and sampling above, the digitisation of a continuous-time signal $x(t)$ can be described as the measurement of the instantaneous amplitude value for each sample at the time $nT$ ($n = 0, 1, 2, \cdots$, and $T = 1/f_s$). Compared to continuous-time LTI systems, the definition and properties of discrete-time LTI systems are similar, although the form of expression is somehow different. For example, the Fourier transform for a discrete-time system, called the *discrete Fourier transform (DFT)*, is expressed as:

$$X[\omega_k] = \sum_{n=0}^{N-1} x[t_n] e^{-j\omega_k t_n}, k = 0, 1, 2, \cdots, N-1 \tag{3.17}$$

in which:

$N$ is the number of samples.

$x[t_n]$ is the input signal amplitude at $n^{\text{th}}$ sampling instant $t_n$, $t_n = nT$.

$T$ is the sampling period.

$X[\omega_k]$ is the spectrum of the input signal $x[t]$;

$\omega_k$ is the k$^{\text{th}}$ frequency sample, $\omega_k = \frac{2\pi}{NT}$

The corresponding *inverse discrete Fourier transform (IDFT)* is given by:

$$x(t_n) = \frac{1}{N} \sum_{k=0}^{N-1} X(\omega_k) e^{j\omega_k t_n}, n = 0, 1, 2, \cdots, N-1 \tag{3.18}$$

The concepts and properties of discrete-time signals will be utilised throughout this thesis, as the auralisation framework is based on acoustic models with digital signal processing techniques and data stored in digital forms. The rest of this section will discuss each element following the signal flow throughout an auralisation framework, including sound generation, transmission, and reproduction.

### 3.1.2 Sound Source Modelling

The aim of sound source modelling is to obtain a ‘dry’ sound that can be fed into the signal flow blocks in the following steps. Here a ‘dry’ sound can be considered as a source signal free of reverberation and other cues introduced by the sound transmission process [7]. From the perspective of auralisation, source modelling methods can be categorised into two types: *recording based source modelling* and *sound synthesis* techniques. For recording based source modelling, recordings in an anechoic chamber are often used in order to achieve 'dry' sound conditions. In order to catch the spatial properties of a sound source, efforts should be applied on capturing the directivity patterns correctly for the given recorded sources. This is often not an easy task as directivity patterns for some sources are not constant. For example, the radiation patterns of a violin vary significantly according to the frequency content of the tones played [9]. At frequencies below 600Hz, a violin radiates omnidirectionally. Above 600Hz, certain trends are apparent up to 1600Hz in the horizontal plane and 1400Hz in the longitudinal plane as shown in Figure 3.2. The directivity patterns become complex and vary significantly for frequencies above these regions.

Apart from directivity patterns, another challenge for recording based source modelling is the simultaneous capture of multiple sources, e.g. symphonic music. Although in theory it is possible to record an ensemble by instruments one-by-one, and synchronize different takes and parts later to combine as an ensemble [62], this is a tedious process which is time-consuming and not cost-effective. An alternative to this one-by-one method is an *orchestral recording* made with an orchestra in an anechoic chamber [7]. The microphones are located near the instruments or instrument groups in order for the least cross talk possible. However, there is a loss of versatility and flexibility compared to the one-by-one based method because the recordings are taken from the ensemble as a whole. In addition most anechoic chambers are too small for taking orchestral recording of a large ensemble.

For sound synthesis techniques, the plausibility are more concerned than the directivity patterns as the timbre may vary a lot when different sound synthesis methods are used. Here the term *plausibility* can be considered as the

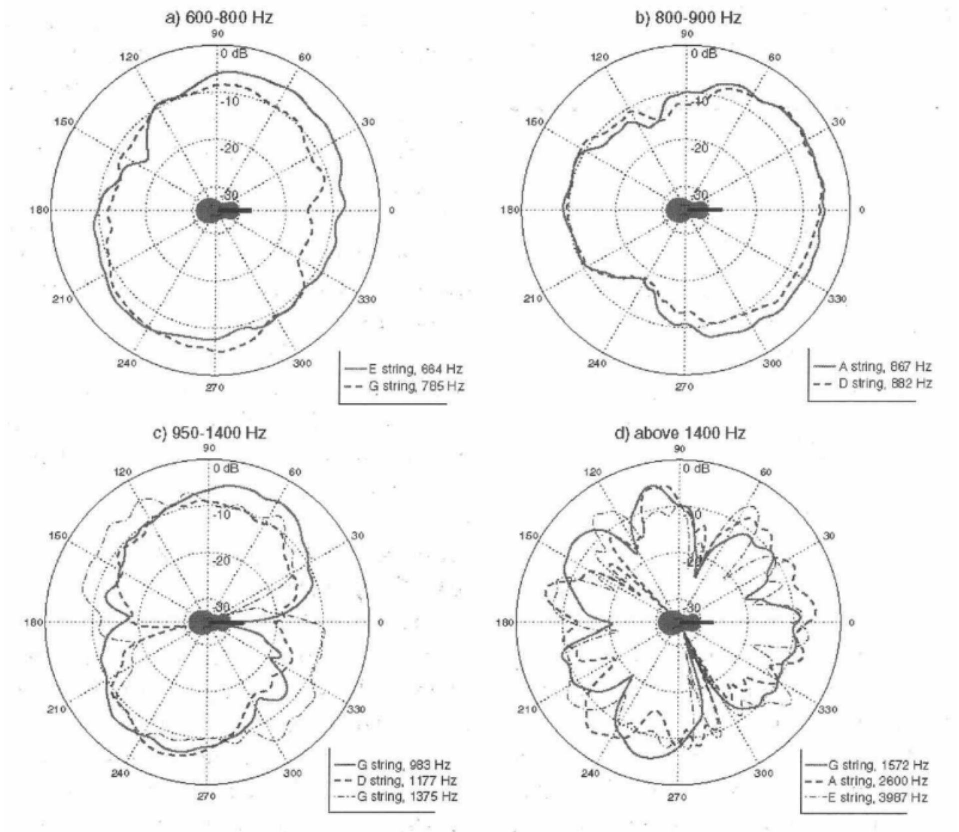


**Figure 3.2:** *Directivity trends of a violin in the longitudinal plane, reproduced from [9].*

agreement of the heard scene with a listener's expectations [63] or compatibility with an external reference (e.g. a playback recording or a real sound) [64]. A 'plausible' auralisation means that people should consider the sound scene credible for the real-world circumstance in which the sound event happens. A detailed overview of some typical sound synthesis techniques will be presented in Chapter 4. Generally, synthetic sounds can be useful or even indispensable under some circumstances, particularly when it is difficult to obtain anechoic recordings, e.g. moving sources in an outdoor environment, walking sounds on different surface materials, gunshot sounds from different weapons, etc. When sound synthesis techniques are applied, it is useful to evaluate the plausibility of the synthesised sounds to satisfy the expectations of most users, although the expectation may vary significantly from person to person. This is similar to the concept of sound quality evaluation introduced in Chapter 2, which can be conducted by subjective listening tests, or calculation of objective metrics if there are suitable ones available.

### 3.1.3 Sound Propagation Modelling

Upon obtaining appropriate source models as excitation signals, the next step is to find a proper method to represent the sound propagation effects within the space to be auralised. As discussed in Section 3.1.1, the signal flow for the auralisation of airborne sounds can be modelled as an LTI acoustic system, which can be described with respect to the system's impulse response. Therefore, one approach of sound propagation modelling is to obtain appropriate impulse responses ensuring faithful recreation of the acoustic characteristics inherent to the environment.

The most straightforward way to achieve this is to take measurements of the impulse responses in the target space. There are numerous studies on specific measurement techniques in terms of the excitation signals, source specifications, and receiver specifications used, including mono-recording, stereo-recording, and multi-channel recordings such as sound field microphones [65]. However, it is not always possible or economically feasible to make accurate impulse response measurements for all environments. For instance, it is impossible to take recordings in a space still at the design stages, or a heritage space that was different from its past configurations at some point. Hence, it is necessary to use numerical methods to simulate the impulse responses based on known information about the space.

There is a wide range of different numerical methods, all of which have pros and cons. Generally, these methods can be divided into two categories, which are geometrical acoustic models (GAMs) and wave-based models. A brief overview of some typical approaches within these two categories are presented in this chapter. Such an overview can provide useful cues and theoretical background knowledge for the design and development of appropriate sound propagation models for the traffic noise auralisation framework proposed in

Chapter 6.

### Geometrical Acoustic Models

In geometrical acoustics, sound is assumed to propagate as rays. This can be analogous to light being modelled as rays in an optical system. As geometrical acoustics does not describe wave phenomena such as diffraction and interference, these wave effects are often neglected or compensated partially in additional processing steps. This assumption can be considered valid at high frequencies where the wavelengths of sound are sufficiently small compared to the surface dimensions and the overall dimensions of the space [66]. At lower frequencies, GAMs may cause relatively large errors because of the lack of wave phenomena. The threshold of the frequency above which GAMs are acceptable can be estimated by the *Schoroeder frequency* [1], which is given by:

$$f_{schoroeder} = 2000 \left( \frac{T_{60}}{V} \right)^{1/2} \tag{3.19}$$

in which:

$f_{schoroeder}$ is the Schoroeder frequency in Hz.

$T_{60}$ is the reverberation time of the space in seconds.

$V$ is the volume of the space in cubic metres.

For common room-based listening environments, $f_{schoroeder}$ often falls into the range 100–200Hz. Thus, GAMs are often used for modelling the room propagation effects in mid- and high-frequency range above this lower boundary of 200Hz.

There are a wide range of modelling techniques based on geometrical acoustics, either reflection-path-based, such as image source [67], ray tracing [68], and beam tracing method [69], or surface-based such as radiosity [70], and acoustic radiance transfer [71], etc. A complete overview of different geometrical acoustics can be found in [66]. In the context of the development of sound propagation models for traffic noise auralisation as part of this thesis, two of the most widely used geometrical acoustics methods will be further discussed – *image source* and *ray tracing*.

In an *image source model*, sound reflections are treated as emitting from phantom ‘image sources’ which are the mirror positions of the original sound source against all the surfaces in a model. The image source representing the sound reflected specularly against a single boundary is called a *$1^{st}$-order image source*. Then these image sources are reflected against all the surfaces, resulting in $2^{nd}$-order image sources, $3^{rd}$-order, etc. The energy decay of each order image source is determined by the distance between the image source and the receiver, and also the boundary conditions of the interacting surfaces in the model. This process is repeated until a termination condition is met, e.g. up to the threshold of reflection order, or the threshold of energy decay. The impulse response can be obtained by summing all the impulsive signals emitted from each source with appropriate time delay and energy attenuation with respect to the distance between the sources and the receiver. Figure 3.3 illustrates an example of this process in a 2-D rectangular space. In this figure, the original source is denoted by $S$, and the receiver is denoted by $R$. The boundaries are marked in bold. The dashed lines represent the reflected sounds from the $1^{st}$- to $3^{rd}$- order image sources. The reflection order related to each space is denoted by shade where white denotes $0^{th}$ order and dark grey denotes $4^{th}$ order. Further examples of valid image source positions are denoted by black crosses [4].

In practice, image source methods are usually used to simulate early reflections within a space with simple geometry, such as a rectangular room. This is because the computation time goes up exponentially with the increase of image source order. In other words, the complexity of an image source algorithm is $O(N^k)$ where $N$ is the number of boundaries and $k$ represents the $k^{th}$-order image source. Therefore, it is difficult to find all the possible image sources if the image source order is too high or the space shape is too complex

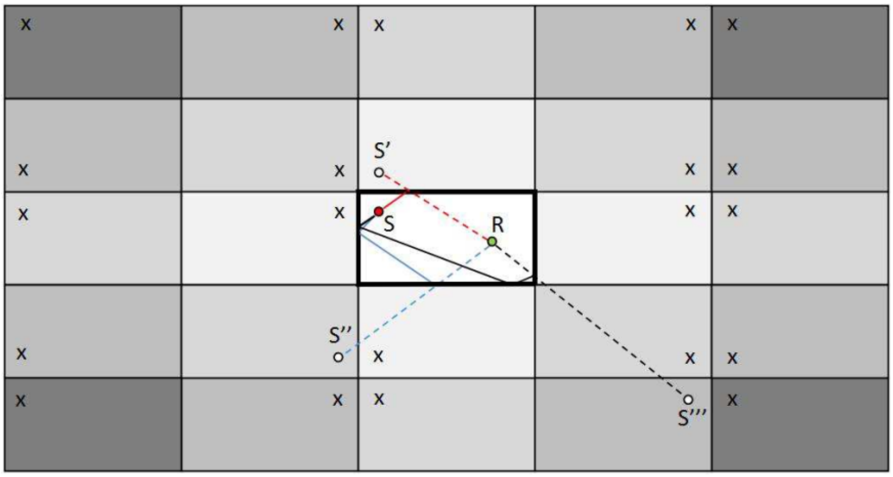


**Figure 3.3:** *Interpretation of the Image Source Method in a rectangular 2-D space, reproduced from [4]. Source and receiver positions are marked as S and R. Examples of $1^{st}$ – $3^{rd}$ order reflections are given alongside their respective image sources denoted as S', S", and S"'. Higher order image sources are denoted '×'.*

in geometry. For some cases when the space is simple in geometry and low-order early reflections are of high interest, it is viable to use the image source algorithms computing at interactive rates, allowing alteration in the acoustic model in real-time [72].

In a *ray tracing model*, sound emitting from a source is assumed to be bunches of sound rays with initial energy travelling in straight lines towards various directions. When a ray hits a boundary, it loses some energy according to the boundary conditions, and is then reflected specularly. The receiver is represented by a spherical space called *receiver volume*. If a ray passes through the receiver volume, the energy and time-delay of that ray will be stored. This process will cease when a termination condition is met, e.g. threshold of energy decay of the ray before arriving at the receiver volume. The impulse response can be obtained by summing up all the data in the receiver volume. Figure 3.4 illustrates an example of this process in a 2-D space. In this figure, sound rays are emitted from the source towards various directions. The dashed lines represent the reflected sound rays with different energy marked with different thickness. The thicker the dashed line is, the more energy it represents. The receiver volume is represented by a round circle in the 2-D space, which corresponds to a sphere in the 3-D space. As can be seen from this figure, the accuracy of ray tracing depends on two main factors. First, as there is limited number of total rays emitted from the source, the number should be sufficient that most of the reflections from effective directions can intersect with the receiver volume; second, the receiver volume should be set appropriately so as not to underestimate or overestimate the ray intersections.

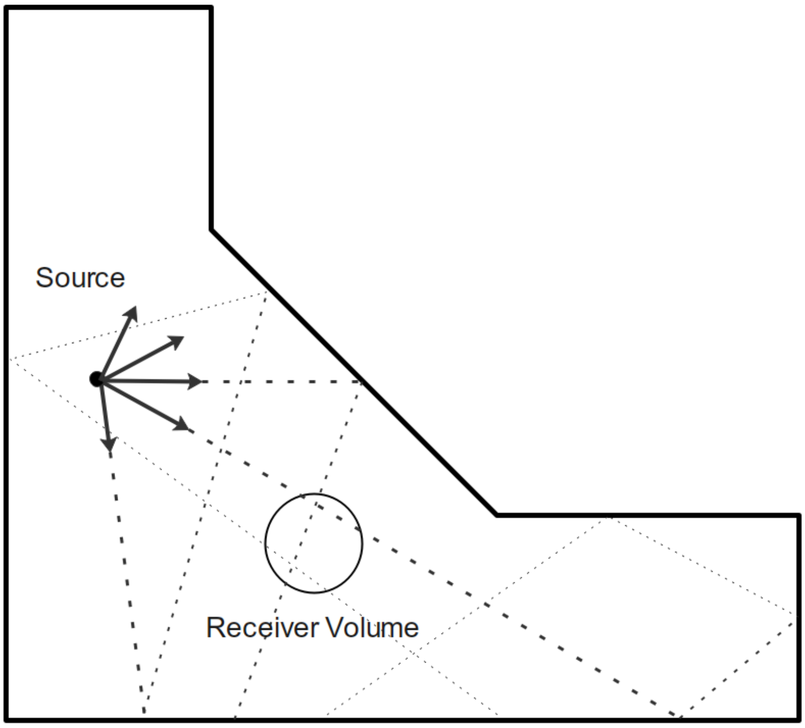


**Figure 3.4:** *Interpretation of the ray tracing in a 2-D space.*

Compared to the image source, ray tracing is less accurate in general, primarily because it is a stochastic process rather than deterministic. For the image source, in theory it is possible to find all exact specular reflections for a given set of source, receiver, and boundary positions. For ray tracing, however, it is random according to the direction of rays, which means that every time the algorithm runs, the rays emit to different directions. Some valid rays might not cross the receiver volume in a specific execution, leading to an underestimation of the result. Although the accuracy of ray tracing methods can be improved by increasing the number of rays and setting the receiver volume appropriately, there is a lack of general guidance for fine tuning such methods [73]. The main strength of ray tracing is that the computation time for high order reflections is much shorter than the image source as the complexity increase is generally linear with the number of rays utilised once the energy decay threshold of the sound rays is determined.

Since both of these GAMs have their pros and cons, it is sensible to combine them together in order to take advantage of their relative strengths. This inspires the development of *hybrid methods*, which make a distinct improvement over the traditional GAMs. The most common hybrid method based in geometrical acoustics is a combination of image source for early reflections and ray tracing for late reflections. This has been applied in some commercial acoustic simulation and auralisation software, such as ODEON [54] and CATT-Acoustic [55].

As GAMs treat sound waves as rays with energy carried, the phase information and wave effects of sound waves are not considered. These effects can be crucial for some cases, particularly for low frequency sounds. For these cases, models that can better describe the physical phenomena are required, which belong to the category of wave-based models.

### Wave-based Models

Wave-based models aims to find a mathematical solution to the *wave equation* across a region of space. The wave equation considers a description of waves as derived from classical mechanics or fluid dynamics, expressed as:

$$\frac{\partial^2 p(t, x_1, x_2, \cdots, x_n)}{\partial t^2} = c^2 \nabla^2 p(t, x_1, x_2, \cdots, x_n) \tag{3.20}$$

in which:

$p$ is the air pressure.

$c$ is the speed of sound.

$\nabla^2$ is the Laplacian operator, $\nabla^2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \cdots + \frac{\partial^2}{\partial x_n^2}$, $x_1, x_2, \cdots, x_n$ and where $n$ is the number of dimensions considered (generally $n = 3$).

Hence, the linear wave equation in three-dimensions is defined as:

$$\frac{\partial^2 p(t, x_1, x_2, \cdots, x_n)}{\partial t^2} = c^2 \frac{\partial^2 p(t, x_1, x_2, \cdots, x_n)}{\partial x^2} \tag{3.21}$$

As the solution to the wave equation is across the whole region of space, the wave phenomena such as diffraction and interference can be preserved inherently and expressed by the solution. Therefore, the results can be theoretically correct throughout the whole frequency range, which is superior to geometrical acoustics methods which are acceptable only for mid- and high-frequencies. Nevertheless, the computational cost is usually much higher for wave-based methods when compared to GAMs. The computational cost depends on the size and density of meshes and nodes used to discretise the space, and the numerical method used to solve the wave equation at these meshes. In theory, the higher target frequency that is required to be solved, the denser the meshes will be needed to obtain that solution. Thus, wave-based methods are often used for modelling low frequencies with feasible cost [74]. For some cases when wave phenomena for mid- or high-frequencies are also of high interest (e.g. acoustic diffusers for high frequencies), wave-based methods are also indispensable in order to simulate these particular wave phenomena. Some popular wave-based methods include *Finite Element Method (FEM)*, *Boundary Element Method (BEM)*, *Finite-Difference Time-Domain (FDTD)*, and *Pseudo-Spectral Time-Domain (PSTD)*, etc. Some detailed overview of wave-based methods for room auralisation and urban sound auralisation can be found in [74, 75]. In order to understand more clearly the strengths and weakness of wave-based methods versus geometrical acoustics methods, three typical wave-based methods, FEM, BEM, and FDTD will be further discussed in terms of their concepts and characteristics in this section.

The *Finite Element Method (FEM)* is an approach to generate discrete algorithms for partial differential equations [76]. The idea is to breakdown the whole space into a series of discrete elements with finite size. Each element is modelled as a damped mass-spring system. The neighbouring elements are concatenated with each other by nodes. In this way, it is possible to model acoustic pressure in a sound field as the displacement of the interconnected masses from the corresponding equilibrium positions. Figure 3.5 illustrates the idea of FEM in a 2-D space, in which $F$ is the driving force applied to the system, $m$ is the mass of an element, $b_i$ is the damping factor, and $k_i$ is the stiffness of spring.

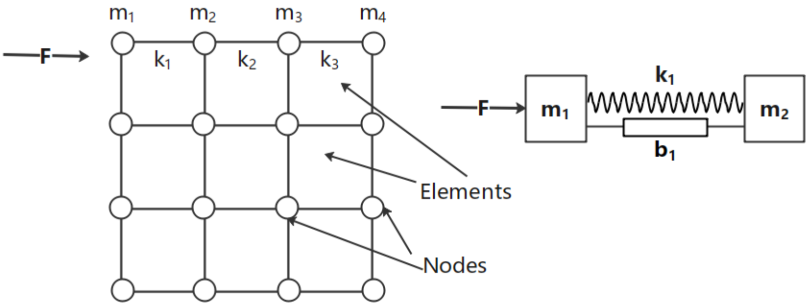


**Figure 3.5:** *Interpretation of FEM with damped mass-spring systems in a 2-D space, cited and reproduced from [5]. Each element is modelled as a damped mass-spring system shown as the right part in the figure. Elements are connected at nodes and can be programmed to respond under different loading conditions shown as the left part in the figure. F is the force applied to system, $m_\mathrm{i}$ is the mass of an element, $b_\mathrm{i}$ is the damping constant, and $k_\mathrm{i}$ is the spring constant.*

Mathematically, the solution to the wave equation by FEM modelling can be expressed as:

$$\mathbf{x} = (\mathbf{K} + j\omega\mathbf{b} - \mathbf{M}\omega^2) \setminus \mathbf{F}(\omega) \tag{3.22}$$

in which:

$\setminus$ represents the pseudo-inverse in matrix calculation.

$\mathbf{x}$ represents the multidimensional air pressure matrix.

$\mathbf{K}$ represents the stiffness matrix.

$\mathbf{b}$ represents the damping matrix.

$\mathbf{M}$ represents the mass matrix.

$\mathbf{F}$ represents the driving force matrix in respect of $\omega$ in the frequency domain.

As can be seen from the equation 3.22, FEM solves the wave equation based on a single frequency. In order to get the complete spectral response in the frequency domain, this process should be repeated throughout the target discrete frequency range. The threshold of the solution in the frequency domain depends mainly on the density of the nodes. For simulation of higher frequencies, the node density should be high enough to represent all the geometrical details. In practical situations, FEM is usually used for simulation of low frequencies in small rooms so that the results can be obtained after taking reasonable computation time [77, 78].

The *Boundary Element Method (BEM)* aims to solve the wave equation formulated in boundary integral equations form, which is a numerical tool for the analysis of boundary value problems for partial differential equations. The key idea of BEM is that the simulated acoustic field can be represented by superposition of fields due to elementary sources located on the boundaries of the space [79]. An integral equation taking the values of variables specified for each of these boundary elements represents an exact solution to the governing wave equation. Once the integral equation is obtained, in post-processing stages, the initial wave equation can be used to calculate the solution at any desired point in the simulation domain [79]. In acoustics, the traditional approach is to numerically approximate the Kirchoff-Helmholtz (K-H) integral equation which is derived from the wave equation using Green's theorem [80], expressed as:

$$c(x)p(x) = -\int_s j\rho\omega v_n(x_s)G(x_s|x) + p(x_s)\frac{\partial G(x_s|x)}{\partial n}ds \tag{3.23}$$

in which:

$x$ is the observed location, and $x_s$ is the unit source location.

$c(x)$ is a constant at location $x$.

$p(x)$ is the complex pressure amplitude at location $x$.

$\rho$ is the density of air.

$\omega$ is the angular frequency in the frequency domain.

$v_n(x_s)$ is the normal surface velocity at location $x_s$.

$G(x_s|x)$ is the free space Green's function relating locations $x$ and $x_s$.

The terminolology $\frac{\partial}{\partial n}$ represents the partial derivative of the function with respect to the unit outward normal at the point $x_s$ on the boundary.

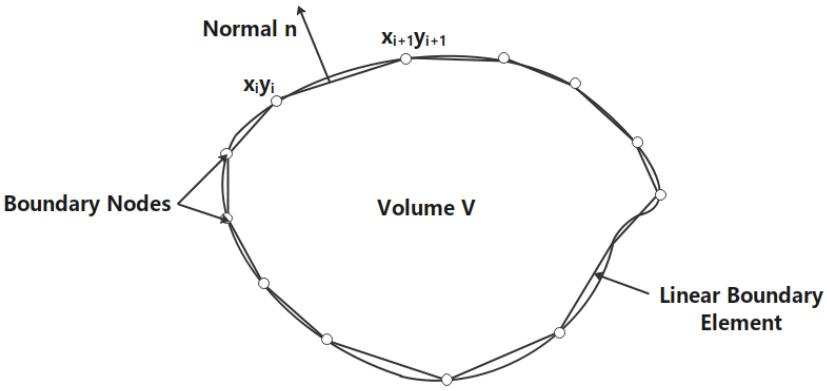


**Figure 3.6:** *Interpretation of BEM with discretised boundary elements and nodes in a 2-D space.*

Compared to FEM, the number of elements can be significantly reduced as only the boundaries in the simulation domain are discretised rather than the whole space, as shown in Figure 3.6, which leads to reduced computation time and memory requirement. However, the matrices used in BEM are usually dense and not symmetrical, for which the computation costs are still considerable [81]. As the common rule of thumb is to use at least six elements per wavelength for a BEM model, the frequency threshold of the solution depends on the size of the elements [82].

The *Finite-Difference Time Domain (FDTD)* method models the simulation domain as a grid of interconnected nodes. While FEM solves the wave equation at a specific frequency, FDTD may perform across a wide range of frequencies in a single simulation. The main idea of FDTD is using forward, backward, or centered finite difference approximations to represent the derivatives of a function. Recalling the 3-D wave equation 3.21, when the finite difference approximations applied, the FDTD method can be expressed as [4]:

$$
\begin{aligned}
\frac{p^{n+1}_{l,m,q} - 2p^{n}_{l,m,q} + p^{n-1}_{l,m,q}}{T^2} = c^2( & \frac{p^{n}_{l+1,m,q} - 2p^{n}_{l,m,q} + p^{n}_{l-1,m,q}}{h^2} + \\
& \frac{p^{n}_{l,m+1,q} - 2p^{n}_{l,m,q} + p^{n}_{l,m-1,q}}{h^2} + \\
& \frac{p^{n}_{l,m,q+1} - 2p^{n}_{l,m,q} + p^{n}_{l,m,q-1}}{h^2})
\end{aligned}
\tag{3.24}
$$

in which:

$T$ is the sampling period representing a small step in the time domain.

$h$ is the spatial sampling distance representing a small step along the x-axis.

$n$, $l$, and $q$ are integers that describe the position of the system in discrete-time and space, respectively.

$p^{n+1}_{l,m,q}$,$p^{n}_{l,m,q}$, and $p^{n-1}_{l,m,q}$ are the air pressure at the location $[l, m, q]$ for the next, current, and previous time steps, respectively, at time $t = nT$.

$p^{n}_{l+1,m,q}$, $p^{n}_{l,m,q}$, and $p^{n}_{l-1,m,q}$ are the air pressure for the current time step at the location $[l + 1, m, q]$, $[l, m, q]$, and $[l - 1, m, q]$, respectively.

$p^{n}_{l,m+1,q}$ and $p^{n}_{l-1,m,q}$ are the air pressure for the current time step at the location $[l, m + 1, q]$ and $[l, m - 1, q]$, respectively.

$p^{n}_{l,m,q+1}$ and $p^{n}_{l,m,q-1}$ are the air pressure for the current time step at the location $[l, m, q + 1]$ and $[l, m, q - 1]$, respectively.

The calculation process of FDTD is recursive. For each step, the terms of the next time are unknown, while the terms of the current and previous steps are obtained during the calculation process. As a wave-based method, FDTD shares the same limitations as the other wave-based methods, e.g. requirement of extensive computational resources, the trade-off between the density of nodes and frequency resolution, and error caused by numerical dispersion, etc. [83]

Apart from these three typical wave-based methods, there are a variety of other wave-based approaches for acoustics, such as Pseudo-Spectral Time-Domain (PSTD) [84], Digital Waveguide Mesh (DWM) [85], and Functional

Transform Method (FTM) [86], etc. An overview of some state-of-art application of wave-based methods for auralisation studies will be presented in Section 3.2. As each method has its advantages and disadvantages, there is no 'one size fits all' solution for specific wave-based methods. With respect to auralisation, wave-based methods are indispensable when the accurate perception of low frequency sounds are of high interest, e.g. auralisation of public transport buses and lorries in narrow city streets with high realism. The considerable computation issue of wave-based methods can be partly solved by using parallel GPU and CPU programming [87], but there are still developments to overcome to make wave-based methods more feasible and cost-effective.

### 3.1.4 Spatial Audio Reproduction

After modelling sound sources and sound propagation effects, the next stage in an auralisation framework is to reproduce the simulated data in an appropriate audible way. In order to get more authentic listening experience, it is imperative to use spatial audio techniques so that natural directional cues can be reproduced, which correspond to our own auditory experience in our daily lives. Two different types of spatial audio rendering and reproduction techniques can be considered, which are *loudspeaker array systems* and *binaural audio techniques* using headphones.

#### Loudspeaker Array Systems

Loudspeaker array systems for spatial audio rendering can be categorised into two general types. Some loudspeaker array systems aim to recreate the whole target sound field within a space so that listeners can move and rotate their heads freely within the sound field. An alternative to the complete sound field reconstruction is to render the sound field accurately only around the listening position called the 'sweet spot'.

Systems aiming at recreating the whole target sound field within a space are usually based on *wave field synthesis (WFS)* [88] methods. The theoretical background of this concept is the Kirchoff-Helmholtz (K-H) integral expressed as Equation 3.23. According to the K-H integral equation, complex wave fields in a volume can be decomposed into simple wave functions such as plane or spherical waves emitted from a surface. The decomposition process is achieved by analysing the signals from a spatially distributed microphone array measurement using the spatial Fourier transform. Figure 3.7 illustrates the concept of WFS, in which the sound field generated from the virtual guitar sound source is decomposed and substituted by a linear loudspeaker array.

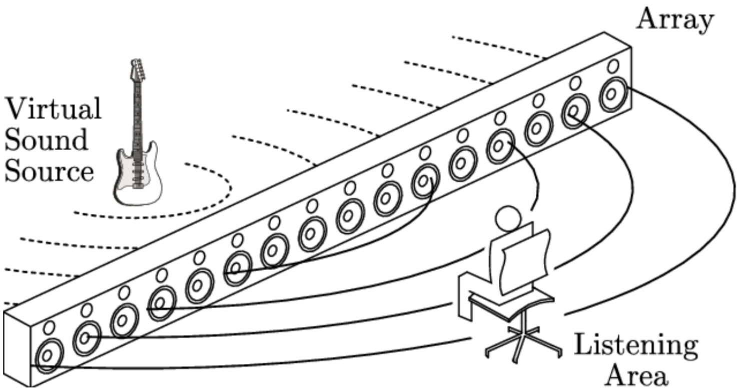


**Figure 3.7:** *Illustration of wave field synthesis, reproduced from [10]. The sound field generated from the virtual guitar sound source is decomposed and substituted by a linear loudspeaker array.*

The higher the spatial sampling of the microphone array, the more precise the reconstructed wave field can be. Unfortunately, this is not always feasible in practice, as too many loudspeakers might be needed, and such arrays are also limited by the physical size of the loudspeakers used. For practical use, the spatial resolution is often simplified according to the characteristics of human hearing. For example, as the human hearing system is more sensitive to sounds on the horizontal plane compared to the vertical plane, it is rational and feasible to use 2-D loudspeaker arrays based on cylindrical waves decomposition to provide essential spatial information [89]. Amplitude corrections for the loudspeaker array are always needed when decomposing the 3-D environment into 2-D cylindrical waves as the distance law is modified during the process. Although simplified this way with compromise in spatial resolution, the number of loudspeakers can be still very high, e.g. 500 loudspeakers required for a 30m$^3$ room for precisely reconstructing the sound field up to 10kHz with cylindrical waves [7].

Compared with WFS, loudspeaker array systems that work for a 'sweet spot' are more feasible and cost-effective as fewer loudspeakers are needed. One of the popular techniques belonging to this category is *Ambisonics* [90]. Ambisonics is a sound capture/rendering method based on a spherical harmonic decomposition of a sound field. The simplest form of Ambisonics is called *B-Format* spatial audio which encodes an audio signal into the 1$^{st}$-order spherical harmonics [90].

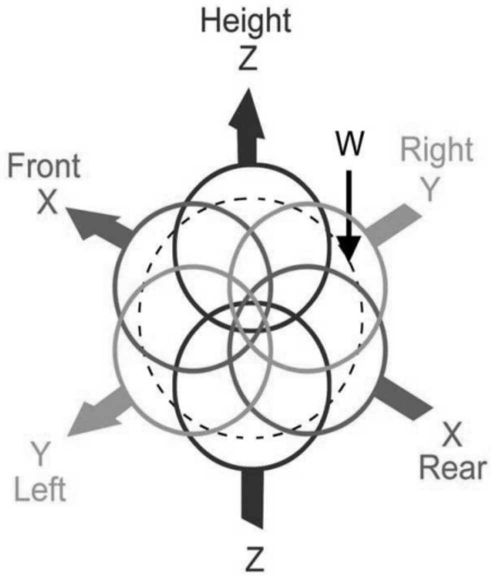


**Figure 3.8:** *Interpretation of the directivity patterns for B-Format channels X (left-right), Y (front-rear), Z(up-down), and W (omnidirectional), reproduced from [11].*

B-Format spatial audio consists of four channels denoted as X, Y, Z, and W. The first three channels refer to the pressure gradient orientated in the directions of left-right, front-rear, and up-down, respectively, while W refers to the pressure recorded in an omnidirectional channel. Figure 3.8 illustrates the directivity patterns of B-format decomposition. For a source signal $S$ encoded into B-Format, which is located at horizontal angle $\theta$ and elevation angle $\phi$ on a unit sphere, the gains for each channel can be expressed as:

$$\begin{aligned} W &= S\frac{1}{\sqrt{2}} \\ X &= S\mathrm{cos}\theta cos\phi \\ Y &= S\mathrm{sin}\theta cos\phi \\ Z &= S\mathrm{sin}\phi \end{aligned} \tag{3.25}$$

Higher order Ambisonics uses higher orders of spherical harmonics and require more channels. Once the encoded signals are obtained, they can be decoded to a specific loudspeaker array. For each loudspeaker, all the channels are summed with different gain and phase according to the direction of the loudspeaker. The loudspeaker array is often arranged evenly on a spherical surface to eliminate the localisation trade-off caused by differences in distance and spatial density of loudspeakers. Compared to WFS, Ambisonics is more viable and cost-effective as fewer loudspeakers are required. However, as the sound reconstruction is theoretically accurate only at the 'sweet spot', there is limitation with respect to free movement of the listener or catering for multiple listeners. Although the 'sweet spot' can be expanded by increasing the order of the system or optimising the decoding scheme [91], this inherent drawback impedes the widespread use of Ambisonics. Currently, one of the most popular applications of Ambisonics is to provide spatial cues and a suitable test environment to improve binaural audio rendering techniques, which means to mix and present the Ambisonic signals on headphones to get a similar audio perception as would be heard via actual Ambisonics loudspeaker array.

### Binaural Audio Techniques

The theoretical background to binaural audio is that if sound pressure variations at the eardrums can be properly captured and reproduced, the resulting sound will be perceived to be as authentic as the corresponding actual sound event. In order to achieve the best possible sense of authenticity, binaural recordings can be implemented using an individual or on a dummy head. A dummy head is designed to replicate an average-sized human head with two ears, a nose, a mouth and upper torso, with tiny microphones located in the ear canals, as shown in Figure 3.9. As head shape and size vary for different individuals, it is impossible to use dummy head recordings to reproduce the exact binaural listening experience for a specific person. Instead, it can be used to capture a general binaural effect that is assumed to be acceptable for most people.

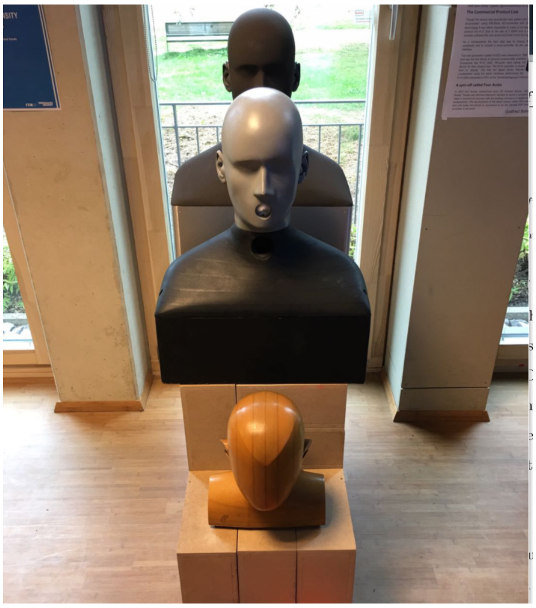


**Figure 3.9:** *Example dummy heads demonstrated in Laboratory on Acoustics in RWTH Aachen University (picture taken by the author).*

In terms of signal processing, these binaural offsets are described as the Head-Related-Transfer-Function (HRTF), which is the Fourier transform of the measured Head-Related Impulse Response (HRIR) at the ear drum for the excitation placed at the source under anechoic conditions. There are various open source HRTF databases online, such as the KEMAR HRTF database [92], the CIPIC HRTF database [93], the FABIAN HRTF database [94], and the SADIE II datebase [95], etc. These HRTF databases are measured based on a sound source at a fixed distance from the dummy head (within 1 to 2 metres). With a database of HRTFs, sound incidence from a specific direction can be simulated by convolving a mono source $s(t)$ with a pair of HRIRs:

$$\begin{aligned} p_{left}(t) &= s(t) * \mathrm{HRIR}_{left}(t) \\ p_{right}(t) &= s(t) * \mathrm{HRIR}_{right}(t) \end{aligned} \tag{3.26}$$

Binaural audio techniques are rendered over headphones. Although more accessible and more cost-effective than loudspeaker arrays for users, there are some inherent issues for headphone systems that should be taken into account. Firstly, wearing comfort is an important factor for the quality of hearing experience. It might be uncomfortable/unnatural for some listeners with sounds played via headphones. It has been reported that wearing headphones for too long may cause health problems, such as transient sensorineural hearing loss [96]. Secondly, the localisation might be perceived as 'in-head' when non-personalised HRTFs are used. This is because the HRTFs data measured from a dummy head may not suited well for some individuals, which degrades the perceived authenticity. 'In-head' problems can be partly solved by using individualised HRTF databases [7], although this is difficult to implement as it is usually time-consuming and difficult to obtain individualised HRTFs via on-site measurement for a group of subjects. Thirdly, commercial headphones are often equalised with respect to a reference sound field, which could be a free-field, a diffused field, or an artificial sound field defined by the manufacturer specifically for a given headphone model [97]. As there is not a standardised reference sound field for headphone equalisation, the perceived quality of sound may vary more or less depending on the brand of headphone or equalisation used for a given binaural listening experience.

## 3.2 Applications of Auralisation

As the term auralisation was first proposed by Kleiner in 1993 in the context of room acoustics [46], there is no doubt that auralisation has been widely used in this field, particularly for rooms where the acoustics are of high interest, e.g. concert halls and opera houses. With the rapid development of relevant modelling techniques and computing power in recent years, the application of auralisation is extending towards other fields, such as environmental sound outdoors, and real-time applications associated with Virtual Reality (VR) and Augmented Reality (AR) systems for creating interactive and immersive virtual environments.

### 3.2.1 Auralisation of Room Acoustics

For auralisation of room acoustics, sound sources are typically set as ‘dry’ music and/or speech signals recorded in an anechoic chamber. Emphasis is put on capturing or modelling apparent spatial room impulse responses, and convolving them with these ‘dry’ signals to produce an authentic audible experience of being virtually inside the auralised room. GAMs are often implemented to numerically simulate these spatial impulse responses. With commercial software based on such methods, e.g. ODEON, CATT-Acoustics, and EASE, etc., it is convenient to build up a model for complicated geometries. One of the typical applications of such a model in practice is to investigate and demonstrate the acoustic implications caused by proposed renovations and modifications to an architectural design [98], and/or to examine and optimise the effect of *public address (PA)* systems with respect to different loudspeaker positions [99] within the space. In the context of archaeology, auralisation of room acoustics can be used to represent, understand and experience past environments, showing how sound contributed to events and activities within historic spaces, which may reveal some aspects of human culture with audible cues [85, 100].

Auralisation of room acoustics is still an active research area, with consistent emergence of more efficient methods for different applications. Some state of the art developments pertaining to room acoustics auralisation can be categorised as follows:

- Application of wave-based methods combined with geometrical acoustics methods: Some recently published hybrid methods that combine wave-based methods for low-frequencies and geometrical acoustics for mid- and high-frequency sound simulation can be found in [4, 101, 102]. To extend the frequency range of wave-based methods, the technique to extrapolate spatial impulse responses to cover higher frequencies proposed by Southern et al. [103] can be considered.

- Source directivity patterns incorporating wave-based methods: Some frameworks that incorporate source directivity patterns for different wave-based methods for auralisation have been published recently, including directivity encoding schemes for BEM [104], FDTD [105, 106], and PSTD [107], etc.

- Perceptive and objective evaluation of auralisation models: Brinkmann et al.[108] did round robin tests to evaluate the state of the art of room acoustic modelling software both in the physical and perceptual realms. They found that most simulations using the GAMs generate obvious errors once the assumptions of geometrical acoustics are no longer met, and the difference between simulated and measured impulse responses for the same scene was clearly audible. A potential solution to this issue is calibrating the GAMs according to the on-site measurements at different points, from which the simulated binaural room impulse responses (BRIRs) were reported to be perceptually equivalent to the measured BRIRs [109].

### 3.2.2 Auralisation of Road Traffic Noise

Compared to its wide range of applications in room acoustics, auralisation for outdoor environmental sound is relatively less common. There are several reasons for this. Firstly, the range of sound sources for outdoor auralisation is richer than that of indoors which mainly consists of speech and music. It can be more difficult or even impossible to get anechoic recordings for some outdoor sounds, e.g. a bird chirping, or noise emitted by a flying aeroplane. Secondly, outdoor sound propagation effects are potentially more complex than those found indoors due to a series of extra considerations such as air turbulence by wind, meteorological effects, and ground effects for different types of terrains, etc. [75]. Thirdly, it can be more difficult to get outdoor recordings with a sufficient signal-to-noise ratio (SNR) for post-processing because of wind noise and/or some uncontrollable situations in an outdoor environment (e.g. it is difficult to obtain the impulse response of a street with vehicles passing by and many people present during the day-time).

On the other hand, the demand for more powerful tools that can be used for outdoor sound auralisation is increasing, especially in relation to urban *soundscapes* which refer to the holistic acoustic environment as perceived by humans in situ. The auralisation of urban soundscapes not only makes it possible to support urban planning and environmental noise assessment, but also might better enable the public's engagement on topics relating to a better understanding of human-environment relationships. Conventionally, environmental sound is usually evaluated quantitatively via objective metrics such as SPLs. Although these metrics can be used to describe the time-varying and/or time-averaged characteristics of sound levels with different resolution in the time-frequency domain, and can be visualised as noise maps, bringing a more direct impression, it is difficult to translate these numerical descriptions of a given environment in sound levels, or colours on a noise map, into an audible experience. The accurate evaluation of comfort or annoyance levels regarding the content of the sound is even more challenging. To tackle this issue, studies on the auralisation of environmental sounds and/or the overall acoustic environment of an outdoor space can be considered. This includes the auralisation of jet aircraft [110], railway noise [111], urban streets [112, 113], and forests [114, 115], etc. One of the most demanding auralisation targets related to environmental sound is traffic noise, as road traffic is a common noise source in our daily lives, a potential threat to public health, and although managed by mitigation to reduce overall SPLs, if at all, there is insufficient consideration on related correlation with perceived annoyance [116].

Road traffic noise is the combination of noise generated by a flow of one or more vehicles. The noise levels become high, having complex patterns of variation, when there is a large volume of traffic and different driving conditions. So far, there are no standardised methods for traffic noise auralisation, and the resultant modelled scenes may vary significantly, depending on the methods and algorithms utilised. Some specific considerations should be taken pertaining to moving sources representing vehicle pass-bys, and the relevant sound propagation in an outdoor environment which may cause specific and noticeable audible effects. As it is difficult to get anechoic recordings of a vehicle pass-by sound, synthetic methods should be taken into account for source models.

### Sound source modelling

Some of the recently published synthetic source models for traffic noise auralisation can be categorised as micro-, meso-, and macro-scopic models according to the level of detail required to auralise traffic flow scenes.

#### 1) Micro-scopic models

Micro-scopic models focus on the sound emitted from each vehicle from a low-level perspective. Flow noise is simulated by calculating the emission from each vehicle for small time blocks, and these sounds are summed together according to the distribution of vehicles over a specific period of time. The low-level perspective here means that, for each vehicle, the rolling noise and the propulsion noise are calculated separately, and each noise is determined by multiple parameters such as gear setting, engine order, engine speed, tyre structure, and the mechanical impedance of road surface, etc. An advantage of Micro-scopic models is that it is possible to allow full control of the signal characteristics. However, it is usually very computationally demanding and time-consuming to run a micro-scope model, limiting its suitability for real-time applications. Moreover, it may be troublesome to get access to the full range of parameters required to run a micro-scope model.

For example, Maillard and Jagla [112] proposed a method of synthesising engine sound and tyre noise using a granular synthesis technique. Grains are extracted from the analysis of a recorded engine sound signal corresponding to continuously varying engine speeds, and tyre noise recordings are obtained using the Close-Proximity Method (CPX), a standard measurement method for noise levels emitted from a passenger car tyre when rolling over a road surface. Each grain is assigned with a vehicle speed or engine speed as the first control parameter, in addition to engine load as a second control parameter. This method is computationally effective for real-time usage of engine sound and tyre/road noise synthesis, but its flexibility in terms of synthesising different types of vehicle sound is relatively limited as it is difficult or impossible to take recordings of all the vehicles required.

Another example of such a micro-scopic model for engine sound synthesis is proposed by Piernen et al [117] in which the engine sound is synthesised by a spectral modelling synthesis (SMS) approach. The time-varying characteristics of the tonal signal and the stochastic signal are assigned with control parameters including engine speed, engine load, engine order, gear setting, and emission angle, etc. They took a series of in-situ measurements for a specific vehicle with different engine conditions in terms of engine speed, engine load, and emission angle, and used the spectral information in these recordings to extract about 180 input parameters to generate the engine sound using the SMS model. Although the SMS model gives full control of the signal characteristics, it is based on recordings of a specific vehicle model, with limited flexibility when synthesising sounds emitted from different engines.

Hoffman [118] proposed a micro-scopic model for tyre rolling sound synthesis based on the physical behaviour of the tyre. In this model, tyre sounds are synthesised by an established model (SPERoN) which is a combination of a physical model and a statistical model. The physical model calculates the contact forces of the tyre-road contact according to the design structure and material of a tyre, while the statistical model is established by implementing multivariate linear regression analysis to a series of pass-by recordings, in order to predict the spectrum of tyre noise in respect of airflow-related mechanisms, tyre friction, and tyre cavity, etc. Although it is claimed that the perception of simulated and recorded signals correlate well when compared in listening tests, it is difficult to get all the required data for running this model because most are unpublished or undocumented, such as tyre vibration patterns, airflow-related mechanisms, tyre friction, and tyre cavity patterns, etc. Moreover, as the main purpose of this model is to support tyre design instead of auralisation, it is computationally demanding and time-consuming to run such a model based on the physical behaviour of the tyres, limiting its application in a wide range of auralisation contexts.

**2) Meso-scopic models**

Meso-scopic models focus on the sound emitted from each vehicle from a high-level perspective. The status of each single vehicle is determined by some empirical equations rather than a large number of low-level parameters. These empirical equations are derived from large measurement datasets and statistical analysis, and make use of high level parameters (e.g. vehicle category, vehicle speed) to describe the status of each single vehicle. Hence, the output sounds generated by Meso-scopic models do not exactly correspond to a specific vehicle sound, but provide an approximation in sound power levels and perceived plausibility. Some factors that influence the source timbre, such as acceleration/deceleration, driving patterns, road type, and road surface conditions, can be partly compensated or corrected by involving additional empirical equations. Compared with Micro-scopic models, Meso-scopic models have limited flexibility, relying on the effectiveness of derived empirical equations and measurement datasets. However, these models are less computationally demanding, and are less time-consuming to generate, which opens up the possibility of real-time applications with fewer parameters required.

One typical example is the Harmonoise model [119] for sound source modelling. In the Harmonoise model, the SPLs of tyre rolling sound are given as functions of frequency, vehicle speed, and vehicle category. There are five main categories of vehicles defined in the Harmonoise tyre noise model (light vehicles, medium-sized vehicles, heavy vehicles, other heavy vehicles, and two wheelers). An empirical equation representing the near field SPL of tyre/road noise for each vehicle category with regression parameters is presented as:

$$L_t(f) = a_t(f) + b_t(f)log(\frac{v}{v_{ref}}) \tag{3.27}$$

in which $v_{ref}$ is the reference speed set as 70km/h, and $v$ is the vehicle speed considered as a constant value, validated within the range of 30–130km/h. The regression parameters $a_t$ and $b_t$ are given in one-third octave bands for the frequency range 25Hz to 10kHz. The energy of the total tyre noise is considered to be allocated to two point sources for light vehicles, which are located at 0.01m and 0.3m/0.75m above the ground, respectively. Some other influencing factors, such as the road surface material and wetness, the road age, the source directivity patterns, and the acceleration/deceleration status, etc., are treated as additional corrections ($\Delta L_i$, i=1,2,3...) to the reference condition. Although the Harmonoise tyre noise model was originally developed for predicting the spectra of long-term time-averaged sound pressure levels, the idea of this engineering model (using an empirical equation with regression parameters) has been implemented in some relevant auralisation studies for tyre noise synthesis for pass-by vehicles, such as [117, 120]. It is mainly used because the timbre of tyre rolling noise can be generally considered as broadband noise, if the minor tonal part caused by tyre resonances can be neglected. Therefore, it is feasible and convenient to create an audible sense of tyre noise for vehicle pass-bys based on empirical equations. The regression parameters $a_t$ and $b_t$ and the additional corrections ($\Delta L_i$, i=1,2,3...) can be achieved or selected by different methods, e.g. derived from different sets of in-situ recordings corresponding to different pass-by conditions, which makes the model quite flexible.

### 3) Macro-scopic models

Macro-scopic models do not focus on the sound emitted from single vehicles. Instead, a traffic lane is treated as a linear source without considering the specific conditions of each vehicle. The sound signal of the linear source is simulated by filtered broadband noise with *modulation transfer functions (MTFs)* [121] to generate rippled spectra, so as to create the perception of fluctuation due to each vehicle pass-by. Compared to micro- and meso-scopic models, macro-scopic models can dramatically save computational power and time at the sacrifice of plausibility and flexibility. Therefore, these models are mainly used for background traffic noise, e.g. road traffic far away from a listener. For example, in [122], an approach that considers the traffic as accumulated noise is used and individual pass-by events, e.g. Doppler effect, are not modelled explicitly for each vehicle. A combination of MTFs and pitch shifting algorithms as used to simulate fluctuations due to traffic flow inhomogeneities. This auralisation has been validated by testing against mixed outputs from a previously developed and validated demonstrator.

## Sound propagation modelling

In terms of sound propagation models for traffic noise auralisation, it is often more difficult to take measurements of impulse responses in an outdoor environment. Therefore, numerical methods are usually used to simulate outdoor impulse responses representing the acoustic characteristics inherent to the environment. Most of the existing numerical methods used in room acoustics can be applied with modification for the characteristics of the outdoor environment. GAMs or wave-based models are used in some studies to investigate outdoor sound propagation in an urban environment, such as image models for urban squares [123] and street canyons [124], ray-tracing models for urban squares [125], PSTD models for urban streets [126, 127], and a digital waveguide mesh (DWM) model for a forest [128], etc.

Apart from the numerical simulation of impulse responses, there are also some engineering models that use empirical equations derived from statistical analysis of recorded datasets for predicting sound levels in an outdoor environment. These engineering models focus on the variation of sound levels in terms of different outdoor environments from a point-to-point propagation perspective, and establish a series of empirical equations to approximately predict a variety of sound propagation effects, such as reflection, scattering, ground effects, vegetation influences, meteorological effects, etc. Some typical examples of these engineering models include ISO 9613-2 [129], NORD2000 [130], and HARMONOISE [131], etc. Although these models are for the approximation of sound propagation effects under different scenes, and in theory less physically accurate than numerical simulations of the corresponding environment impulse responses, there are fewer computational resources required for the calculation of the empirical equations than the implementation of a numerical simulation. Therefore, these engineering models are specifically suitable for predicting large-scale outdoor sound propagation, e.g. generating a noise map for a city.

### 3.2.3 Auralisation for Virtual Reality and Augmented Reality

Virtual reality (VR) refers to a computer-generated simulation in which a person can interact within an artificial three-dimensional environment using electronic devices, such as special goggles with a screen or gloves fitted with sensors [132]. In this simulated virtual environment, the user is able to have a realistic-feeling experience. Augmented reality (AR) is based on the perception of a real-world environment, which can be defined as a real-time direct or indirect view of a physical real world environment that has been enhanced/augmented by adding virtual computer generated information to it [133]. A combination of VR/AR is called *mixed reality (MR)* which can used to enhance the digital interaction experience of the virtual/real world [134]. Currently, the simulated visual experience of being physically present in a VR environment can be commonly achieved via VR headsets by rendering high resolution, 360 degree stereoscopic visual scenes. Auralisation can be used as a powerful tool to create audible cues to enhance the sensation of 'presence' in a virtual environment and the plausibility of the virtual objects created within it. Two basic requirements for rendering the auditory environment for a VR/AR system are *dynamic* and *immersive* presentation. In other words, the system should facilitate user interaction in real-time without notable artificial auditory effects. However, this is not an easy task because of the limitation on computation power and the availability of data to render complex audiovisual scenes.

In order to achieve dynamic auditory rendering, the signal delivered to the VR/AR audio system should be processed 'in near real-time'. This will correspond to the acceptable limits for latency and update rate for the system. The higher the update rate and the lower the limit for latency is, the smoother and more continuous the perception of the auditory scene rendering will be. Nevertheless, a high update rate means that very little time can be allocated for acoustic simulations of sound sources, propagation effects, and spatial audio rendering, limiting the capability of rendering complex acoustic scenes. Therefore, the acceptable limits for latency and update rates are often contradictory to each other in practice, with a compromise between the complexity of the auralisation models and the sensation of the dynamics of the scene. Information from psychoacoustics are widely used for reference when setting specific limits of latency and update rates. For example, an update rate of 60Hz and total delays of 50ms are considered acceptable for rendering moving sound sources in a virtual environment [7]. Here total delay includes the acoustic simulation and signal processing time for running the auralisation model, in addition to the latency caused by hardware communication such as the VR headset, the sound card, etc.

Immersion for auditory environment rendering is also a complicated problem. Theoretically, acoustic rendering of faithful physical models for an environment should correspond flawlessly to a real-word scenario, and the listeners can therefore respond naturally to what is presented to them. However, it is often difficult or impossible to render a complete physically-based auralisation framework in real-time because of the limitations on computation time and dynamic rendering, as well as the availability of resources for setting up these physically-based models with sufficient data and storage space. Therefore, in practice, the audio capabilities of VR systems aim to create a perceptually plausible auditory experience, rather than one that is considered as being physically accurate [135]. This aim also corresponds well with AR systems as there is no absolutely physically accurate acoustic properties for a virtual object which does not exist in real-world. Unfortunately, there is still a lack of study on how to evaluate the plausibility of an auralised auditory scene in VR/AR, or to judge whether an auralised scene is acceptable/unacceptable. Knowledge from psychoacoustics can be used for reference to determine what information might be included for the inputs and outputs, and what can be fairly discarded, but this can provide only general ideas rather than specific guidelines for implementation of an auralisation framework to a specific acoustic problem. In fact, one of the greatest difficulties for acousticians and audio engineers is to establish a hard line between the subjective evaluation and the objective physical characteristics of a sound or sound environment. Generally, the evaluation of auralisation plausibility can be divided into two aspects, which are the perceived accuracy of spatial cues, and the perceived realism of timbre and sound quality.

In normal-hearing individuals, the vast majority of spatial information is derived from ITD, ILD, and the spectral differences between the acoustic signals that arrive at each ear, as discussed in Chapter 2. As sound localisation on the vertical plane is much weaker compared with that on the horizontal plane, some spatial information in the vertical plane can be ignored to simplify the spatial audio rendering process for the applications where the motion of the sound emitter in the vertical plane is not of high interest. In fact, for some HRTF database measurement, the spatial resolution is not evenly distributed across the whole azimuth and elevation range. For some directions where the human auditory system is not as sensitive to directional cues, the measurement points are set more sparsely to simplify the measurement process and the database [92]. When rendering a sound from a direction which is not included in the HRTF database, a series of HRTF interpolation techniques [136–138] can be used to simulate the impulse response pairs of the target direction, or the information from neighbouring directions may be used directly as an approximation of the target direction. It is still an open question on how to evaluate the perceived difference and the degree of acceptance between such different approximation methods.

The perceived realism of timbre and sound quality is dependent on complicated factors, and there are two main difficulties for quantitative evaluation. Firstly, the perceived realism of timbre and sound quality may be related to not only auditory perception, but also other perception aspects, e.g. visual, haptic, etc. For example, Viollon et al. [139] found that when people hear natural sounds (e.g. singing birds) with associated visual scenes (e.g. woods) presented, the sounds will be rated as more pleasant than the same sound presented without visual scenes. Maffei et al. [140] showed that the perceived loudness and annoyance level for transparent noise barriers is lower than that for opaque barriers. They also found that the impact of the perception of a sound event may be influenced by changing the colour of a visual component, with no associated change in any related acoustic property [141]. These findings reveal that human perception is multisensory by its nature and that the environment is perceived and represented holistically [142]. Therefore, a holistic approach from different perspectives is required for standardising the evaluation of perceived quality of sound. Secondly, there is a lack of objective metrics to describe the perceived quality of sound. The most commonly used method for perceived quality of sound is using rating scales with a series of adjective pairs assigned to the characteristics of the sound, e.g. pleasant–annoying, bright–dull, natural–artificial, etc. This is called the *semantic differential (SD)* method which has been used in several acoustic studies [17, 143, 144]. However, there are no standard methods for selecting the adjective pairs, and the descriptive words may vary for different auralised sounds or a sound auralised by different methods, making it hard to compare the quality of sound auralised by different auralisation models. Moreover, these terms may have different meanings for different people based on their own experience and cultural backgrounds. Similar to the concept of product sound quality introduced in Chapter 2, an alternative to the SD method is based on objective metric calculation to describe characteristics of sound plausibility. However, there have been no such standardised metrics so far. Thus, it is still worthwhile to investigate the relationship between different objective metrics and subjective responses in terms of plausibility.

In summary, auralisation can be used as a powerful tool for VR/AR systems by creating auditory cues that enhance the user experience. When designing and implementing an auralisation framework, specific considerations should be taken on the balance between the physical accuracy, perceptual plausibility, and available computational resources, according to the application of the auralisation framework. There is also still a lack of guidance on how to evaluate the plausibility of such auralised sounds. While subjective evaluation can be used to provide useful information, it is also worthwhile to explore objective metrics in a wider context.

## 3.3 Summary

This chapter has covered the fundamentals of auralisation, including the background theory of treating airborne sound auralisation problems as LTI systems, and the three basic elements within an auralisation framework consisting of sound source modelling, sound propagation modelling, and spatial audio reproduction, in addition to some considerations on applications of auralisation in room acoustics, road traffic noise, and with VR/AR systems. These content in the relevant fields are conducive to develop a holistic understanding of what can be achieved with auralisation and what are involved within its concept.

After an overview of auralisation techniques applied in room acoustics and outdoor environments, it is found that auralisation is less commonly used in outdoor environments, although there is increasing demands for such tools for outdoor scenes. As discussed in Chapter 1, the most direct motivation behind this study is to develop and apply auralisation techniques for outdoor sound environments. Traffic flow noise, which is a common sound source in an urban environment, has been chosen as the target sound scene to be auralised to explore the potential solutions for the auralisation of outdoor sounds. As discussed in this chapter, it is found that there have been no standardised methods for traffic noise auralisation, and the resultant outcomes may vary significantly when different auralisation frameworks are used. After an overview of some of the recently published synthetic source models for traffic noise auralisation, these models are categorised into micro-, meso-, and macro-scopic models. After an overview of the most used sound propagation modelling methods, it is found that apart from the GAMs and wave-based models, there have been some engineering models that are worth considering to approximate outdoor sound propagation effects. These engineering models are less computational demanding compared with GAMs or wave-based methods. Although auralisation can be used as a powerful tool for VR/AR systems, after an overview of its current application, it is found that there is still a lack of guidance on how to evaluate the plausibility of the associated auralised sounds. Some subjective evaluation methods can be taken for reference when implementing auralisation in VR/AR systems.

The next chapter will focus on the the concepts and techniques of procedural audio, which can be used for sound source modelling in an auralisation framework for the cases in which it is difficult or impossible to take anechoic recordings. These techniques will be helpful for the development of the traffic flow auralisation framework proposed in Chapter 6, where the sound propagation model and binaural sound reproduction techniques discussed in this chapter will also be implemented.

# Chapter 4

# Procedural Audio

Having established the theoretical background of auralisation in Chapter 3 covering the three main aspects including sound source modelling, sound propagation modelling, and sound reproduction techniques, this chapter will now focus on some specific sound source modelling cases in which it is difficult or impossible to obtain anechoic recordings for auralisation purposes. In fact, this is a common situation encountered for auralisation of environmental sounds, such as the wind blowing, birds chirping, and fire crackling, etc. Under these circumstances, it is necessary to find solutions to acquire 'dry' source signals to be fed into the sound propagation and reproduction blocks, according to the specific requirement of the auralisation scenes. Procedural audio, which can be concisely understood as programmatic and data-driven synthetic sound signals created in real-time, provides potential towards this goal. However, there will always be compromises between the realism, flexibility, and computational resources when implementing procedural audio methods. It is crucial to get a balance between these contradictory features to tailor reasonable and feasible solutions for a specific auralisation scene. This chapter will first give an introduction to the concepts and definition of procedural audio, together with the characteristics of this technique and the main advantages and disadvantages compared with sample-based audio in which recordings are always needed as the input signals. Then an overview of some sound synthesis techniques following the idea of procedural audio will be presented, followed by a brief introduction to some applications of procedural audio in video games and soundscape studies to demonstrate what can be achieved and where the limitations are for the widespread application of procedural audio.

## 4.1 Introduction

### 4.1.1 Definition of Procedural Audio

So far, there has been no formal definition of the term *procedural audio*, although the idea of *procedural* has been widely utilised in the audio engineering field, 'relating to or comprising memory or knowledge concerned with how to manipulate symbols, concepts, and rules to accomplish a task or solve a problem' [145]. Regarding auralisation, the 'problem' here can be considered as producing audible sounds that fit one or more constraints including plausibility, flexibility, and real-time performance, etc. One of the earliest definitions of this term is proposed by Farnell [146], stated as:

> *"Procedural audio is sound qua process, as opposed to sound qua product. Behind this statement lies a veritable adventure into semiotics, mathematics, computer science, signal processing and music. Let us reformulate our definition in verbose but more precise modern terms before moving on. Procedural audio is non-linear, often synthetic sound, created in real time according to a set of programmatic rules and live input."*

The context above might be somehow abstract and confusing at first. Here, in order to clearly understand what the term means, it is better to first recognise what kind of audio techniques do not correspond to this concept.

Playback recordings of a sound source directly as part of an audible scene or sound environment does not correspond to the idea of procedural audio.

In fact, when considering auralisation for the use of rooms acoustics, anechoic recordings of music or speech are very commonly as sound sources because they are the typical sound sources heard in rooms in everyday life. With appropriate use of recording techniques, it is possible to capture all the required features of a sound source in the recorded files, which leads to a high realism when played back. Each time a recording is used, the same audio content in the same order (from start to end) at a fixed rate (identical to the sampling rate) will be replayed and/or fed into the steps that follow in terms of sound propagation and spatial audio reproduction for auralisation. Hence, these recordings are treated as assets, or in other words, 'sound qua product' instead of 'sound qua process' according to the definition above.

Sounds produced by Musical Instrument Digital Interface (MIDI) [147] can be considered as procedural audio under some circumstances based on the nature of how the sounds are actually generated. With a MIDI sequencer, no actual recordings are transmitted. Instead, information pertaining to the notation, pitch, velocity, and vibrato, etc., is used. A database of recorded short clips of samples of the target sound source is indispensable which can be then rearranged at playback in a specific order according to the information stored in the music sequencer. If the playback order is fixed, for example, a melody written by a human composer in which the timbre, the velocity, and the loudness of the notes have been determined, it is not an example of procedural audio. If the playback order is variable according to some descriptive rules written in code, e.g. a melody written by algorithmic mathematical method running in real-time, this can be considered as procedural audio because it is driven by a set of programmatic rules with live input to produce sounds. The output sounds then change according to the variation of the driving parameters. This is similar for some other sample-based audio techniques using short sound clips, e.g. granular synthesis [148], in which recorded samples are split into small pieces lasting typically 10–50ms called 'grains'; multiple grains may be layered on top of each other, and may play at different speeds, phases, volume, and frequency, among other parameters. If the grains are played in a fixed order, it is, again, not considered as a procedural audio method. If the grains can be manipulated individually and rearranged with different orders, loops, and layers, etc., according to a series of algorithms for specific applications, this does follow the concept of procedural audio, but with recordings involved. In fact, in some research these sample-based audio techniques are referred to as *dynamic sound* techniques which are strictly distinguished from procedural audio because recordings are always required to run such models [149]. This thesis will follow this concept and will clearly distinguish dynamic sound techniques with procedural audio. In other words, any recording-based method, such as granular synthesis, is not considered as procedural audio in this thesis.

### 4.1.2 Characteristics of Procedural Audio

The main characteristics of procedural audio have been summarised in its definition according to Farnell [13], including the terms *non-linear*, *synthetic*, *real-time*, and *programmatic* and *live input*, etc. Here *non-linear* means the sequenced sound can be played in any order, e.g. it can jump between samples, move at different rates or in different directions, rather than in a fixed order from the beginning to the end at a fixed playback rate. It is therefore possible to change the order, pitch, and/or timbre of the output sound by tuning the algorithm. This non-linearity brings the possibility for procedural audio to adapt to a variety of application scenarios in which interactive and flexible sounds are required according to the decisions, actions, and responses of the listener.

The interactivity and flexibility feature is further enhanced by utilising *synthetic* sounds with dynamic waveforms, spectrum, and amplitude characteristics, either produced by ‘true synthesis’ (created entirely from nothing but time dependent equations and input parameters) or ‘analysis and synthesis' (created according to equations and data extracted from usually spectral analysis of relevant recordings). Although a synthetic sound might not be perceived being as realistic as its recorded counterpart, due to it not always being possible to find a synthesis method suitable for imitating the full characteristics of a target sound source, more freedom and a greater range of potential sounds can be obtained by procedural audio methods. These sounds can then be implemented and integrated into the wider context of an auralised scene or environment.

*Real-time* refers to the acceptable limits of latency between the actual start of the rendered audio material and hearing the resultant sound. The limits may vary according to a listener's expectation in terms of application, from several milliseconds in the case of a single sinusoidal sound synthesis to a couple of seconds for the case of algorithmic composition [13]. It is worth noting that some algorithms are designed to render audio in an offline mode, which means the audio system takes a relatively long time to process and produce a sound with a significant audible latency. Although these sounds have some of the features described above such as being data-driven and flexibile, and may be used for auralisation purposes in some cases, they do not strictly belong to the range of procedural audio techniques because of their non-real-time performance.

*Programmatic rules* (algorithms in code) and *live inputs* (driving parameters) are vital for procedural audio methods. Algorithms should be designed according to a simplification of the specific properties and behaviours of the target sound source. Some of the most widely used algorithm design methods for procedural audio will be introduced in Section 4.2. Live inputs can be categorised by different level of details. High-level input parameters are often designed to be manipulated and decided by the listener so as to enhance the experience of interactivity, while low-level input data can be extracted from the relevant algorithms to make the system more autonomous. For instance, when generating the noise of a fan via procedural audio, the high-level input parameter might be set as the speed of the motor blades which can then be adjusted by the user. Other inputs, such as the volume level of fan noise, and the filter coefficients used to process the audio signals to get different spectra corresponding to different rotation speeds, should be treated as low-level parameters which are integrated into the algorithm.

### 4.1.3 Comparison with Sample-based Audio

In sample-based audio (also called recording-based audio in some research) techniques, actual recordings are indispensable as key assets for rendering sound sources. From the perspective of auralisation, if the acoustic scene is complex where multiple sound sources are included, recordings of each single source would be required, which is usually managed by a large database consisting of audio files. In contrast, the core assets for procedural audio are the algorithms and code which can be executed to render the resulting sounds in real-time.

Compared to sample-based audio, the requirement of the computation memory and storage space for procedural audio can be significantly reduced when rendering complex sound scenes, while the CPU workload may increase according to the complexity of the algorithms used. This is because a specific audio file needs a fixed storage space on the disk and has a fixed cost of memory when it is loaded, no matter what the sound represents and how complex the content is. For procedural audio, in contrast, it is algorithms consisting of several lines of codes that are stored and loaded in the memory instead of specific audio files, and a series of sounds can be obtained by running the algorithms with different input data. The more complex the algorithm is, the more computational resources are required. This is so called *variable cost* which is one of the advantages of procedural audio compared to sample-based audio techniques. With variable cost, it is possible to render complex sound scenes under resource-limited conditions, e.g. mobile platforms or wearable devices. In order to take the most advantage of variability, knowledge from psychoacoustics is often taken for reference when simplifying an algorithm. For example, peripheral sounds, or sound levels below the masking threshold as introduced in Chapter 2, can be reasonably removed as part of a series of simplifications to tune the algorithm to save computational cost. This is generally not feasible for sample-based audio which has a fixed cost in terms of memory and CPU workload when replayed.

Another unique advantage of procedural audio compared with sample-based audio is its variety and flexibility. Here variety and flexibility imply the potential for sound variation in terms of its timbre, frequency, and level, etc., so as to adapt to different scenes in practical applications. Recordings are often played precisely the same way as recorded. Even though they can be lightly modified via some signal processing techniques such as resampling, pitch shifting, and filtering, the potential dynamic range and flexibility is much more limited than procedural audio methods. For example, the sound of flying bullets that might be heard in a game can be synthesised by procedural audio with a physical-based synthesis model [150]. It is possible to create a series of gunshot sounds which adapt to a variety of different shotgun models by tuning the descriptive parameters of the gunshot structure. To achieve a similar effect using sample-based audio techniques, a series of recordings of different gunshot models are required, which is difficult to implement and will occupy much more storage space than a piece of code in the physical-based synthesis procedural audio model.

Nevertheless, procedural audio is not a one-size-fits-all solution for sound source modelling problems. Indeed, there are areas where it may fail and can never replace recorded sound. Firstly, there are aesthetic issues with procedural audio. When using models and algorithms to create sounds, it is necessary to make appropriate pre-assumptions to simplify the target sound source or sound event for practical implementation, with considerations on the limitations of computational resources. As some details are neglected or not fully reproduced due to simplification, it is often considered that sounds created using procedural audio methods are not always plausible, or authentic enough, when compared to their counterpart recordings from the political perspective, even though these synthetic sounds might be highly realistic from the technical perspective [146]. Here the words 'realistic', 'authentic', and 'plausible' are descriptive adjectives that indicate the properties of a sound in a virtual scene such that people might consider the sound credibly associated with the real-world circumstance that gives rise to its occurrence. It is usually considered that there are fewer plausibility issues for playback recordings if all the properties of the sound can be inherently captured and replayed, when compared to synthetic sounds using procedural audio techniques. Furthermore, the metrics and thresholds for evaluating whether a synthetic sound is plausible are not well defined, and depend heavily on the content of the sound and its function in a sound environment. For example, peripheral sounds can be recognised with time-averaged statistics and so they might be still considered plausible although with the lack of details presented [151]; the noise masking effect may have a significant impact on the subjective evaluation of a sound scene, so a sound scene evaluated as highly plausible may be poorly rated in terms of plausibility when it is masked by a noise masker, and vice versa [152]. Knowledge from psychoacoustics can provide some cues for adjusting dynamic level of details to simplify the procedural audio algorithms to plausibly generate the target sounds.

Secondly, the advantage of variable cost may sometimes become a dilemma as it is not always possible to predict the required computational resources at early design stages. It is sometimes the case that the complexity of an algorithm is beyond the allocated resources. There are two general ideas to tackle this issue. One is to develop some less computational complex alternatives in advance. If the original algorithm has run out of allocated resources, the algorithm can switch to a simplified version, sacrificing plausibility to some extent. Another solution is to include a pre-computation stage, with computation processes with high complexity prepared at a ‘warm up’ stage before running the whole algorithm. This is a compromise of real-time performance of the whole sound scene.

## 4.2 Sound Synthesis Methods for Procedural Audio

As algorithms are the core assets for procedural audio models, this section will give an overview of some popular sound synthesis methods based on which specific algorithms can be developed as relevant to the context of the thesis. Each synthesis method has its pros and cons, and can be suitable for specific application scenarios. In order to implement a procedural audio model, it is necessary to choose appropriate sound synthesis methods and integrate them accordingly to develop a variety of algorithms for creating different sound sources capable of adapting to different acoustic scenes.

### 4.2.1 Additive Synthesis

Additive synthesis is a method that adds together, generally, sine waves or other sinusoidal components of different frequencies and levels to produce a final sound [153]. The theoretical background of additive synthesis is the Fourier Series, which says that in theory any periodic wave can be described by specifying the frequency and amplitude of a series of sine waves. These sine waves are called *partials* in an additive synthesis model. The amplitudes and frequencies of these partials are considered to vary continuously and slowly so that short-time Fourier analysis can be utilised to determine the amplitude and frequency envelopes. Mathematically, an additive synthesis model can be expressed as:

$$y(t) = \sum_{i=1}^{N} A_i(t) \text{sin}[\theta_i(t)] \tag{4.1}$$

$$\theta_i(t) = \int_0^t \omega_i(t)dt + \phi_i(0) \tag{4.2}$$

where

$A_i(t)$ is the amplitude of the $i^{\text{th}}$ partial over time $t$.

$\theta_i(t)$ is the phase of the $i^{\text{th}}$ partial over time $t$.

$\omega_i(t)$ is the radian frequency of the $i^{\text{th}}$ partial at time$t = t$.

$\phi_i(0)$ is the phase offset of the $i^{\text{th}}$ partial at time t = 0.

As can be seen from the mathematical definition above, each partial represents a component with one single frequency and phase, having a start and end time independent from others. For generating sounds with rich frequency content, a vast number of partials are required. The more partials involved, the more computational resources an additive synthesis algorithm requires. In practical applications, additive synthesis is often used for synthesising sounds with much related to harmonics, e.g. musical instruments, for which the perceived timbre is primarily dependent on the relationships between the frequency of the harmonics and their evolution over time.

### 4.2.2 Subtractive Synthesis

Subtractive synthesis is a method in which an audio signal is attenuated by filters to remove certain frequencies in order to obtain the desired timbre of the sound [153]. It can be seen as analogy to a sculpting process – starting with a huge hunk of rock, chiselling away gradually, revealing the sculpture expected. The fundamental frequency of the source is first tuned to the desired pitch, and the filter is designed to reproduce the spectral envelope of the target sound. By implementing time-varying low-pass, high-pass, band-pass, or notch filters with different cut-off frequencies and filter coefficients, numerous sounds and sound effects can be obtained. Mathematically, a subtractive synthesis algorithm can be expressed as:

$$y(t) = \int_0^t h(t, \tau)x(\tau)d\tau \tag{4.3}$$

Where $x(\tau)$ is the input audio signal, $h(t, \tau)$ is the impulse reponse of the filter at time $t$, and $y(t)$ is the output signal. In practical applications, subtractive synthesis is often used for synthesising musical instruments [154], or as a sub-process combined with other synthesis methods for generating non-music sounds.

### 4.2.3 Spectral Modelling Synthesis

Spectral modelling synthesis (SMS) is a combination of additive synthesis and subtractive synthesis. It models time-varying spectra as: 1) a collection of sinusoids controlled through time by piecewise linear amplitude and frequency envelopes (referred to as the *deterministic part*), and 2) a time-varying filtered noise component (referred to as the *stochastic part*) [155]. By combining the mathematical definition of additive synthesis in Equation 4.1 and subtractive synthesis in Equation 4.3 together, the output signal of an SMS algorithm can be expressed as:

$$y(t) = \sum_{i=1}^{N} A_i(t)sin[\theta_i(t)] + \int_0^t h(t, \tau)x(\tau)d\tau \tag{4.4}$$

Where representation of all the symbols is the same as in Equations 4.1–4.3. The first part of Equation 4.4 represents additive synthesis for rendering the tonal part of a sound, while the second part is responsible for the noisy part of a sound. White noise is usually used for $x(\tau)$.

Whether a synthetic sound produced by an SMS algorithm is plausible or not depends primarily on how the additive synthesis and subtractive synthesis algorithms are designed. Figure 4.1 shows the flow chart of an SMS algorithm. As can be seen from this figure, an input signal marked as ‘Original Sound’ is used. Therefore, SMS is an ‘analysis and synthesis’ method which generates sounds according to equations and data extracted from the analysis of relevant recordings, rather than a 'true synthesis' method. The original sound signal is fed and analysed by a short-time Fourier transform (STFT), and processed by some algorithms for sinusoidal peak detection, peak tracking, and spectral envelop modelling (e.g. the YIN algorithm [156]), so as to provide useful time-varying cues for the additive synthesis and subtractive synthesis in the steps that follow. In order for real-time implementation of the SMS algorithm, it is necessary to involve a pre-computation stage in which several recordings are analysed in advance to extract and store the useful information in a lookup table for running the additive synthesis and subtractive synthesis interactively in real-time after this pre-computation stage.

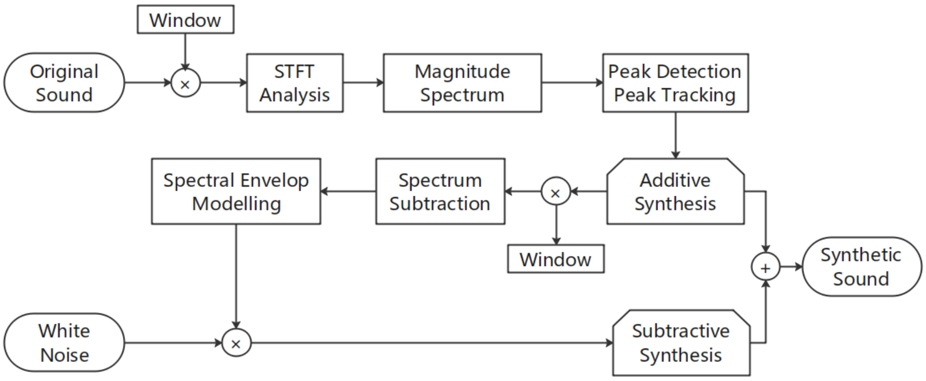


**Figure 4.1:** *The flow chart of an Spectral Modelling Synthesis (SMS) algorithm, reproduced from [12].*

### 4.2.4 Physical Modelling Synthesis

Physical modelling synthesis is a method that generates sounds by computing mathematical models derived from physical objects and phenomena [153]. This method can emulate the subtle variation of the object behaviour according to physical laws, leading to astonishing realism if the physical behaviour is manipulated to a sufficient level of detail. This is, however, very hard or infeasible for implementation under many circumstances as huge amounts of low-level physics, acoustics and mathematical parameters are needed to be defined. It can be difficult or not cost-effective to get all the required input parameters. Moreover, the computation process may take a very long time if the solver to the physical model is too sophisticated. Therefore, physical modelling synthesis is suitable for the cases in which physical laws are relatively simple and the equations for description of the physical phenomena can be established accordingly, while ‘cheap’ solutions to these equations can be achieved in terms of computational power and memory requirements.

The following context shows a typical example of an early physical modelling synthesis algorithm called the *digital waveguide* for synthesising plucked strings or struck strings [157], in order to illustrate how to develop a physical modelling synthesis algorithm from real physical phenomena. The theoretical background is wave propagation across a damped vibrating string.

Figure 4.2 shows a lossless vibrating string with stiffness $K$ and unit mass $\varepsilon$. Recalling the 1-D wave Equation 3.21 in Chapter 3, the physical behaviour of the string can be expressed as:

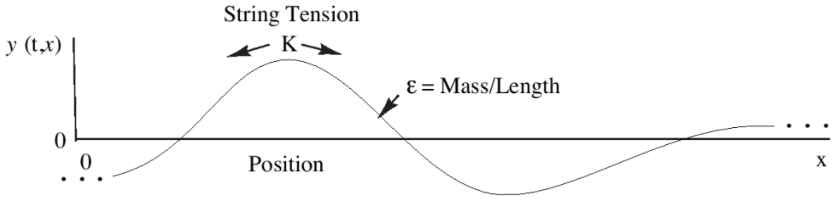


**Figure 4.2:** *Illustration of a vibrating string for physical modelling synthesis, reproduced from [12].*

$$\frac{\partial^2 y(t,x)}{\partial t^2} = c^2 \frac{\partial^2 y(t,x)}{\partial x^2} \tag{4.5}$$

where $c = \sqrt{K/\varepsilon}$ is the speed of wave transmission across the string. This equation can be solved using the travelling wave solution to the 1-D wave equation [12], defined as:

$$y(t,x) = y_r(t - x/c) + y_l(t + x/c) \tag{4.6}$$

Where $y_l$ and $y_r$ are arbitrary twice-differentiable functions denoting wave movement to the left and right, respectively. For discrete-time signals, this continuous travelling wave solution can be expressed as:

$$\begin{aligned} y[nT, mX] &= y_r[nT - mX/c] + y_l[nT + mX/c] \\ &= y_r[nT - mT] + y_l[nT + mT] \end{aligned} \tag{4.7}$$

Where $X$ is the spatial sampling interval, $x = mX$; $T$ is the time sampling interval, $t = nT$ such that $X = cT$. Equation 4.7 can be further simplified as:

$$y[n, m] = y^+[n - m] + y^-[n + m] \tag{4.8}$$

Where $y^+$ and $y^-$ are new notation for these travelling waves for convenience. This implies that the original wave function can be expressed by summing two parallel digital delay lines representing the left-going and right-going travelling wave components. Figure 4.3 shows an example of an ideal plucked string which has been sampled and decomposed into two travelling waves. Here 'ideal' means an initial string displacement and a zero initial velocity which fully determines the resulting motion in the absence of further excitation. The amplitude of each travelling wave delay line is half the amplitude of the initial string displacement and the sum of the upper and lower delay lines give the actual displacement value.

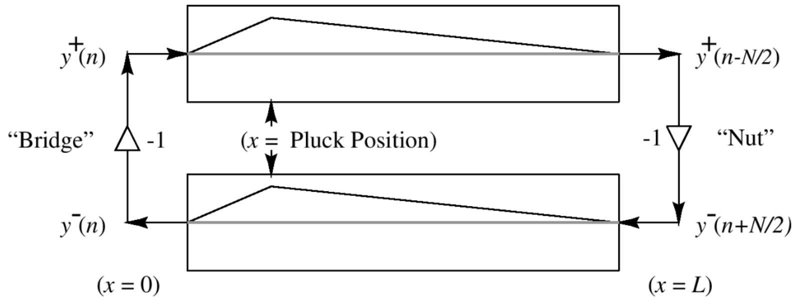


**Figure 4.3:** *Illustration of a plucked string decomposed of two travelling waves, reproduced from [12].*

While damping factors are considered, such as the frictional force which can be approximated as proportional to the particle velocity in the string and independent of frequency, the wave equation becomes:

$$K\frac{\partial^2 y(t,x)}{\partial x^2} = \varepsilon\frac{\partial^2 y(t,x)}{\partial t^2} + \mu\frac{\partial y(t,x)}{\partial t} \tag{4.9}$$

where $\mu$ is the first-order damping factor derived from experimental values. The travelling wave solution to this damped system becomes:

$$y(t,x) = e^{-(\mu/2\varepsilon)x}y_r(t - x/c) + e^{(\mu/2\varepsilon)x}y_l(t + x/c) \tag{4.10}$$

This solution can be discretised as:

$$y(n,m) = g^{-m}y^+[n-m] + g^m y^-[n+m] \tag{4.11}$$

Figure 4.4 demonstrates the digital simulation diagram of the damped plucked string. As can be seen from this figure, the loss factors $g$ for each sample have been pushed through to a single point (the 'Nut' end in this figure). By doing so, the output displacement at any position along the string will not correspond exactly to the sum of the upper and lower delay lines. In other words, the output becomes 'non-physical'. However, this actually has little change on the output value. The physical accuracy will not be compromised because the $N$ round-off errors per period arising from repeated multiplication by $g$ have been replaced by a single round-off error per period in the multiplication by $g^N$ [12].

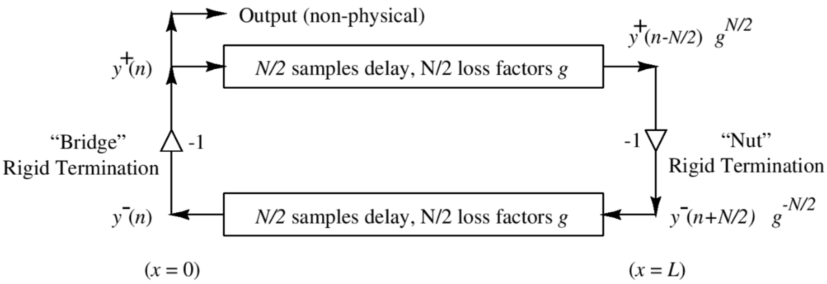


**Figure 4.4:** *Digital simulation diagram of a damped plucked string with rigid termination boundary condition, reproduced from [12].*

The two reflecting terminations (gain factors of -1) may be commuted so as to cancel them. Then the right-going delay may be combined with the left-going delay to give a single, length $N$, delay line, as shown in Figure 4.5.

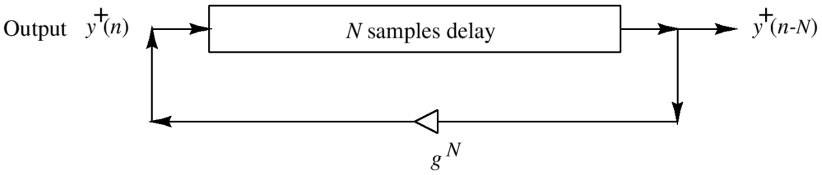


**Figure 4.5:** *Combining the right-going delay with the left-going delay in Figure 4.4, reproduced from [12].*

In real vibrating strings, damping typically increases with frequency. When considering the frequency-dependent damping factor, the wave equation becomes:

$$K\frac{\partial^2 y(t,x)}{\partial x^2} = \varepsilon\frac{\partial^2 y(t,x)}{\partial t^2} + \mu\frac{\partial y(t,x)}{\partial t} + \mu_3\frac{\partial^3 y(t,x)}{\partial t^3} \tag{4.12}$$

where $\mu$ is the first-order damping factor and $\mu_3$ is the third-order damping factor , both derived from experimental values.

The loss factors $g$ should be digital filters having gains which decrease with frequency. A simple first-order low-pass filter ican be used as an approximation to this ideal behaviour, for which the signal flow is shown in Figure 4.6 and corresponding transfer function is expressed as:

$$G(z) = 0.5 + 0.5z^{-1} \tag{4.13}$$

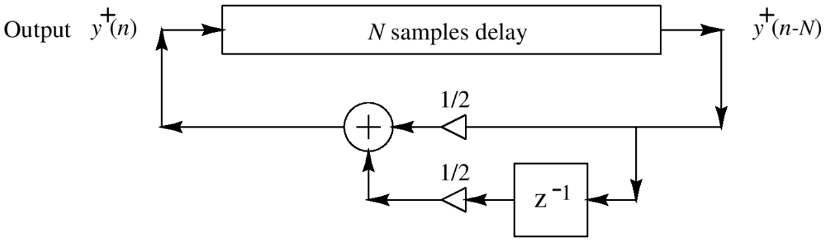


**Figure 4.6:** *Illustration of the digital waveguide algorithm including an N sample delay line and a first order low-pass filter, reproduced from [12].*

This implementation of the digital waveguide solution to the 1-D wave equation is also known as the *Karplus-Strong algorithm* [157] used for plucked string sound synthesis. Other FIR loss-filters under the stability constraint can also be utilised to make the algorithm more adaptive [158].

As can be seen from this example, there are a lot of low-level details required when establishing a physical modelling synthesis model. In some cases, the physical system can be sufficiently simplified without significant loss of timbre quality, so as for a computationally efficient implementation. The simplifications should be tailored according to the specific requirements and assumptions regarding different physical phenomena. For example, the mechanism of footsteps sounds can be simplified as the result of multiple micro-impact sounds between a shoe and the floor [159], water sounds can be considered as the sum of multiple submerged oscillating bubbles [160], and gunshot sounds can be simplified as a Friedlander wave plus a shock wave [150], etc.

### 4.2.5 Wavetable Synthesis

Wavetable synthesis is similar to additive synthesis, but uses more complex source waves than pure sine waves, such as square waves, sawtooth waves, or samples produced by real musical instruments taken from recordings. These complex waveforms are stored as ‘cells’ in a lookup table. The synth scrolls through the cells in the table one-by-one to output the sample as a sound. The continuous transition between two cells is realised by interpolation between their shapes or cross-fading [161]. Different pitches are created by speeding up or slowing down the table lookup rate.

Wavetable synthesis is convenient and cost-effective for practical use. However, it is not as flexible as physical modelling synthesis because the lookup table is indispensable. The timbre and nature of synthetic sounds created depends primarily on the content of cells. Under some circumstances the stored cells might not be useful for synthesising a target sound. Moreover, the memory requirement might be very high for a wavetable with a large amount of cells. In real applications, wavetable synthesis is often used together with other sound synthesis methods to produce natural sounds.

### 4.2.6 FM Synthesis

FM (Frequency Modulation) synthesis is a sound synthesis method that uses a signal called a *modulator* to modulate another signal called a *carrier*. The frequency content in the carrier is mostly related to the pitch of the target sound, while the modulator is used to alter the carrier at a specific modulation rate and modulation intensity. Figure 4.7 illustrates the concept of frequency modulation. The upper figure shows the high frequency carrier wave, the middle figure shows the low frequency modulating signal, and the last figure shows the resultant frequency modulated wave.

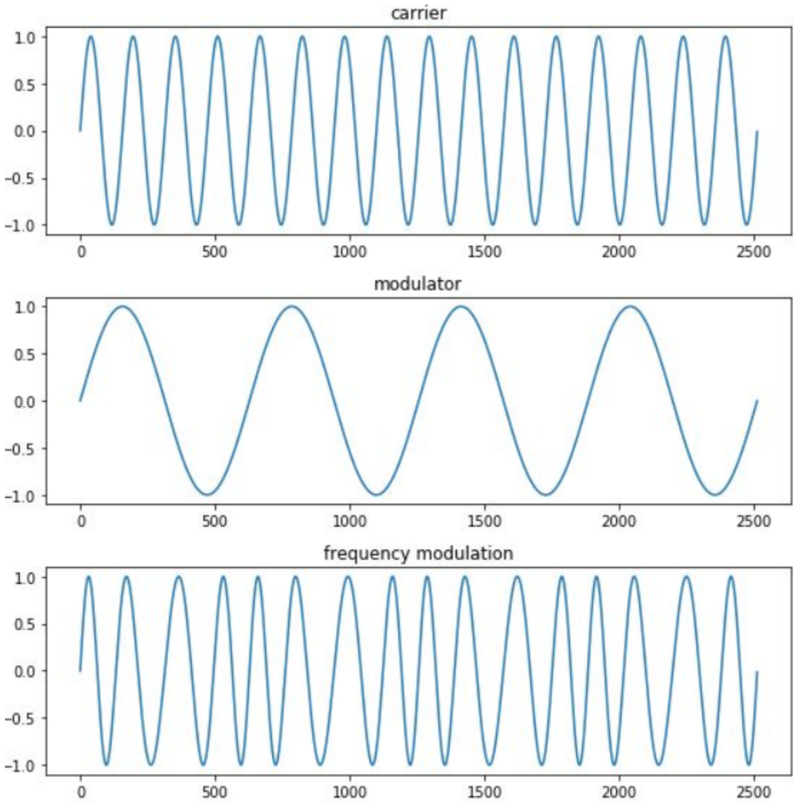


**Figure 4.7:** *Illustration of FM synthesis. The upper figure shows the high frequency carrier wave, the middle figure shows the low frequency modulating signal, and the bottom figure shows the resultant frequency modulated wave.*

As can be seen from the Figure 4.7, the frequency of the modulated wave varies according to the frequency of the modulator, and the frequency pattern of the carrier can be still be found in the modulated wave. If the carrier frequency falls into the audible range, this can lead to the perception of a 'vibrato' effect. When the frequency of the modulator increases, the 'vibrato' effect will disappear and the modulated sound can be perceived as a new timbre because of the change in its waveform, as shown in Figure 4.8.

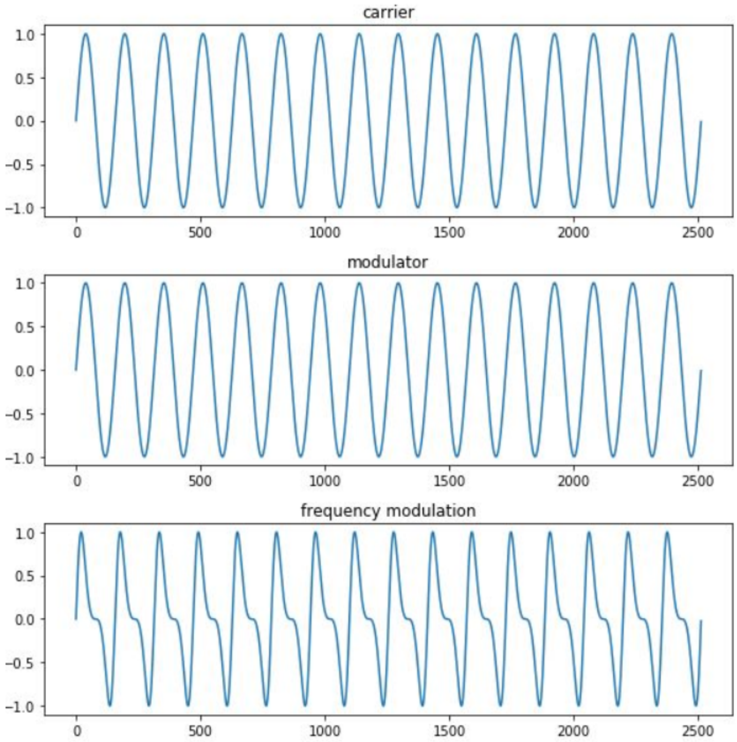


**Figure 4.8:** *Illustration of FM synthesis and the consequence of increased modulator frequency resulting in the synthesis of a new timbre different from either carrier or modulator.*

When sinusoids are use for mathematical description of the carrier and the modulator, the FM synthesis algorithm can be expressed as:

$$y(t) = A\text{cos}(\omega_c t + I\text{sin}\omega_m t) \tag{4.14}$$

Where:

$y(t)$ is the modulated signal.

$A$ is the amplitude of the modulated signal.

$\omega_c$ is the carrier frequency.

$\omega_m$ is the modutator frequency.

$I$ is the *modulation index* which is the ratio of the *peak deviation* to the modulating frequency. Peak deviation refers to the maximum amount by which the frequency of the signal may deviate from the carrier frequency.

Using the trigonometric identities, the Equation 4.14 can be expanded as:

$$
\begin{aligned}
y(t) &= A\cos(\omega_c t + I sin\omega_m t) \\
&= A\sum_{-\infty}^{\infty} J_n(I)\cos((\omega_c + n\omega_m)\cdot t)) \\
&= A(J_0(I)\cos(\omega_c t) + \sum_{n=1}^{\infty} J_n(I)\cos((\omega_c - n\omega_m)\cdot t) + (-1)^n\cos((\omega_c + n\omega_m)\cdot t))
\end{aligned}
\tag{4.15}
$$

Where $J_n(I)$ are the values of the *Bessel functions of the first kind* which is a pre-defined continuous function, at position $I$. Figure 4.9 gives the illustration of the first five Bessel functions of the first kind. As can be seen from Equation 4.15, there are series of frequency components $\omega_c \pm n\omega_m$. These frequencies are called *sidebands* which are a group of frequencies higher or lower than the carrier frequency. The amplitudes of sidebands are determined by the values of the Bessel function of the first kind. Figure 4.10 shows an example of the sideband caused by the FM process. As can be seen from this figure, for a fixed carrier frequency and a fixed modulation frequency, the sidebands are also fixed in frequencies, while the amplitude of each sideband is determined by the modulation index $I$. For small modulation index values ($I < 1$), the amplitude of the carrier is larger than the amplitudes of higher order sidebands. If the modulation index increases to a high value (over 1), the amplitudes for the higher order sidebands may be larger than the carrier. This brings the opportunity for FM synthesis suitable for generating sounds with rich harmonic content.

## 4.3 Applications of Procedural Audio

As procedural audio is both highly interactive and flexible, offering compromises in terms of plausibility to facilitate computational efficiencies, it is highly suitable for audio scenes in which a variety of peripheral sounds and/or sound effects are required to enhance the user experience. Hence, there is no doubt

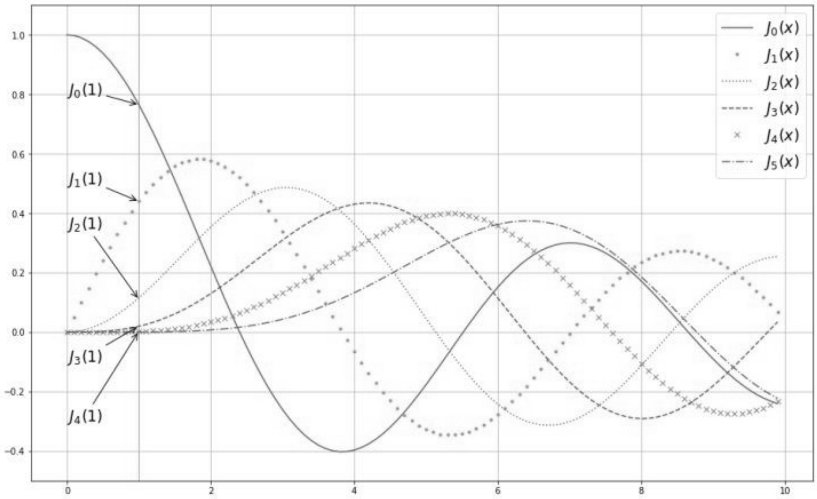


**Figure 4.9:** *Illustration of the first five Bessel functions of the first kind, reproduced from [13].*

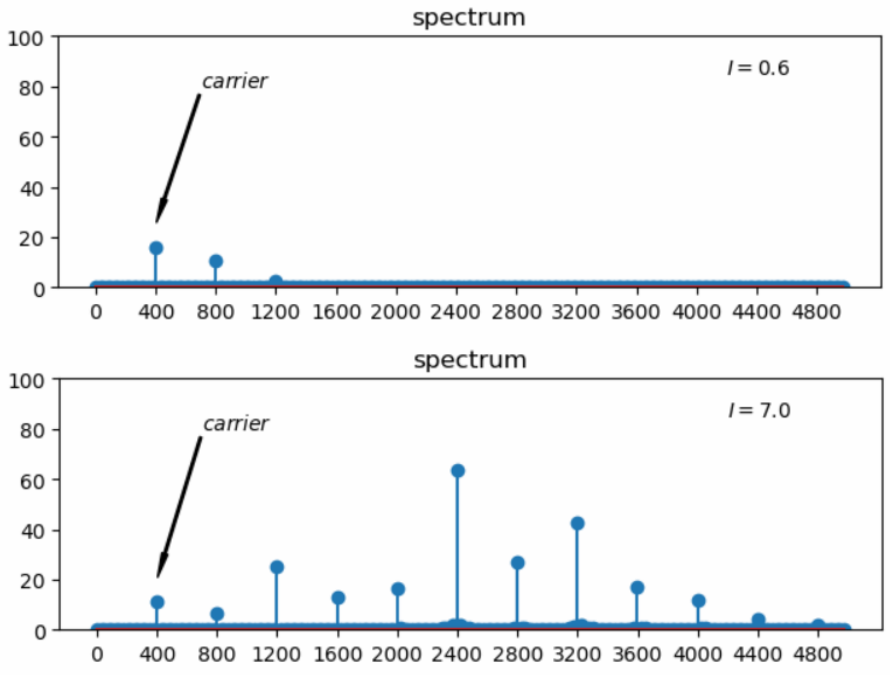


**Figure 4.10:** *Illustration of the sidebands caused by FM process with a small modulation index (above) and a large modulation index (below).*

that procedural audio has been most successfully applied in generating sound effects and music for video games, although it is still not as popular as sample-based audio techniques for modern commercially published games [146, 162].

In fact, the application of procedural audio can be traced back to the early 1980s when purely synthesised sounds were utilised for most video games due to the limitation on the computational resources at that time. Music and sound effects in these games were generated in real-time by hybrid analogue or digital oscillators, envelope controllers, filters, and amplifiers. Such music and audio effects were designed and integrated into different models of *sound chips*, such as 8039 8-bit Sound CPU, SID C64 sound chip [163], etc., shown as Figure 4.11. There have been some famous pieces of video game music and sound effects written via these sound chips, e.g. Super Mario Bros, Giana Sisters, and Bubble Bobble, etc. [164]. With the development of computer hardware and sample-based audio techniques, especially using compressed audio formats such as MP3 and AAC [165], recording-based audio has become more popularly used for games. This is mainly because there are more computational resources available now for personal computers and home video game consoles, as well as due to the general belief that procedural audio does not have as good sound quality, coupled with the lack of powerful tools for implementing procedural audio in commercial games today [166].

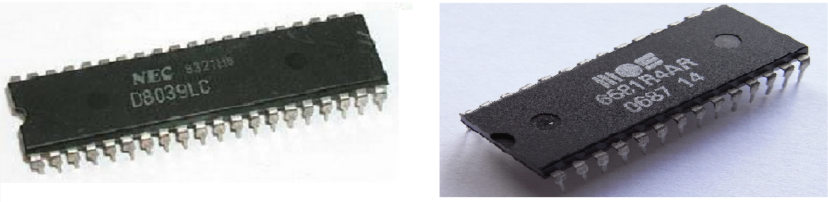


**Figure 4.11:** *Illustration of sound chips in the early days of game audio (left: 8039 Sound CPU, right: SID C64 sound chip, in the 1980s), reproduced from http://www.arcadeshop.com/ and https://www.c64-wiki.com/.*

In the past few years, there has been a renaissance in procedural audio and synthetic sounds in the sound design for video games and animation. On one hand, more kinds of game platforms have appeared in the market and become more and more popular, e.g. mobile devices, VR headsets, and online game services, etc. Computational power varies in different platforms, and sometimes it is impossible or not cost-effective to render all the details of a sound environment using purely recording-based audio techniques. On the other hand, there is an increasing number of *open-world games* and *sandbox games* [167] in the video game industry recently (e.g. Grand Theft Auto, Minecraft, etc.). These games are designed as non-linear with large open areas that can be explored by the players, giving them a great degree of creativity and freedom. It is not viable to pre-sequence everything in terms of graphics and audio for these open-world games. Instead, pragmatic rules are inevitable to render a specific environment both visually and audibly and this can be realised by procedural methods.

One of the main applications of procedural audio in video games is generating sound effects to avoid perceived repetition, e.g. walking/running footsteps [159, 162], sword swings [168], etc. If recording-based audio is used, a dataset consisting of multiple recordings for the different object statuses is required, and much time and effort would be invested in applying randomisation and post-processing to the resulting samples. With procedural audio, it is possible to generate numerous sound effects by tuning the synthesis algorithms appropriately, which could be designed as several presets with a 'one-click' solution. Another widely applied area for procedural audio is creating ambient sound effects for which features are varying over a short period but stable on a large time scale, such as the sound of wind blowing [169], fire burning [170], and water drops [171], etc. As these ambient sounds draw mainly the peripheral attention of the player, there is more tolerance on the compromises of plausibility and sound quality as long as the player is not asked to pay specific attention to these sounds. In fact, according to the study conducted by Böttcher et al. [166], when being actively involved in playing or purely observing a video recording of a game, the majority of players do not notice any difference between sample-based audio and procedural audio if they are not told to pay particular attention to the sounds.

Apart from video games, the application of procedural audio has been extended to other virtual environment scenes, such as replacing some 'Foley sounds' as used in the film industry [172], and rendering the acoustic environment for social meeting spaces in VR [173], etc. This technique has been also used for study to the area of *soundscapes*. A specific explanation of the term *soundscape* will be presented in Chapter 5. Here it can be considered as the overall experience of the sum of sounds in a specific scene. A valid soundscape synthesis where the user can interact with and change the soundscape might offer a way towards a useful creative design tool. As there is a common belief that procedural audio does not have as good sound quality as sample-based audio [146], procedural audio is often used for creating ambient background sound such as wind, water, and background traffic noise, etc. The foreground sources which plays the roles of *keynote* or *signal* according to the classification of soundscapes proposed by Schafer [174], are still synthesised by sample-based audio in existing soundscape synthesis tools currently, such as the soundscape synthesizers proposed in [175–177]. However, it is not always possible to judge clearly if a sound source is foreground or background, e.g. in the case of multiple vehicle pass-bys. In fact, it is usually a subjective process to determine whether recordings or synthetic sounds should be used for soundscape synthesis with complex factors to be considered in addition to plausibility. For example, it is not always possible or cost-effective to take on-site recordings that fulfil the specific requirement of a synthetic soundscape. Even though Finney and Janer [178] claimed that it might be feasible to use recordings from community-provided audio databases (e.g. Freesound.org [179]) to synthesise some specific soundscapes instead of real in-site recordings, the flexibility is quite limited compared to sound synthesis by procedural audio. Therefore, it is worth further exploring the potential of procedural audio in soundscape synthesis to understand more clearly to what extent plausibility can be achieved and under what circumstances this technique can be utilised.

## 4.4 Summary

This chapter has introduced the concepts and characteristics of procedural audio, and compared them with sample-based audio considering both advantages and disadvantages. As algorithms are the core assets for procedural audio models, an overview of some sound synthesis methods is presented, including additive synthesis, subtractive synthesis, spectral modelling synthesis, physical modelling synthesis, wavetable synthesis, and FM synthesis. The pros and cons of each synthesis method have been discussed, and it is found that when using procedural audio approaches, it is necessary to choose appropriate sound synthesis methods and integrate them accordingly for the specific sound scene to be rendered.

After an overview of its usage in video games and soundscape studies for creating sound effects and ambient sounds for interactive applications, it is found that there is a potential for procedural audio in a wider context because of its flexibility and interactivity, e.g. being embedded in an auralisation framework from an engineering perspective. As it is not always possible or cost-effective to take on-site recordings for auralisation purposes, especially in an outdoor environment, it is worth exploring the potential of procedural audio to understand more clearly about to what extent plausibility can be achieved and under what circumstances this technique can be utilised.

The next chapter will focus on the concepts and methodologies of environmental noise, with specific attention to traffic noise from the perspective of noise evaluation and soundscapes, in order to make a clear understanding of where and how the potentialities of procedural audio might be used in this field.

# Chapter 5

# Environmental Sound

This chapter will discuss the topic of environmental sound from a practical perspective. This means examining what is involved when discussing environmental sound, and how it is monitored, evaluated, and treated. The purpose of this chapter is to clarify the available evidence regarding the impact of environmental sound, especially traffic noise which is one of the most common outdoor sound sources in our daily lives. This will be explored in terms of relevant noise evaluation and prediction methodologies, and to consider how the design of our sound environment may be improved by using auralisation and procedural audio techniques as discussed in the previous chapters.

Here the term *noise* is defined as 'unwanted sound', or sound 'out of place' [25], e.g. a sound that is considered by a listener to be excessively loud, unexpected or unpleasant [180]. Following this definition, the term *environmental noise* is often used to describe the unwanted or harmful outdoor sound created by human activities, such as noise emitted by road traffic, rail traffic, air traffic, and from sites of industrial activity, etc. [26] Hence, there is no doubt that attitudes towards environmental noise tend to be negative, which is different from the double-sided assessment of natural environmental sound, sometimes considered to be positive, such as wind blowing, raining, and birds singing, etc. However, it has also been argued that a certain amount of environmental noise is required to give meaning to a location, and to make it inhabitable in its representation of the presence of human society [181]. A totally ‘silent’ environment is more unsettling than a noisy one, representing a dangerous environment not fit for habitation [26]. In fact, from the perspective of *soundscapes* research, environmental noise caused by human activities might also be considered as a resource rather than a waste, and it is important to design a soundscape with all the environmental sound blended to reflect the traditional or cultural elements of an area [124]. A further discussion on relevant soundscapes methodologies will be presented in Section 5.4 in this chapter.

According to the definition of noise used here, what can be noted is that it is a subjective process to judge whether a sound is ‘unwanted’ or ‘out of place’. The effects that a specific environmental sound can have are actually dependent on the life experiences and outlook of an individual. Whilst the effects of environmental sound are not fully understood by personal circumstances, there has been a lot of research on the negative effects of environmental noise which may lead to increased stress and its associated physical symptoms in some individuals [182]. Sudden loud noises and low level continuous noise are both known to cause sleep disturbance [183]. There is also some evidence to suggest a potential connection between environmental noise and cardiovascular and mental health problems [184]. These studies show the statistical significance between environmental sound and health-related consequences, which have been taken as a reference by some governmental authorities worldwide when making provisions for prevention and mitigation of environmental noise. Although legislation and management have been gradually improved and extended to cover different types of environmental sound, there are still blurred or blank areas of policies due to the fundamental problem of noise definition: what might be considered as noise by one individual is not necessarily noise to another individual or in another circumstance. These less-defined areas of policy can potentially be improved through the use of auralisation techniques which bring an audible experience of the time-varying content of environmental sound, and providing cues for a listener from a more perceptual perspective.

This chapter is developed from the content in Chapter 2 which covers the fundamental aspects of acoustics including sound generation mechanisms, propagation effects, and listening perception. The categorisation of environmental sound are discussed, followed by an overview of the effects of environmental noise. Then the methodologies of environmental noise evaluation and prediction are examined, which provide a context for the discussion of environmental noise from the perspective of soundscape theories and methodologies by the end of this chapter, revealing the potentiality, feasibility, and challenges for auralisation techniques in terms of environmental sound.

## 5.1 Categorisation of environmental sound

The most simple and straightforward way to categorise environmental sound is to follow the definition of the term ‘environmental noise’ introduced at the beginning of this chapter, that is, considering environmental sound as ‘wanted sounds’ and ‘unwanted sounds’, respectively. Here ‘wanted sounds’ can be considered the delightful, informative and/or meaningful sounds perceived within a sound environment, which support our environmental perception and cognition. In contrast, ‘unwanted sounds’ are often excessively loud, unexpected, or annoying, which have negative impacts on our perception of the sound environment or distract us from our ongoing tasks. Notwithstanding, this categorisation is too abstract and vague to be applied in practice, as a specific sound source under one circumstance may be considered either as ‘wanted’ or ‘unwanted’ by different individuals according to their life experiences and outlook. Therefore, it is necessary to find some more detailed taxonomies that can be used as guidance for the study of recording, detection, and evaluation of environmental sound in practice.

There have been many studies on the categorisation of environmental sound from both perceptual and engineering perspectives. Taxonomies with a perceptual perspective focus on the connection between perceptual properties related to human information processing, and acoustic features of a sound object or sound event. For example, Gygi et al. [185] proposed a clustering method for sound source type based on a multidimensional scaling analysis of the relationship between acoustic factors, such as temporal envelope and pitch measures, and the perceived similarity ratings for 50 environmental sounds. Taxonomies from an engineering perspective use perceptual factors as guidance and categorise environmental sound from the mechanism of sound generation. For instance, a widely used classification method for everyday environmental sound consists of three main groups, which are natural (animal sounds, wind blowing, water flowing, etc.), human (non-mechanical human activities, footsteps), and mechanical (traffic noise, industrial sound, airplane noise, etc.). Salamon et al. [14] propose a breakdown of this broad engineering categorisation into a series of lower-level hierarchies, together with an extra top level group introduced as music, as shown in Figure 5.1. This taxonomy has been widely applied in a series of urban sound studies, such as sound labelling [186], sound event classification [187, 188], and environmental noise monitoring [189]. Another example of a taxonomy from an engineering perspective is the physically-based sound-producing event method proposed by Gaver et al [15], in which a hierarchy consisting of a series of descriptive basic ‘sonic events’ caused by liquids (dripping, splashing, etc.), gases (explosion, wind, etc.), and solids (impact, rolling, scraping, etc.) is used to categorise different environmental sound, as shown in Figure 5.2. In this thesis, the taxonomy proposed by Salamon et al. [14] will be utilised as it is aligned with the framework of auralisation and compatible with most proposed taxonomies, providing sufficient low-level details about sound sources in an outdoor environment from an engineering perspective, such as ‘engine’ and ‘wheels’ for vehicle pass-by noise.

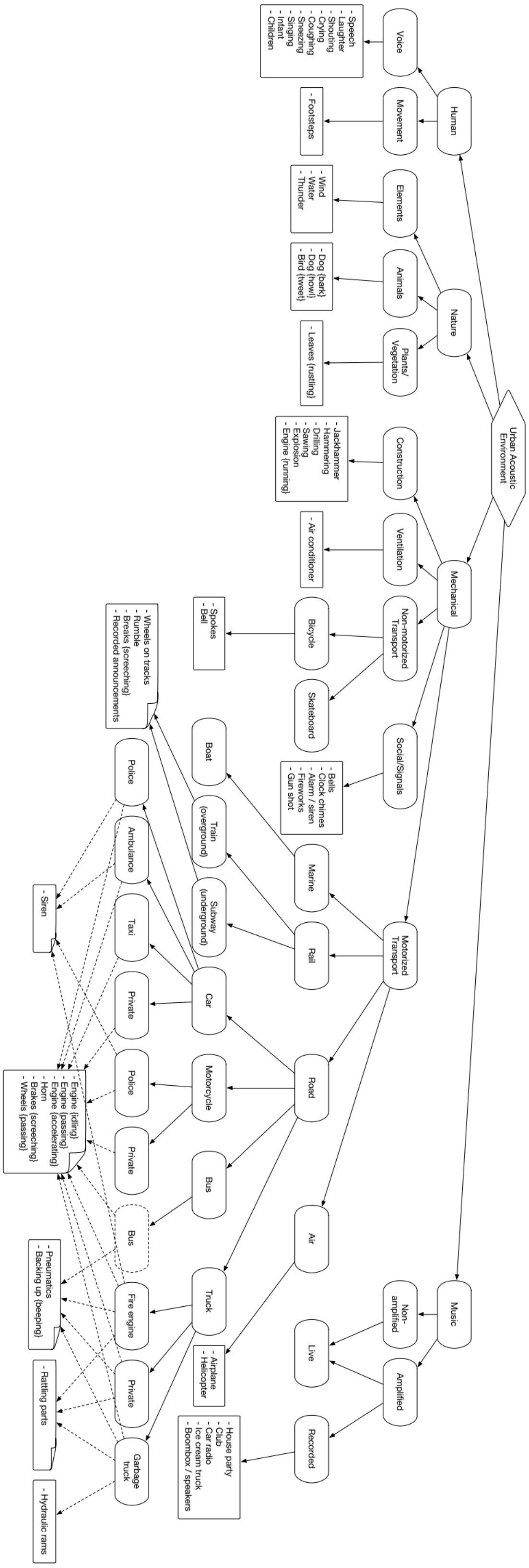


**Figure 5.1:** *Urban sound taxonomy proposed by Salamon et al. [14]. This is a breakdown of environmental sound via an engineering categorisation method, with a series of lower-level hierarchies, together with an extra top level group introduced as music used for clustering common environmental sound in our daily lives.*

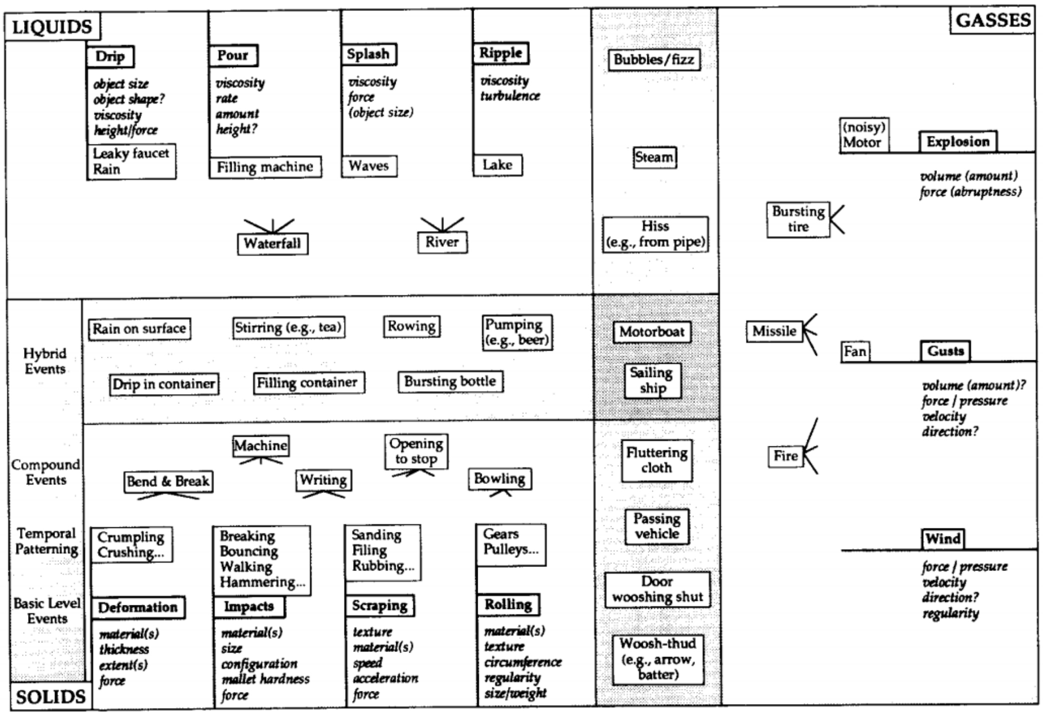


**Figure 5.2:** *Environmental sound taxonomy proposed by Gaver et al., reproduced from [15]. A hierarchy consisting of a series of descriptive basic 'sonic events' caused by liquids (dripping, splashing, etc.), gases (explosion, wind, etc.), and solids (impact, rolling, scraping, etc.) is used to categorise different environmental sound.*

## 5.2 Environmental Noise Impacts

Although the impacts that a specific environmental sound might have on an individual are actually dependent on the life experiences and outlook of that individual, and may also vary from person to person, a series of studies have proved that there is a statistical correlation between the occurrence of environmental sound and various health-related consequences, where the negative influences caused by environmental noise are of increasingly high interest. These negative influences can be separated into two categories: auditory effects and non-auditory effects.

### 5.2.1 Auditory effects

The adverse auditory effects of environmental noise are primarily related to hearing loss [190]. There are two types of hearing loss, which are permanent hearing loss called *permanent threshold shift (PTS)*, and temporary hearing loss called *temporary threshold shift (TTS)*. PTS is non-reversible, whilst TTS can be still fully recovered once the sound level is decreased [191].

There is a high risk of permanent hearing loss when an individual is exposed to a sound environment for which the SPL is higher than 90dB(A) for a long time (in relevant studies and regulations [192, 193], this is quantified as an 8-hour continuous equivalent sound level). For some individuals, the risk threshold for long-term noise exposure may become lower than 90 dB(A), e.g. around 80 dB(A) [193]. Permanent hearing loss is a cumulative process that is often neglected by people until it is severe enough to interfere with their daily activities, such as difficulty in clearly hearing speech during normal communication. Permanent hearing loss usually begins in the high-frequency range of 3k–6kHz, typically around 4kHz – the most sensitive frequency range for our hearing system. The impaired frequency range will gradually extend as the exposure time or the sound level increases, in addition to other influence factors such as ageing [193].

Temporary hearing loss refers to the decreased sensitivity of the ear to certain frequencies for a short period of time, with the recovery dictated by the age of the individual and the nature of the sound source [194]. TTS is relatively less common than PTS in terms of environmental noise, as it often happens due to biological reasons such as a blocked ear canal, middle ear infections, and/or head trauma, etc. Indeed it can happen while exposed to extreme loud noises (over 130 dB(A)) such as a rocket launching, rock music concerts, or sports events, and these are more likely to be occupation related noise rather than common environmental noise sources for most people.

Apart from hearing loss, another phenomenon related to hearing impairment is *tinnitus*, which refers to the experience of sounds like ‘ringing’ or ‘buzzing’ in the ear or head without any external physical causes [195]. There might be complex reasons for tinnitus, and in many cases it is difficult to find an exact cause. In terms of environmental sound, it is interesting to note that one of the aspect of tinnitus therapy is playing white noise or broadband environmental noise (e.g. background traffic noise) via wearable sound generators at appropriate sound levels as a masker to reduce the perception of tinnitus, which is called *tinnitus masking (TM)* [196]. In fact, TM has been used in clinics for over 40 years, and has been proved to be effective for patients in the short term (typically 3–6 months) [196, 197]. This reveals that environmental noise may have not only negative impacts, but also positive effects under some circumstances.

### 5.2.2 Non-auditory effects

Non-auditory effects caused by environmental noise are not as well understood as auditory effects. On one hand, there is a wide range of studies in a variety of relevant fields pertaining to the non-auditory effects of environmental noise, including studies in physiology, psychology, cardiology, neural science, and sociology, etc. As different aspects of non-auditory effects are explored in different disciplines and different methodologies are used to explore specific relevant research questions, it is difficult to summarise all findings more holistically. On the other hand, most of these studies focus on a wider context of noise rather than purely environmental noise, e.g. occupational related noise. Results from these studies cannot be directly applied in the context of environmental noise, although some evidence can be extracted and taken as reference because of the similarity between a specific noise context and the general conditions of the perception of that sound, e.g. working conditions with long-term exposure to road traffic noise [198–200].

On the basis of the categorisation proposed by the World Health Organisation (WHO) [193] and the British Medical Bulletin [201], some of the main categories of non-auditory effects caused by environmental noise can be summarised in the list below.

- **Sleep disturbances** – There are various kinds of sleep disturbance that may be attributed to noise, which include difficulty in falling asleep;

awakenings; and alterations of sleep stages or depth, etc. According to the study conducted by Xiang et al. [202], sleep quality in a silent room (close to 0 dB(A)) can be improved in terms of total sleep time, deep sleep time, and rapid eye movement (REM) sleep time. However, there is still a lack of quantitative evidence showing how background noise in common bedrooms (typically 22–48 dB(A)) influences sleep, and to what extent the sleep quality is affected.

- **Increased risk of cardiovascular disease** – In some studies the risk of hypertension and cardiovascular disease is reported to be higher in populations living in noisy areas around airports and noisy streets [203, 204]. However, in some other investigations there is no obvious association between blood pressure and noise levels [205]. Hence, the overall evidence suggests that there is a weak association between long-term noise exposure and blood pressure elevation or hypertension [201]. Environmental noise may be a minor risk factor for coronary heart disease [201].

- **Mental health effects** – Whilst environmental noise is not considered as a main cause of mental illness, it may worsen mental disorders for some susceptible individuals [193]. However, the relationships between noise annoyance, noise sensitivity and mental morbidity is complex and not yet well differentiated.

- **Impaired task performance** – Several studies have indicated that environmental noise can have adverse effects on cognitive and motivational performance for children both in the classroom and at home [193] . It is assumed there might be a similar impact on adults, but there is lack of quantitative study to support this claim.

- **Negative social behavior and annoyance** – Noise can be considered as an environmental stressor which may provoke aggression, unfriendliness, and other antisocial behaviours. It has been proved that there will be a decreased willingness to help others during exposure to noise as well as for a time period after exposure [193].

- **Interference with spoken communication** – This problem occurs when simultaneous noise masks speech in the critical frequency range (300–3kHz), degrading the intelligibility of speech. It often happens in populations living in noisy areas around airports and busy streets when making conversation, watching television or online videos and disturbed by aircraft noise or traffic noise [201].

## 5.3 Environmental Noise Evaluation and Prediction

### 5.3.1 Environmental Noise Evaluation

There have been several standardised environmental noise evaluation metrics based on sound levels introduced in Chapter 2, e.g. sound pressure level (SPL), sound intensity level (SIL), and sound power level, etc. In most of the international standards or regional regulations, SPLs with different weighting curves are used for guidance when it comes to noise control. As discussed in Chapter 2, these weighting curves are used to approximate the loudness perceived by the human hearing system at different frequencies. Some of the standardised weighting curves, including A-filter, B-filter, and C-filter, are shown in Figure 2.12. Among them, A-weighting is the most commonly used for everyday noise evaluation, while C-weighting is often used in addition to A-weighting for monitoring industrial noise and some occupational related noise [206].

To describe noise impacts in more detail (i.e. long-term properties), there have been several descriptive metrics based on SPLs, standardised by International Standards Organisation (ISO) in ISO 1996: Description, measurement and assessment of environmental noise [207], as listed below:

- $L_{A_{eq},T}$ – the A-weighted, equivalent continuous sound pressure level averaged over the measurement period, $T$. It is defined as:

$$L_{A_{eq},T} = 10log\frac{1/T\int_{t_1}^{t_2} p_A^2(t)dt}{p_{ref}^2}(dB(A)) \tag{5.1}$$

Where $p_A(t)$ is the A-weighted instantaneous sound pressure at running time $t$, and $p_{ref} = 2\times\ 10^{-5}$Pa is the threshold of the human hearing system in terms of sound pressure. By Fourier analysis it is possible to obtain the extended $L_{A_{eq},T,f}$ readings in octave or 1/3 octave bands centred on $f$ (Hz).

- $L_{A_{max}}/L_{A_{min}}$ – the maximum/minimum value that the A-weighted sound pressure level reaches during a given measurement period. This can be extended into $L_{A_{max/min}}$ *Slow* or $L_{A_{max/min}S}$ meaning the measurement is averaged over one second, and $L_{A_{max/min}}$ *Fast* or $L_{A_{max/min}F}$ meaning the measurement averaged over 0.125 seconds.

- $L_{90}/L_{50}/L_{10}$ – the noise level exceeded for 90%/50%/10% of the measurement period. $L_{90}$ is used to quantify the background noise level in a particular environment; $L_{50}$ is used to describe the median of the fluctuating noise levels; $L_{10}$ is used in road traffic noise assessment.

- $L_{DEN}$ – the weighted noise level over a 24 hour period ('DEN' stands for the abbreviation of 'Day-Evening-Night'). It is defined as:

$$L_{DEN} = 10log_{10}\frac{1}{24}(d\times 10^{\frac{L_{Day}}{10}+K_d}+e\times 10^{\frac{L_{Evening}+K_e}{10}}+(24-d-e)\times 10^{\frac{L_{Night}+K_n}{10}}) \tag{5.2}$$

wherein $L_{Day}$, $L_{Evening}$ and $L_{Night}$ are the long-term average day, evening, and night SPLs in dB(A) including adjustments for sound sources and sound character, respectively. $d$, $e$, and $24-d-e$ are the number of daytime hours, evening hours, and night hours, respectively. $K_d$, $K_e$

and $K_n$ are empirical adjustments for daytime, evening, and night-time, respectively, with considerations on the increased likelihood of annoyance or interference with activities during these periods.

On the basis of these descriptive metrics, several measuring and evaluation methods have been developed, and some of them have been standardised. As the content of environmental noise varies significantly and can be complex in terms of temporal and frequency variations in different scenarios, there is no one-size-fits-all method for the measurement and evaluation of all kinds of noise. Generally, the choice of methodologies should be dependent on the nature of the noise and the type of activity to which it is relevant. Some standardised noise measurement and evaluation methods are summarised as below.

### ISO 1996

ISO 1996, titled 'Acoustics: Description, Measurement and Assessment of Environmental Noise' [207] is one of the most widely used standards for assessing environmental noise. The instrumentation for such acoustical measurements should be verified and calibrated according to standards IEC 61672-1 [33], IEC 61260 [208] and IEC 60942 [209]. The measurement procedures are designed according to the characteristics of different types of environmental noise sources such as road traffic, rail traffic, air traffic, and industrial plants. Both long-term measurement and short-term measurements are included, and the results are presented as $L_{eq}$, $L_{max/min}$, and $L_{DEN}$ in octave or 1/3 octave bands. It is noted that this standard may underestimate the annoyance caused by low frequency noise (typically 20–160 Hz) due to the A-weighting applied to the measurements [210]. Under circumstances when low frequencies are of high interest, variations on the standardised measurement procedures should be considered in relation to specific noise problems.

### BS 4142

The British national standard BS 4142, titled 'Method for Rating Industrial Noise Affecting Mixed Residential and Industrial Areas' [211], has been widely used as a guidance for noise control and noise abatement for a long time in the UK. In BS 4142, the *specific* noise level of a target sound source is calculated by subtracting the *residual* and *background* noise from the *ambient* noise level. The ambient noise level is the overall noise level measured in $L_{A_{eq}}$, including the target sound source in the space. The residual and background noise are measured as $L_{A_{eq}}$ and $L_{90}$ for all the noise in the space excluding the specific sound source. The specific noise level obtained this way is called the *rating noise level*. Some extra adjustments should be considered if the target sound source has particular properties in the time-frequency domain, such as rich harmonic content, impulsiveness, or intermittent behaviour. The rating noise level is compared to the background noise level, which yields the likelihood of annoyance due to the target sound source. Compared to ISO 1996, BS 4142 is more convenient for applications regarding noise complaints by the community, as it provides cues to build the relationship between the time-averaged noise levels and the perceived annoyance. However, the robustness and effectiveness of this relationship can be debatable under some circumstances, e.g. when both the background noise and the target sound source levels are low, which is common for rural areas [212]. Again, just as ISO 1996, using BS 4142 may also underestimate the annoyance caused by low frequency noises, due to the A-weighting measurement used.

### DIN 45680

The German national standard DIN 45680, titled 'Measurement and Assessment of Low-frequency Noise Immissions' [2], has been used as a guidance for evaluating annoyance caused by low frequency noise. In this standard, the non-tonal noise signal in 1/3 octave bands between 10Hz and 100Hz is measured in dB without any weighting filters. This frequency range exceeds human hearing threshold in low frequencies because it is considered that low frequency sounds can have non-auditory effects on people and cause annoyance. A series of limiting values for the night-time sound levels in each 1/3 octave band is presented in DIN 45680, as shown in Table 5.1. Although this standard is designed for evaluation of non-tonal environmental sound, some research shows that it also corresponds well with the evaluation of disturbance when some tonal sounds are involved, such as entertainment music in the low frequency range [213]. This reveals the potential application of DIN 45680 in a wider context for evaluating low frequency sounds in an outdoor environment.

Table 5.1: DIN 45680 night-time low frequency noise limits, reproduced from [2]

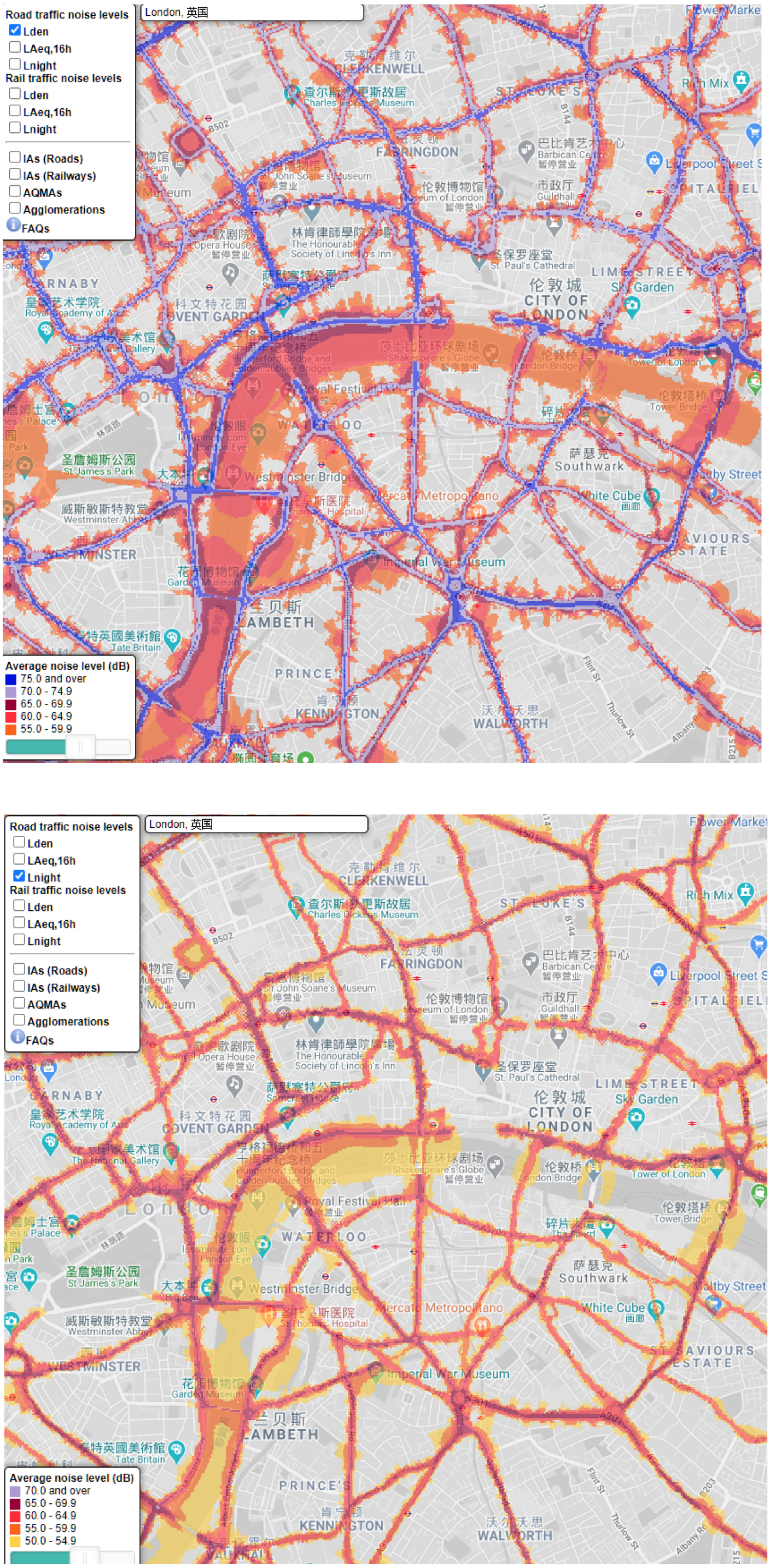
| **Frequency (Hz)** | 10 | 12.5 | 16 | 20 | 25 | 31.5 | 40 | 50 | 63 | 80 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Sound Level (dB)** | 95 | 86.5 | 79 | 71 | 63 | 55.5 | 48 | 40 | 33.5 | 33 | 33.5 |

### 5.3.2 Environmental Noise Prediction

In reality, it is often not feasible or cost-effective to take direct measurements at all potential receiver positions to obtain complete noise level data within a space. Some extrapolation and calculation methods are required to estimate the sound levels in some locations where it is impossible to carry out measurements because of excessive background noise. There have been many *noise prediction models* developed for this purpose, such as ISO 9613 [129], Nord2000 [130], Harmonoise [131], CNOSSOS-EU [214], and ASJ RTN-Model [215], etc. These models are designed from an engineering perspective, predicting the outdoor propagation effects of road traffic, rail traffic, and aircraft noise from measured data by empirical equations so as to extrapolate or calculate sound levels at target locations. These empirical equations are derived from existing measurement data or numerical approximation of the relevant theoretical equations, in terms of geomorphological conditions, meteorological conditions, and wave phenomena such as effects of diffraction and scattering zones. Most such engineering methods are limited to calculations of SPLs in dB(A) for a specific frequency range, e.g. 25–10kHz in the Nord2000 [130] and Harmonoise [131] models. What should be noted is that the empirical equations and input parameters in different engineering models may be different from each other, more or less, based on the specific content of sound source for which a model is appropriate for and the level of detail required in the modelling method. Even though some engineering models share very similar principles and methodologies for modelling techniques with similar configurations for the input, their calculated results may not agree with each other in some cases [216]. Furthermore, most engineering models are developed on the basis of a given local environment with the configurations calibrated by local traffic conditions, meteorological conditions, and geomorphological conditions. Therefore, it is essential to verify and fine-tune an engineering model before applying it in reality within a specific area.

One of the most straightforward applications of such noise prediction models is *noise mapping*, which refers to a map of a region marked by colours and contours according to the SPLs distributed within the area. In order to develop a noise map for a given area, a number of measurement positions should be selected where $L_{A_{eq},T}$ will be measured over a long-term period (e.g. from several months to one year). Other descriptive metrics such as $L_{Night}$ and $L_{DEN}$ can be derived from the relevant $L_{A_{eq},T}$ values. In order to obtain the noise levels between the measurement positions, noise prediction models based on engineering approaches are used to generate a map for the whole area. Figure 5.3 shows an example of the noise map near the City of London area produced by an engineering consultation company following the guidance of the European Noise Directive (END) Directive 2002/49/EC project [217]. The left part is the road traffic noise map in $L_{DEN}$, and right part is the noise map in $L_{Night}$.

Whilst noise maps can provide useful cues regarding long-term time-averaged noise levels in a visualised manner, there are some limitations when


London, 英国

Road traffic noise levels
☑ Lden
□ LAeq,16h
□ Lnight
Rail traffic noise levels
□ Lden
□ LAeq,16h
□ Lnight

□ IAs (Roads)
□ IAs (Railways)
□ AQMAs
□ Agglomerations
FAQs

Average noise level (dB)
🟦 75.0 and over
🟪 70.0 - 74.9
🟥 65.0 - 69.9
🟥 60.0 - 64.9
🟧 55.0 - 59.9

London, 英国

Road traffic noise levels
□ Lden
□ LAeq,16h
☑ Lnight
Rail traffic noise levels
□ Lden
□ LAeq,16h
□ Lnight

□ IAs (Roads)
□ IAs (Railways)
□ AQMAs
□ Agglomerations
FAQs

Average noise level (dB)
🟪 70.0 and over
🟥 65.0 - 69.9
🟥 60.0 - 64.9
🟧 55.0 - 59.9
🟨 50.0 - 54.9

**Figure 5.3:** *Noise map near the City of London area produced by an engineering consultation company. The upper picture is the road traffic noise map in* $L_{DEN}$*, and lower picture is the noise map in* $L_{Night}$.

environmental sounds are illustrated this way. Firstly, the types of sound sources involved in the noise mapping are usually limited to rail traffic, road traffic, aircraft, and industrial noise, etc. Other sources, such as entertainment music which might be a predominant sound source in busy streets, are not considered. Secondly, noise mapping can provide only long-term time-averaged sound levels. It cannot be used to describe dynamic sounds with short-term variations, such as a vehicle or a train pass-by, where, in the near field, noise levels rise significantly for several seconds as a vehicle or a train is coming, and rapidly go down to the background noise level after it has passed. Lastly, noise mapping does not necessarily provide an impression on what an area sounds like. The sound environments across different areas can be extremely different when they are heard, but they may have a similar level of long-term time-averaged SPLs. For example, the audible experience in a residential area near an airport is quite different from the sound perceived in a busy street, although they may be coloured the same in a noise map due to the similarity in SPLs. These issues can be potentially tackled using auralisation techniques with appropriate modelling of sound sources, sound propagation effects, and spatial audio reproduction. For example, Finne and Fryd [218] proposed a method to create auralisations of road traffic noise to support the official noise maps used by the Danish Road Administration with audible experience of what a new road project sounds like after construction. The auralisation is based for headphone playback and has been published online for a wide range of participants. Binaural recordings of single vehicle pass-bys are used for rendering sound sources with short-term dynamic variations. Most of these recordings were taken on late evenings and nights to avoid significant background noise (e.g. singing birds, agricultural machinery sounds, etc.). Propagation effects are modelled according to the Nord2000 engineering method [130]. It has been reported that there was a great deal of public interest in the auralisation of this new road project which has improved communication with the public, and this is now becoming a standard part of presenting future road projects in Denmark when noise is considered to be an important issue [218].

## 5.4 Soundscape Methodologies

### 5.4.1 Soundscape Concepts

Apart from conventional noise evaluation and prediction methods based on SPLs, there is a growing field of soundscape studies which focuses on the prediction and evaluation of the holistic experience of all sounds in a given environment from a subjective listening and perceptual perspective, using a more interdisciplinary approach. Generally, soundscape can be considered as the aural analogy of landscape which refers to all the visible features of an area of land. One of the most widely used definitions of the term *soundscape* was proposed by Schafer as follows: [174]:

> *The sonic environment. Technically, any portion of the sonic environment regarded as a field for study. The term may refer to actual environments, or to abstract constructions such as musical compositions and tape montages, particularly when considered as an environment.*

Compared to conventional environmental noise evaluation methods based on SPLs, soundscape analysis looks at how the sounds in an environment are perceived by human beings together with the relevant impacts on a population and society more generally. This is further interpreted and highlighted in other definitions of soundscape, such as, '*the environment of sound (sonic environment) with emphasis on the way it is perceived and understood by the individual, or by a society*', proposed by Truax [219], and, '*the acoustic environment as perceived or experienced and/or understood by a person or people, in context*', defined in ISO 12913-1:2014 [220].

As can be seen from these definitions, when using soundscape methodologies, not only the negative impacts caused by unwanted environmental noise, but also the positive aspects of some sounds in a given environment that may be crucial for the audible perception of the environment for informative, aesthetic, or epistemic reasons, should be considered, from a holistic perspective. Therefore, soundscape is a multidisciplinary area which involves not only quantitative measurement and analysis of physical parameters such as SPLs, but also qualitative study of the subjective impression of sounds in terms of how, when and why specific sounds are preferred or not via subjective evaluation methods such as questionnaires and/or interviews.

To study the content and nature of different soundscapes, it is necessary to break down holistic sound environments into appropriate levels of detail, in other words, defining categories of environmental sound to which different sound sources belong. One of the most used categorisation methods is that proposed by Schafer [174], in which all the sounds in an environment can be decomposed into three main categories: *keynotes*, *sound marks*, and *sound signals*. Keynotes are sounds, 'which are heard by a particular society continuously or frequently enough to form a background against which other sounds are perceived' [174]. For example, vehicle pass-by sounds can be seen as keynotes in an urban area with busy traffic. Sound signals, by contrast, refer to sounds which aim to draw attention for further interaction with the environment or to take action, such as fire alarms or sirens. Sound marks are analogous to landmarks, which are associated with the unique cultural or historical background of the community in the given environment from a listening perspective, for example the bells ringing from York Minster which can be heard over much of the city or York.

Another popular categorisation method found in much literature [8, 221–224] is the division of environmental sound into *natural*, *human*, and *mechanical*. Natural sounds refer to sounds not directly related to human activities, such as animal sounds, wind blowing, and water flowing, etc. Human sounds include sounds that are representative of human activity without mechanical facilities, such as footsteps, speech, and laughter. Mechanical sounds cover mainly industrial activities such as traffic noise, construction noise, and aircraft noise. This categorisation method is congruent with Schafer's taxonomy, and is more convenient as a guide for practical implementation of recording and analysis of a set of soundscapes, especially when capturing a wide range of environmental sound sources, such as the soundscape databases established in [14] and [225].

In 2018, the ISO published the technical specifications *ISO/TS 12913-2: 2018-Soundscape-Part 2: Data collection and reporting requirements*, in which the taxonomy of soundscapes has been further broken down into lower levels of detail, as shown in Figure 5.4. This preliminary standard provides more detailed guidance on soundscape recordings and analysis when a wide range of sound sources are involved. What should be noted is that there is no strict boundary between these different categories, and a sound may belong to multiple categories at the same time depending on context. For example, the sounds of running water in a constructed area may be categorised as both 'nature' and 'sounds generated by human activity/facility' at the same time according to different contexts.

### 5.4.2 Soundscape Evaluation

So far, such soundscape evaluation methods are still under development and being standardised. Although there have been some preliminary standardised evaluation methods published, such as in the *Soundwalk Method A/B/C* in *ISO/TS 12913-2: 2018-Soundscape-Part 2: Data collection and reporting requirements*, the feasibility and the validation of these methods are still debatable and further research is required to provide more robust evidence for their use and efficiency [226]. As soundscape is a multidimensional phenomenon focusing on people's listening and perceptual experience, there are two main methods for soundscape evaluation: subjective and objective.

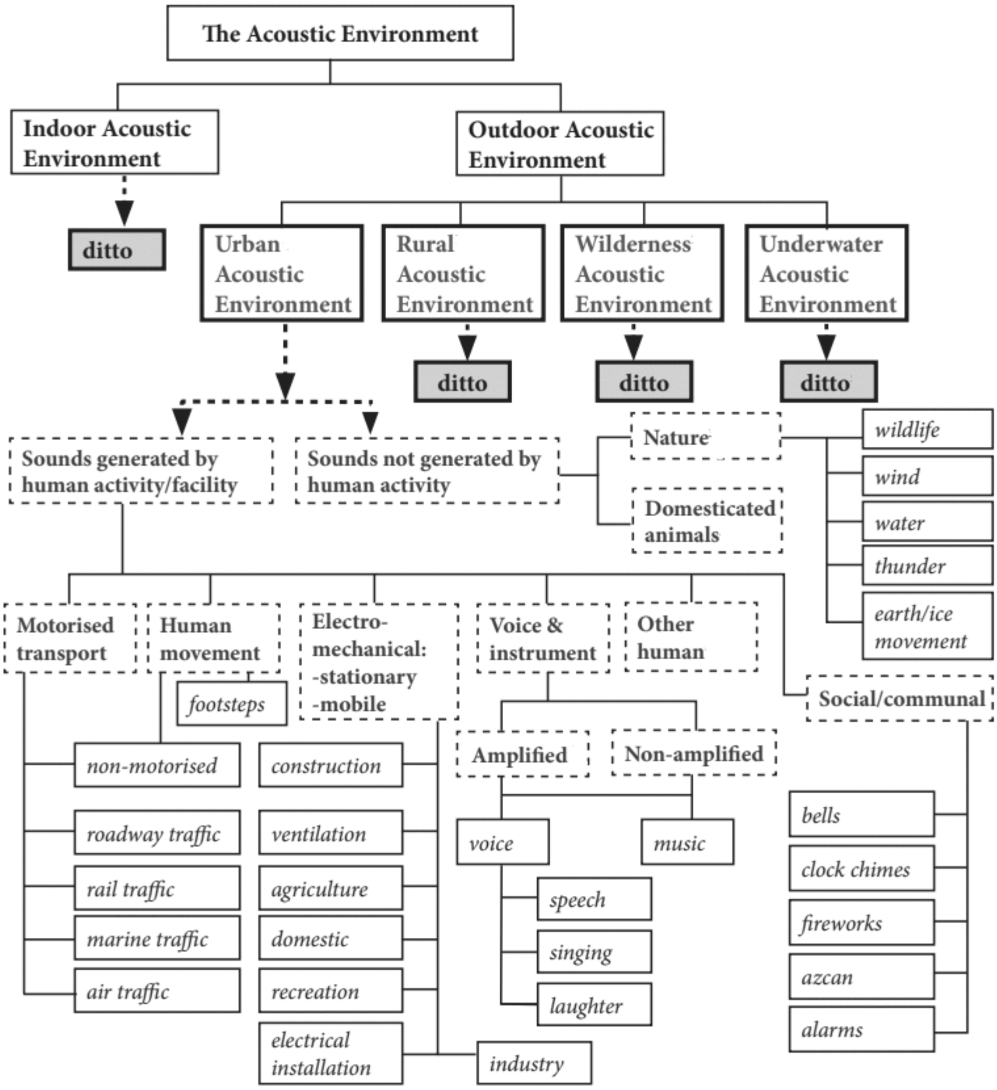
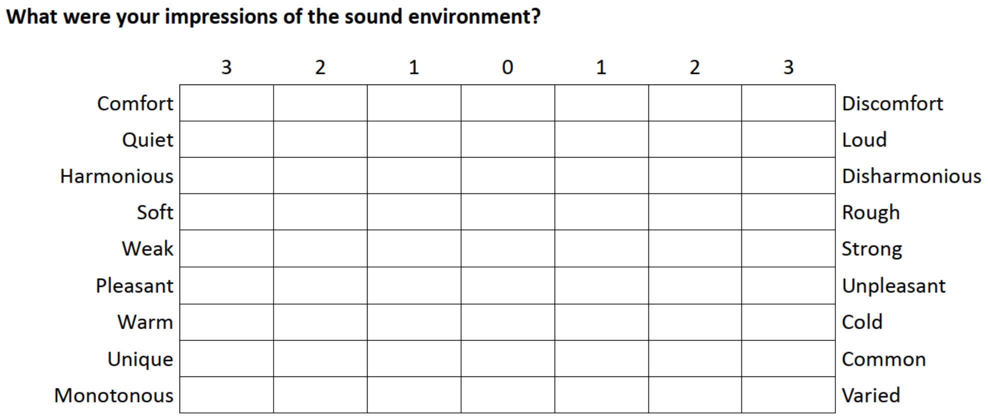


**Figure 5.4:** *Soundscapes taxonomy proposed in ISO/TS 12913-2: 2018-Soundscape-Part 2: Data collection and reporting requirements, reproduced after [16].*

### Subjective Evaluation

Subjective evaluation aims to understand soundscapes via observation of people's emotional and behavioural responses. This process is typically conducted by using questionnaires via a *soundwalk*, which encourages individuals to discriminatively listen to the sounds present within a space or the sound environment over a given period, and to make personal judgements about the sounds that have heard [227]. There are a lot of methods for collecting the data of these personal judgements. One commonly used method is *semantic differential (SD)* [5, 17]. When using this method, a series of bipolar pairs should be chosen to describe the soundscape across multiple subjective dimensions, and with a numerical scale used for each bipolar pair for participants to judge to what extent the soundscape can be described as these polar adjectives. Figure 5.5 provides an example of semantic differential pairs used in a specific soundwalk test [5] for evaluating an outdoor sound environment in an urban area. Another example of SD pairs chosen by Kang and Zhang [17] in the study of soundscapes of public spaces is shown as Figure 5.6.



**What were your impressions of the sound environment?**

| | 3 | 2 | 1 | 0 | 1 | 2 | 3 | |
|---|---|---|---|---|---|---|---|---|
| Comfort | | | | | | | | Discomfort |
| Quiet | | | | | | | | Loud |
| Harmonious | | | | | | | | Disharmonious |
| Soft | | | | | | | | Rough |
| Weak | | | | | | | | Strong |
| Pleasant | | | | | | | | Unpleasant |
| Warm | | | | | | | | Cold |
| Unique | | | | | | | | Common |
| Monotonous | | | | | | | | Varied |

**Figure 5.5:** *An example of semantic differential pairs used for the evaluation of an outdoor sound environment in an urban area, reproduced from [5].*

As can be seen from these examples, different terms have been selected according to the context of the different soundscapes. In fact, how to choose suitable descriptive pairs is an inherent challenge for SD approaches. It is still an open question as to how bipolar pairs should be interpreted and selected

| **Semantic Differential Pairs** | | | |
|---:|:---|---:|:---|
| Agitating | Calming | Comfort | Discomfort |
| Directional | Everywhere | Echoed | Deadly |
| Far | Close | Fast | Slow |
| Gentle | Harsh | Hard | Soft |
| Interesting | Boring | Like | Dislike |
| Meaningful | Meaningless | Natural | Artificial |
| Pleasant | Unpleasant | Quiet | Noisy |
| Rough | Smooth | Sharp | Flat |
| Social | Unsocial | Varied | Simple |
| Beautiful | Ugly | Bright | Dark |
| Friendly | Unfriendly | Happy | Sad |
| High | Low | Impure | Pure |
| Light | Heavy | Safe | Unsafe |
| Steady | Unsteady | Strong | Week |

**Figure 5.6:** *An example of semantic differential pairs chosen by Kang and Zhang in the study of the soundscape of public spaces, reproduced from [17].*

as being suitable for a specific context. Furthermore, there is still a lack of study on the robustness of the same descriptive pairs used under different circumstances for different context. The relationship between different pairs can be complex, especially when participants with different nationalities or native languages are involved, making it difficult to compare the results from different soundscape studies, especially when various semantic differential pairs are used. In order to tackle this issue, some methodologies have been proposed, such as the Spatial Audio Quality Inventory (SAQI) which attempts to create a consensus vocabulary that would be suitable for multiple languages [228], and the Self-Assessment Manikin (SAM) which uses pictograms instead of words to represent potential emotional responses to listening to a soundscape or other stimuli in order to help streamline understanding by different participants to some extent [8]. By correlation analysis, it is found that the SAM can be a directly comparable and useful tool to SD for the analysis of subjective soundscape experience [229].

Apart from soundwalk, another subjective evaluation method is *narrative interview*, which aims to collect qualitative data from local experts and daily users of a space without participating in an actual soundwalk [230]. Narrative interviews can be implemented by a series of structured and scheduled interviews with the target participants, and/or open questionnaires for the interviewees. On one hand, a narrative interview may enhance the comprehensiveness and robustness of the results as more local users and experts are involved. On the other hand, the results from narrative interviews can be used as a basis when tailoring questions and choosing specific semantic differential pairs in soundwalk studies. Narrative interview is currently being developed in addition to soundwalk towards a standard subjective soundscape evaluation method in the technical specifications *ISO/TS 12913-2: 2018-Soundscape-Part 2: Data collection and reporting requirements* [16].

### Objective Evaluation

The aim of objective evaluation methods is to seek appropriate objective metrics to describe the perception and behavioural responses of soundscapes based on physical or statistical parameters. A wide range of acoustical and psychoacoustical metrics may be considered according to the context of soundscapes, such as the SPLs, loudness, and reverberation time (RT), etc., as partly introduced in Chapter 2. However, there have been numerous criticisms on the correlation between such acoustic and psychoacoustic parameters and the perceived quality of the sound environment. Therefore, in soundscape studies, more hearing-related parameters have been proposed based on a combination of physical parameters and statistical analyses of the results from subjective evaluation in order to describe the perception of complex acoustic scenes in an objective way.

For example, Fiebig et al. [231] proposed an index of evaluation of complex traffic noise based on regression analysis of subjective evaluation results and a series of psychoacoustic parameters including loudness, sharpness, roughness, and impulsiveness. Lavandier and Defréville [232] proposed a descriptor defined as, ‘unpleasantness of sound’, based on SPLs and the relative duration of categories of sound sources in an urban environment (e.g. buses, motorbikes, children's voices, etc.). Ricciardi et al. [233] proposed an indicator called, 'sound quality', based on SPLs and multiple regression of a series of subjective perceptual variables such as 'visual amenity', 'overall loudness', etc., defined according to a given context.

As can be seen from these examples, although a variety of soundscape descriptors have been proposed for the evaluation of complex sound scenes, the definitions depend heavily on the context of the specific soundscape in the study. It is still an open question as to what extent an established objective metric can be useful in another soundscape context which is either similar or different from the validated condition. In order to solve this problem, a recent research project called, 'Soundscape Indices (SSID)' [234], aims to develop a series of descriptors that can reflect levels of human comfort in a wide range of soundscape contexts. Multidimensional factors from the perspectives of psychology, acoustics, psychoacoustics, neural science and physiology are considered according to a wide range of soundscape contexts. Some more detailed reviews of the impacts of soundscapes from the perspective of such diverse disciplines can be found in [235–238]. The methodologies for urban soundscape recording and questionnaire surveys implemented in SSID are proposed in [239], which aim to develop a large-scale, international soundscape database. Binaural recordings are used in SSID to restore spatial information as if a human listener is present in the recording position. There is still a debate about whether binaural measurements should be mandatory or whether monaural recordings are sufficiently adequate, although it is found that binaural recordings can render corresponding perceptual aspects more consistently than monaural counterparts for soundscape evaluation [240]. The SSID project is still ongoing and there are further steps to be implemented, e.g. identifying key factors and their influence on soundscape quality based using the database of psychological evaluations, developing and validating soundscape indices based on this database and identified key factors, and demonstrating the applicability of the identified soundscape indices in practice, etc.

### 5.4.3 Soundscape Design

Although the concept and evaluation methods for soundscapes are still under development, some elements have been already utilised in practice. Soundscape design aims to enhance the public's engagement with their acoustic surroundings through a better understanding of the nature of certain places and sounds. There are two typical cases for the practical implementation of soundscape design – in the physical world and in the virtual world.

Soundscape design in the physical world aims to use appropriate sound sources or sound propagation influencers so as to create a narrative for the space, improve the acoustic comfort of the space, or create a 'sound mark' based on a soundscape based analysis of the natural aspects of a particular environment. The utilised sound sources might be audio recordings replayed by, sometimes hidden, loudspeakers and public address (PA) systems, or physical real sound emitters/propagation influencers. For example, during the period of 2008 Beijing Olympic Games, a series of insect sounds were replayed by loudspeakers buried under the ground of the Beijing Olympic Village to enhance the 'natural' experience of the area [241]. Water sources, especially fountains, have been utilised in many public squares for soundscape design as maskers, which not only reduce noise annoyance, but also creates a new physical, and an associated soundscape that can benefits landscape and have impacts on public behaviour [242, 243]. Some examples of sound propagation influencers for soundscape design can be found in the application of sound sculptures, which uses physical sculptures with resonating cavities tuned specifically to the site to imitate an instrument played by, e.g. wind, or sea waves [244–246]. Most of these sound sculptures have become soundmarks as well as landmarks which contribute not only to the soundscapes by highlighting the sense of the 'sound of nature', but also have aesthetic, cultural, and economic impact on the local community [247] as an increasing number of tourists are attracted by the sound sculptures, such as Blackpool High Tide Organ (Blackpool, UK), Wave Organ (San Francisco, USA), and Sea Organ (Zadar, Croatia), etc.

Soundscape design based in a virtual world uses auralisation and acoustic modelling techniques to create soundscapes for enhancing the user's experience or engagement with the virtual environment which can be a simulation of the real world. For example, the engineering firm Arup Acoustic reportedly used auralisation with 3-D VR imaging techniques to demonstrate the impact of aircraft noise in the dwelling area near London Heathrow Airport [248] and the noise of the high-speed trains (HS2) [249]. This is a straightforward way to demonstrate the impact of environmental noise caused by large ongoing projects in reality which can influence the modification of proposed project designs through simply listening. It was reported that 25,000 people have listened to the HS2 auralisations during public consultation [249], which shows the strength of auralisation techniques in soundscape design for public's engagement. This is congruent with the findings in Finne and Fryd's study on auralisation for planning new roads [218], revealing that some natural sounds (e.g. birds singing) should be added to road traffic noise for such an auralisation so as to enhance the holistic context of the soundscape.

Another example of using auralisation in the creation of soundscapes for engaging with the public is the 'Listening to the Commons' project [250] from the University of York, which aimed to '*recover the soundscape of debate as experienced by women listening through a ventilator in the historic House of Commons c. 1800-34*'. The results of that auralisation were presented in the 2018 'Voice and Vote' exhibition at which most visitors were not acousticians [251]. It is proved that such an auralisation can be helpful for the visitors to engage with the digital outputs and recover women's experience of politics c. 1800-34.

What should be noted is that there is no standardised method for soundscape auralisation, and the resultant auralised scenes may vary significantly, depending on the methods and algorithms used. It is therefore important to use appropriate auralisation techniques carefully and tactfully according to the context of the sound environment, the reasons for requiring an auralisation in that place, as well as the practical restrictions (e.g. available data, real-time performance). It is also important to verify the plausibility of the final auralisation before wider public demonstration.

## 5.5 Summary

This chapter has covered the topic of environmental sound, including the underlying aspects of the definition, categorisation, impacts, evaluation and prediction methods, and the use of soundscape concepts and methodologies for public engagement. The aim of the literature review in this chapter is to find the potential and challenges of using auralisation techniques to support conventional noise evaluation/prediction methods and soundscape studies, especially for the traffic noise in an urban environment which is of high interest in this study.

After an overview of relevant literature, it is found that although environmental noise has negative influences on our perception of the environment and personal health, it is not sufficient to use conventional noise evaluation and prediction methods based on SPLs to fully understand the impacts caused by environmental sounds. This can be potentially or partly fixed by using auralisation techniques. From the perspective of soundscape, auralisation can also be a powerful tool for rendering audible experiences which is useful for the study of soundscape evaluation and soundscape design. As there has been no standardised method for environmental sound auralisation so far, the resultant auralised scenes may vary significantly, depending on the methods and algorithms used in the associated auralisation framework. These gaps provide the context for the following two chapters of this thesis: to develop a traffic flow auralisation framework (presented in Chapter 6), and to evaluate the plausibility of the proposed auralisation framework (presented in Chapter 7).

# Chapter 6

# Developing an Auralisation Framework for Vehicle Pass-bys using Procedural Audio

Based on the review of auralisation, procedural audio, and environmental sound in Chapters 3, 4, and 5, there seems to be a strong case for the development and deployment of an auralisation framework for environmental sound using procedural audio. Indeed, there have been several studies that take advantage of procedural audio in terms of its dynamic flexibility for creating natural sounds such as rain drops [171], fire [170], wind [169], etc., and human sounds such as footsteps [159]. Most of these approaches are developed for sound design purposes and used in video games and films instead of auralisation purposes from an engineering perspective. On the other hand, most of the existing auralisation studies in environmental noise, e.g. traffic noise [117, 126, 252], aircraft noise [253–255], and railway noise [111, 256] are based on in-situ or anechoic recordings, which has very limited flexibility for wide disseminating and interactive demonstration. Therefore, it is worthwhile to think about taking advantages of procedural audio in terms of flexibility and interactivity as applied to the auralisation of urban soundscapes which hold possibilities beyond the domain of urban planning and environmental noise assessment procedures, to a plethora of potential public engagement and artistic initiatives aimed at strengthening human-environment relationships [5]. The rest of this thesis will concentrate on exploring how such techniques might be combined and implemented in practice, what can be achieved, and to what extent such approaches are effective.

Road traffic noise has been chosen for exploring the auralisation techniques with procedural audio. There are several reasons for this choice. Firstly, it is one of the most common noise sources to be evaluated and managed in our daily lives, and the auralisation of traffic noise is crucial for conventional noise studies and soundscape research. As discussed in Chapter 3, a series of existing studies on the auralisation of road traffic noise has already been implemented, which can be used as a reference when considering procedural audio alternatives. Secondly, it is often not convenient or cost-effective to get all the required recordings for various noise sources in a road traffic scenario as anechoic recordings from different vehicle types running at different speeds and different driving patterns may be needed to make the auralisation variable and interactive. Procedural audio, according to the discussion in Chapter 4, has the advantage of such flexibility, which can be potentially utilised for rendering a wide range of vehicle sounds by tuning algorithms. Thirdly, as discussed in Chapter 4, the cost of computational resources is variable in procedural audio approaches, so there is a possibility to implement auralisation on different platforms with limited computational conditions, e.g. mobile devices and web platforms. This unique feature may be used to potentially extend the application scenarios for auralisation, which needs to be further investigated.

As yet, there are no standardised methods for traffic noise auralisation, and the resultant modelled scenes may vary significantly, depending on the methods and algorithms used. An auralisation framework for road traffic noise at a microscopic urban scene is therefore devised in this chapter to explore the possibility and feasibility of procedural audio in such a virtual acoustic environment, through a combination of sound source modelling, sound propagation modelling, and spatial audio reproduction. Here ‘a microscopic urban scene’ means the effect of individual vehicle and building surfaces should be considered in such an urban scene. A typical microscopic urban scene is the perceived scene of a person when he/she is standing at a city square or walking along a street in the city.

Figure 6.1 demonstrates an example of such a microscopic urban scene, which is a simulated urban intersection rendered in [18]. As can be seen from this figure, there are multiple types of sound sources, including cars, trams, and pedestrians in a microscopic urban scene. For such an auralisation, the effects of different sound sources and building surfaces should be considered individually.

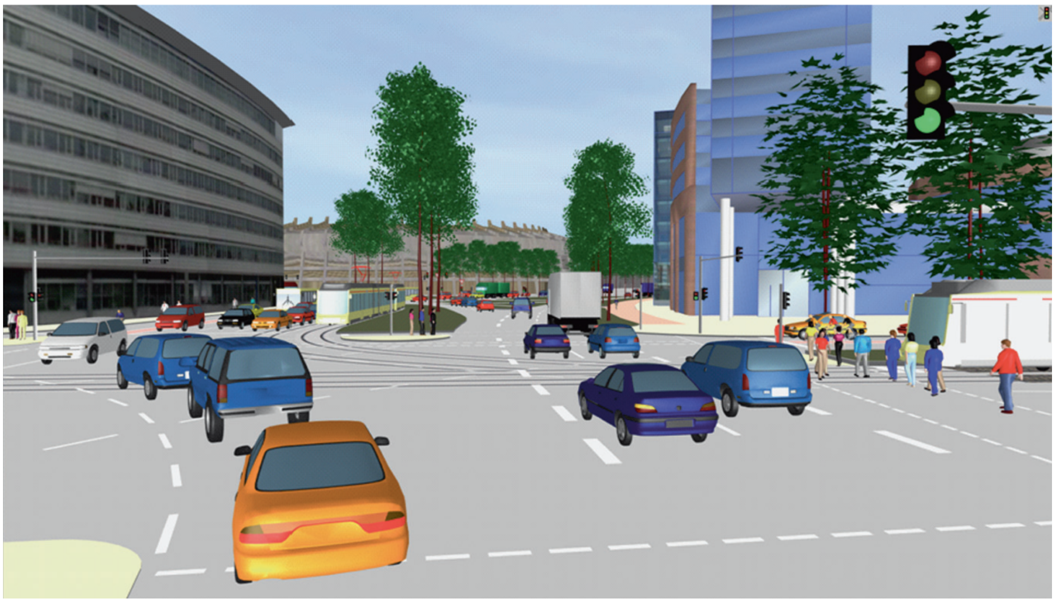


**Figure 6.1:** *An example of a microscopic urban scene rendering, reproduced from [18].*

This chapter will document the process of the development of the auralisation framework for such a microscopic urban scene. Whereas an overview of auralisation and procedural audio has been presented in Chapters 3 and 4, this chapter goes into more detail about what has been adopted and how the related models are developed and implemented.

This chapter is organised as follows: firstly it documents the methodology used to establish the source modelling for a single pass-by vehicle using procedural audio methods. Secondly, it deals with some imperative sound propagation effects for a moving vehicle in a microscopic urban scene, including distance attenuation, early reflections, and Doppler shift. Thirdly, binaural sound reproduction with HRTF processing is introduced for spatial audio rendering. Traffic flow auralisation is achieved by adding multiple auralised single vehicle pass-by sounds with a variety of vehicle types, speeds, and driving patterns, according to the descriptive parameters of the traffic flow. While the process described here is considered flexible for a variety of different scenarios, the algorithms should be tailored and tuned for specific scenarios. A comparison between the developed auralisation framework using procedural audio and its counterpart based on recording-based audio for a particular urban scene will be presented in the next chapter, so as to evaluate the perceived plausibility and validate the feasibility of using procedural audio approaches for road traffic noise auralisation.

## 6.1 Single Vehicle Pass-by Sound Source Modelling

According to some relevant studies in road traffic noise [119, 257, 258], a vehicle pass-by noise is mainly composed of engine sound and tyre noise. These two types of sounds contribute differently to the overall pass-by sound, depending on the vehicle model, operating conditions, and road materials, etc. For example, tyre noise would be the dominant noise source on an open highway, whilst the engine sound plays a more important role when a vehicle climbs a slope at low speed [259]. As there are substantial differences between the mechanism and audible impression on engine sound and tyre noise, these two types of sounds are simulated individually and combined in this thesis, which is congruent with the methodology in some other road traffic auralisation studies [112, 117, 119, 252, 260], as discussed in Chapter 3.

### 6.1.1 Engine Sound Synthesis

A vehicle typically has either a gasoline engine or a diesel engine. In theory, these two types of engines are both *internal combustion engines* which are designed to convert the chemical energy in fuel into mechanical energy. Sound emitted from an internal combustion engine is mainly composed of combustion noise, piston slap noise, gear noise, valve knock, and fuel pump noise, etc. [261]. As discussed in Chapters 3 and 5, synthetic sounds can be used in environmental sound studies to fulfil the requirement to vary the control of different sound parameters. In this thesis, procedural audio techniques are used for engine sound synthesis, by which the model can be highly dynamic, flexible, and interactive.

From an auralisation perspective, two main methods can be considered for modelling the time-varying engine sound. The first type is recording-based, analysis-synthesis methods. A series of recordings should be taken with controlled parameters such as engine speed, engine load, and engine order, etc. The analysis part breaks up the recording into constituent pieces which store the information of the engine sound under different conditions. These elements are concatenated accordingly in the synthesis part via signal processing techniques such as looping, crossfading, and pitch-shifting, etc., to achieve a range of variations. For instance, an enhanced pitch-synchronous overlap-and-add (PSOLA) algorithm is used in the granular synthesis model for engine sound synthesis proposed by Jagla et al. [262]. Although the results have been reported being realistic in some video games [164] and driving simulation applications [263], the most obvious drawback of these recording-based methods is inflexibility. The possible variations depend primarily on the range of recorded source signals, and it can be difficult or time-consuming to obtain all the required recordings for the full range of possible variation. Although there have been studies on interpolating and extrapolating between the recordings using signal processing methods (e.g. Spectral Modelling Synthesis as used in

[117]) to enhance the flexibility, there are compromises in other aspects, such as real-time performance, or audible artefacts under some circumstances.

The second type of engine sound modelling method is the so-called physical-based model, which abstract mathematical models from the engine running mechanism and synthesize engine sounds by simulating the behaviour of the relevant components in an engine at different engine running conditions. As no recordings are involved in the physical-based models, these models are fully data-driven, which means any recording-related restrictions do not constrain flexibility. If a physical-based model can be run in real-time, it coincides with the concept of procedural audio discussed in Chapter 4, which is of high interest in this thesis regarding the potential benefits of procedural audio for auralisation of urban soundscapes.

To implement an auralisation framework for road traffic noise using procedural audio, a physical-based model for engine sound synthesis is developed in this chapter. This model is based on the physical-based modelling method proposed by Baldan et al. [264], which is a modified version of Farnell's procedural audio model for car sound synthesis [13] offering an improvement of timbre at high engine speeds with higher-order harmonics. Before describing the physical-based model developed here in detail, it is useful to have an understanding of how an internal combustion engine works and how engine sounds are produced.

### 1) Engine Sound Mechanism

Figure 6.2 demonstrates the principle of a four-stroke engine which is widely used for typical gasoline engines. There are four distinct strokes in a complete operating cycle, which are called *intake*, *compression*, *power*, and *exhaust*, respectively [265].

In the intake stroke, a mixture of air and fuel is absorbed into the combustion chamber, with the intake valve open. The piston moves downwards within the cylinder until it reaches the bottom position (bottom dead centre, BDC).

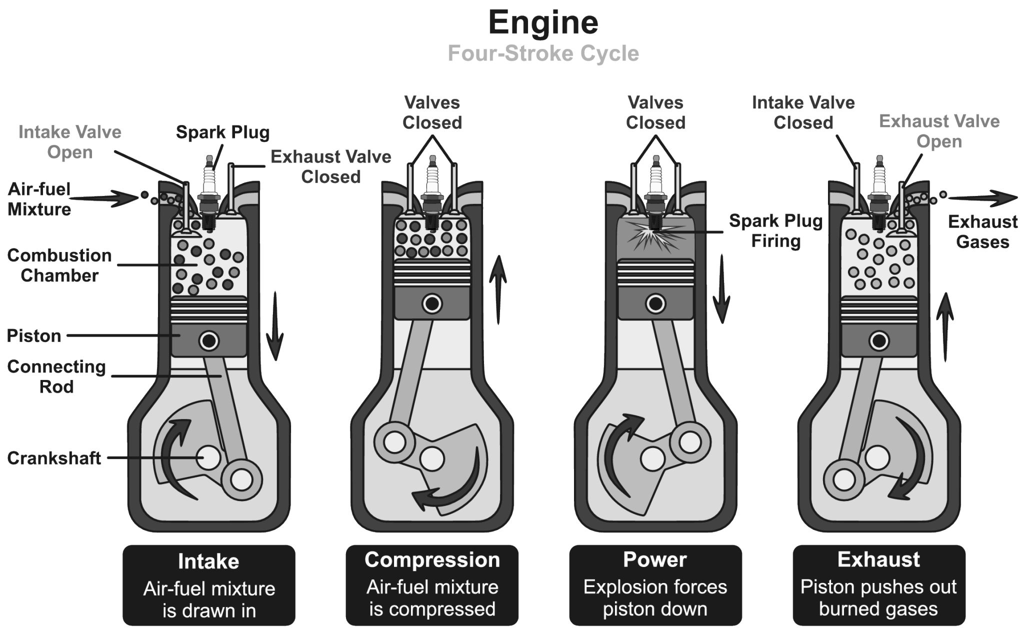


**Figure 6.2:** *Principle of a four-stroke engine, reproduced from [19].*

The intake valve closes gradually whilst the moving direction of the piston is changing, and the combustion chamber is fully sealed at the beginning of the compression stroke. In the compression stroke, the mixture of air and fuel is squeezed in the combustion chamber by the upward moving piston. According to the gas laws, the temperature inside the combustion chamber will increase until the piston moves to the top position (top dead centre, TDC). The ignition occurs with the piston at TDC, which is the beginning of the power stroke. This combustion is a rapid process and energy releases in the form of heat, which is then transferred into the mechanical energy of the piston, forcing it to move downwards quickly. The reciprocating motion of the piston is then converted into rotary motion of the crankshaft by the connecting rod – which is how energy is transferred into rotating the wheels. The exhaust valve will open the second time the piston reaches BDC, and gases are expelled from the combustion chamber and released to the environment via the exhaust system in the exhaust stroke. The piston has two complete back-and-forth movements within the cylinder, and the crankshaft rotates two revolutions (720 degrees) within one complete four-stroke cycle. The intake valve and the exhaust valve only open for one-fourth of the complete four-stroke cycle, on every other cycle of the crankshaft. This four-stroke cycle keeps going as long as the air and fuel are supplied.

Following the basic assumption in some engine sound synthesis models [13, 117, 266], the compression stroke and the power stroke do not significantly contribute to the characteristics of the engine sound since both the intake valve and the exhaust valve are closed. The combustion chamber itself is made of very thick and dense material which works efficiently for sound insulation [13]. Therefore, the acoustical characteristics of an engine are mainly represented by the sound produced by the intake stroke and the exhaust stroke, while the vibration of the engine block caused by the pistons' movement and fuel ignition contributes a 'dull thud' effect to the overall timbre. The exhaust gas plays a more important role in the engine sound properties because it is expelled to the atmosphere at high pressures. Furthermore, the exhaust gas also causes resonance inside the pipes of the exhaust system, which is simulated by a delay line with feedback loops in Farnell's model [13]. In their research, Baldan et al. [264] improved the method of simulating resonance effects by implementing digital waveguides, which are used to approximate the interactions occurring inside real pipes.

### 2) Modelling of valve and piston motions

According to relevant traffic noise auralisation studies [117, 252, 262], the sound pressure emission signal of the engine sound $s_e(t)$ can be assumed to consist of a tonal part and a stochastic part:

$$s_e(t) = s_{e_{tonal}}(t) + s_{e_{stocastic}}(t) \tag{6.1}$$

In some studies [117, 267], Spectral Modelling Synthesis (SMS) is used, which in theory can provide full control over the influencing signal parameters according to Fourier Series. However, the results from these studies show that a large number of sinusoidal components are required for capturing the rich tonal content of the engine sounds at high frequencies. This is not compatible with the idea of procedural audio here in terms of real-time performance according to the discussion in Chapter 4. On the other hand, some physically informed modelling techniques have been implemented, which potentially provide intuitive control of the sound synthesis process as well as real-time performance for procedural audio, such as the models already mentioned, as developed by Farnell [13] and Baldan [264]. In this thesis, a physically informed engine sound synthesis model is proposed, which is a modified version of Farnell's and Baldan's related work, with real-time performance suitable for procedural audio implementation.

The block diagram of the proposed engine sound synthesis model is shown in Figure 6.3. The engine speed $N(t)$ refers to the revolution speed of the crankshaft, which is often represented in revolutions per minute (RPM). The engine load refers to the external restraining torque being applied to the engine, which is often represented in torque ($Nm$). In this model, the engine load is normalised and described in percentage $M(t)\%$. $M(t) = 100\%$ means the engine is at full load, while $M(t) = 0$ means the engine is idling. The crankshaft revolution patterns can be described by these two parameters, in the form of a variable sawtooth wave. Each ramp of the sawtooth wave corresponds to two full revolutions of the crankshaft, according to the mechanism of a four-stroke engine. The frequency of the sawtooth wave is variable according to the instantaneous engine speed and engine load with a low-pass filter simulating the inertia effect to avoid the sound changing too quickly when the engine status changes [13].

Figure 6.4 illustrates how the main components are arranged and connected inside a gasoline engine with overhead camshaft configuration. Most engines have multiple cylinders working synchronously. The crankshaft is connected to pistons by rods. While the engine is running, these pistons move back and forth within cylinders, propelled by the energy generated from the combustion of the mixture of air and fuel in the combustion chamber. The

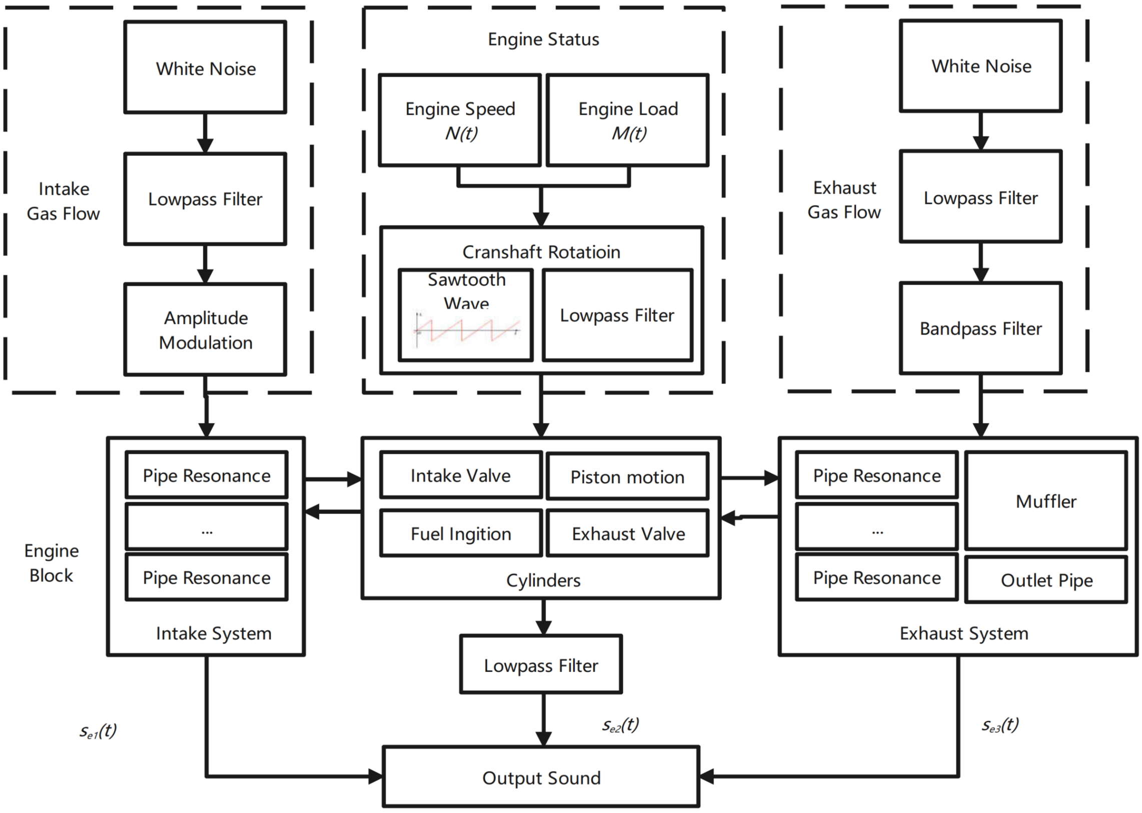


**Figure 6.3:** *Block diagram of the proposed physical-based engine sound synthesis model.*

cylinders are distributed along the crankshaft either in a *flat-plane* configuration or a *cross-plane* configuration. A flat-plane crankshaft has an arrangement of 0 or 180 degrees between crank throws as shown in Figure 6.5(a), while a cross-plane crankshaft has 90 degrees rotation between the neighbouring crank throws as shown in Figure 6.5(b). Both of these configurations can be equally shifted in phase and uniformly distribute power to the crankshaft. From an acoustic perspective, the most significant influence of different crankshaft design is the firing order [268]. In simpler words, for the flat-plane crankshaft design, every time the crankshaft rotates by 180 degrees, one of the cylinders fires, while for the cross-plane crankshaft design, for every 90 degrees that the crank turns, a cylinder fires. Considering the mechanical properties, vehicle performance, and the preferences of consumers, the cross-plane crank with even firing pattern design is mainly used in engines with multiples of 8 cylinders, whilst the flat-plane crank configuration is mostly used in engines with

4 or 6 cylinders.

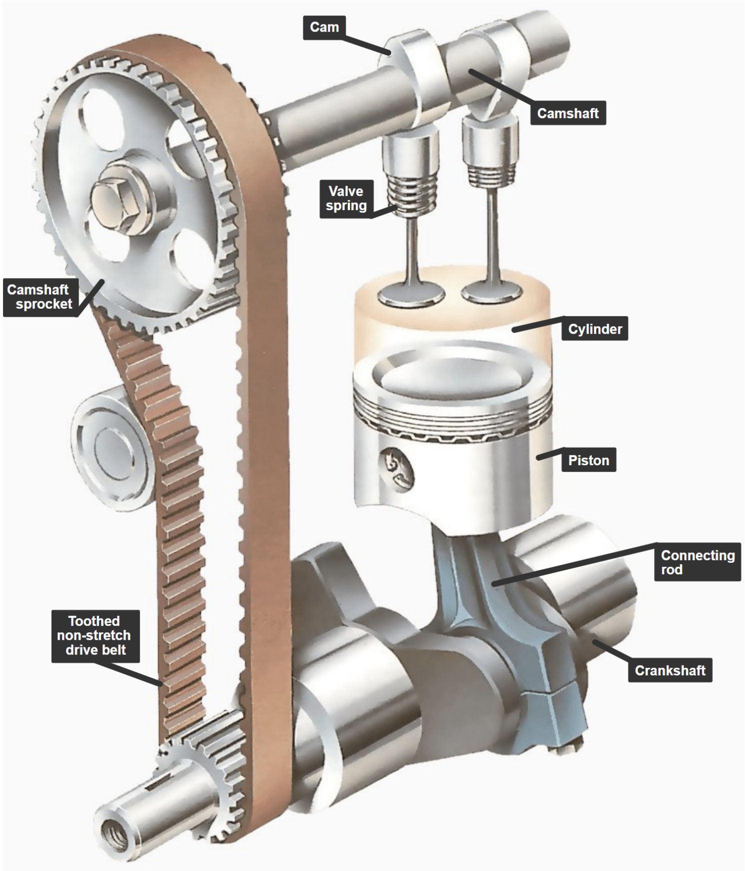


**Figure 6.4:** *Layout of the main components inside a gasoline engine with overhead camshaft configuration, reproduced from [20].*

The motion of the intake valve and the exhaust valve is also controlled by the crankshaft revolution, which is done by the camshaft appropriately connected to the crankshaft and the cams next to the pistons so that these valves can open and close at the correct times concerning the motion of the crankshaft and the pistons, as shown in Figure 6.4. In the procedural audio model proposed by Baldan [264], the motion of these valves are represented by the positive half of a sine wave that corresponds to a quarter of the engine operating phase cycle. It means that the opening/closing time for each valve is fixed when the engine speed is constant. During recent years, the *variable valve timing (VVT)* technique has been widely used in a lot of gasoline engines [269]. From an acoustic point of view, VVT can modify the timbre of the engine sound as the valve opening/closing time is variable according to the instantaneous

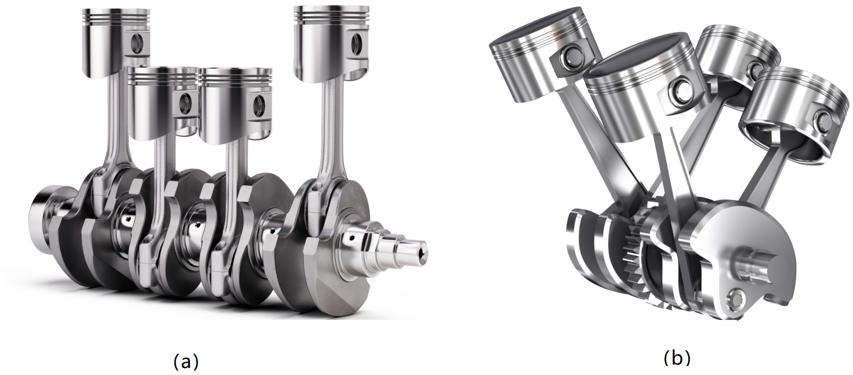


**Figure 6.5:** *(a) Flat-plane crankshaft configuration vs. (b) Cross-plane crankshaft configuration, reproduced from [21, 22].*

engine speed rather than fixed to the phase angle of the crankshaft.

In this thesis, an improved procedural audio model for engine sound synthesis including VVT effects is described as follows:

- The motion of the intake valve:

$$i(x) = \begin{cases} L_i(N)\sin(4\pi x + \phi_i(N)) & 0 < x < \frac{1}{4} \\ 0 & otherwise \end{cases} \tag{6.2}$$

where:

$x$ is the engine phase during one complete four-stroke cycle. Here one complete four-stroke cycle means two full revolutions of the crankshaft, which corresponds to a ramp in the sawtooth wave in Figure 6.3. $0 < x < \frac{1}{4}$ refers to the first quarter of the four-stroke cycle (or first half-revolution of the crankshaft).

$L_i(N)$ is a scaling factor representing the variable lift of the intake valve for VVT.

$\phi_i(N)$ is a phase-shifting factor representing the variable phase of VVT.

At high engine speed, a larger $L_i$ value means that the lift quickens air intake and exhaust, while at lower speeds such lifts decrease to a smaller value of $L_i$, degrading the mixing process of fuel and air. Basically, the phase-shifting factor $\phi_i(N)$ varies the valve time by shifting the phase angle of the camshafts connected to the crankshaft. For example, in some gasoline engines, the camshaft can be rotated in advance by 25 degrees for the intake valve to enable earlier intake for better engine performance, which happens between 1500 and 2000 RPM and over 5000 RPM of the crankshaft [270].

- The motion of the exhaust valve:

$$e(x) = \begin{cases} -L_e(N)\text{sin}(4\pi x + \phi_e(N)) & \frac{3}{4} < x < 1 \\ 0 & otherwise \end{cases} \tag{6.3}$$

  The expression is similar to the intake valve, except that the motion is shifted to the last quarter of the four-stroke cycle.

- The periodic motion of piston:

$$p(x) = A_{pis}\text{cos}(4\pi x), 0 < x < 1 \tag{6.4}$$

  where $A_{pis}$ is the maximum distance from its equilibrium position.

  As the piston is connected to the crankshaft by the connecting rod, its periodic motion can be represented by a cosine wave which has two complete back-and-forth movements within one complete four-stroke cycle.

- Fuel ignition: When combustion occurs, there is a prompt increase of pressure inside the combustion chamber, which changes the moving pattern of the piston represented as the positive half of a sine wave, shifted at the beginning of the expansion phase and scaled by a parameter $t$, which represents the time (relative to the full engine cycle) needed by the fuel to explode[264].

In this thesis, there is an extra consideration of the *ignition timing* which refers to the timing of the release of a spark in the combustion chamber relative to the current piston position and crankshaft angle. Ignition timing is an essential operating parameter that affects spark ignition engine performance and efficiency [271]. In some other procedural audio models [13, 264], it is assumed that the fuel ignition happens just before the top dead centre (TDC). In fact, for most modern engines, the ignition timing is usually variable and can be tuned in real-time by an electronic control unit based on the instantaneous engine speed and engine load [272]. From an acoustical point of view, the variation of ignition timing has an impact on the motion of the piston by an additional displacement, modifiying the timbre of the engine to some extent. The additional displacement of the piston can be expressed as Equation 6.5, which is caused by the combustion within the combustion chamber because of a sudden increment in pressure for each four-stroke cycle :

$$s(x) = \begin{cases} A_{pis}\text{sin}(2\pi(xt + \phi_{ign}(N, M)) & 0 < x < t \\ 0 & otherwise \end{cases} \tag{6.5}$$

where $\phi_{ign}(N, M)$ is the phase shift due to the variation of ignition timing, which is mainly determined by the instantaneous engine speed $N(t)$ and the engine load $M(t)$, as well as some other factors such as the the fuel composition, the type of fuel injectors, and the type/condition of the spark plugs, etc. [271, 272].

### 3) Modelling of intake system and exhaust system

From an acoustic perspective, both the intake system and the exhaust system are composed of a set of pipes connected to the cylinders, in which resonances occur when the air and the gas flow inside the pipes. Following the methodology in Baldan's model [264], these pipe resonances are modelled using digital waveguide modelling, which is a physical modelling synthesis technique made up of delay lines, digital filters, often with non-linear elements, e.g. frequency-dependent losses [273]. Digital waveguide modelling is usually used for synthesising string or wind instruments, such as the plucked string discussed in Chapter 4. Here a 1-D digital waveguide model is used to simulate the aerodynamic interaction within the intake system and the exhaust system. Two delay lines compose the basic unit of the 1-D digital waveguide with different gains, wherein the output of each line is fed into the input of the other, as shown in Figure 6.6. The forward input signal and the forward output signal are denoted as $x_0[n]$, and $y_0[n]$ in discrete time, respectively, with $k$ samples delay. The discrete backward input signal and the backward output signal are denoted as $x_1[n]$, and $y_1[n]$, respectively, with $k$ samples delay.

Constructive and destructive interference occurs due to the original signal interacting with the delayed waves, which represent the resonant modes of the modelled pipe. This way, it is convenient to approximate the behaviour of the sound waves in a real pipe: a sound wave enters from one side of the pipe, propagates through the pipe for a short period of time to the end, with some energy reflected and some dissipated during the process. The gain factors of the forward and backward delay lines are denoted as $g_0$ and $g_1$, respectively. A positive gain factor preserves the phase of the signal, whereas a negative gain factor inverts the phase of the signal.

A set of digital waveguides can be connected, either in series or in parallel, to simulate complex acoustic properties of the intake system and the exhaust system composed of multiple pipes in different configurations. The parallel connection is achieved by summing all the parallel outputs into a single receiving input, or equally dividing a single output across all the receiving inputs. To obtain series connections, the forward output of the first waveguide should be fed into the forward input of the second waveguide, the backward output of the second waveguide should be fed into the backward input of the first waveg-

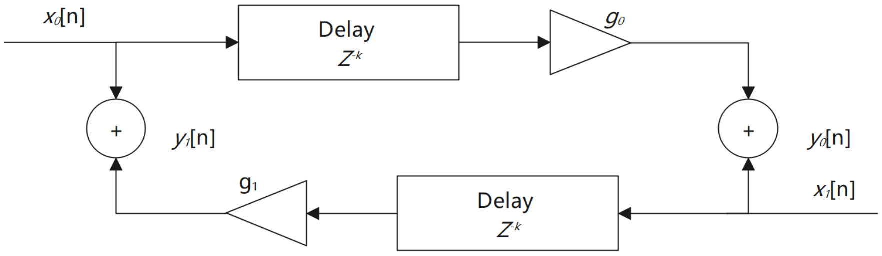


**Figure 6.6:** *The basic unit of the 1-D digital waveguide with different gains. $x_0$[n] is the forward input signal; $g_0$ is the forward gain factor; $y_0$[n] is the forward output signal; $x_1$[n] is the backward input signal; $g_1$ is the backward gain factor; $y_1$[n] is the backward output signal.*

uide. Both the intake system and the exhaust system can be modelled by a set of digital waveguides connected in series or parallel, representing different configuration patterns of the pipes. Figure 6.7 illustrates an example of such a model for pipes connected in series. Each pipe segment is represented by a digital waveguide. The varied cross-sectional areas of the pipes $A_1$, $A_2$, ..., $A_n$ are modelled by different gain factors, which represent the changes of acoustic impedance for a sound wave travelling through different pipe segments. The varied lengths of these pipes $L_1$, $L_2$, ..., $L_n$ are modelled by different delays in samples, which represent the delay for a sound wave travelling through different pipe segments. Following the methodology in the study [264], fixed feedback values (-0.5 for intakes, 0.1 for extractors) are used for the free end, while variable feedback values are set on the valve end, according to the modulation caused by the corresponding motion of the intake valve and the exhaust valve.

The exhaust gas from each combustion chamber is collected into a straight exhaust pipe in parallel and then carried to the muffler installed along the exhaust pipe before it is expelled into the environment. A muffler is a designed structure for reducing the amount of noise at certain frequencies emitted by a vehicle. In Baldan's model [264, 274], a muffler is modelled by four independent, partially reflecting waveguides, in which zero reflection provides no

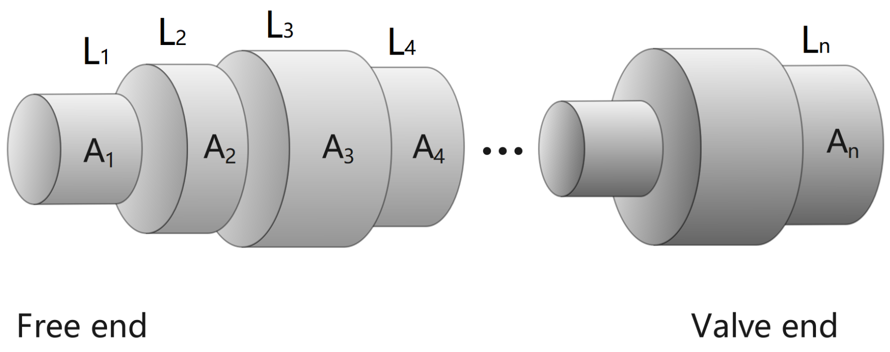


**Figure 6.7:** *Illustration of the intake/exhaust system by pipes connected in series, which can be modelled by digital waveguide elements. The varied cross-sectional areas of the pipes $A_1$, $A_2$, ..., $A_n$ are modelled by different gain factors. The varied lengths of the pipes $L_1$, $L_2$, ..., $L_n$ are modelled by different delays in samples. Fixed feedback values are used for the free end, while variable feedback values are set on the valve end.*

silencing, while a feedback factor of one represents a perfectly silent muffler. This method results in the effect of Helmholtz resonance, which reflects the sound waves at certain frequencies back towards the source and prevents sound from being transmitted along the pipe. For each digital waveguide unit, the delay lines are set so that it has a unique formant (sharp peak) in its frequency response.

In this thesis, sound absorption by the muffler perforation is also considered in addition to the Helmholtz resonances, because perforated pipes have been widely used in modern engine design to improve the wide-band performance of the muffler [275]. The extra sound absorption caused by the perforations is modelled by a series of low-pass filters connected in parallel with the digital waveguides. Yasuda et al. [276] proposed an acoustic-electronic analogy approach, which is used in this thesis to estimate the cutoff frequencies of these low-pass filters to simulate the sound absorption caused by the perforated muffler.

### 4) Physical-based modelling synthesis of engine sound

The 'voice' of an engine is mainly determined by the exhaust gas expelled from the outlet connected to the muffler, modulated by the motion of the exhaust valve. Apart from the exhaust sound, the intake sound and the mechanical vibration caused by the pistons' motion and fuel ignition also contribute to the timbre of the engine sound.

A synthetic engine sound can be created following the flowchart in Figure 6.3. The intake system is fed with low-pass and band-pass filtered white noise, which represents the air/fuel mixture breathing into the intake system. A time-varying low-pass filter and time-varying band-pass filter are utilised to simulate the turbulence of the air/fuel mixture during the intake stroke because of the variation on the components and temperature of the mixture. The filtered noise signal is then fed into a set of digital waveguides, with fixed feedback on the free end and variable feedback on the intake valve end according to the modulation caused by the corresponding motion of the intake valve. The simulated intake noise can be represented by the time domain signal $s_{e1}(t)$ picked up from the free end.

The mechanical vibration is caused by the periodic motion of the pistons plus the transient displacement caused by fuel ignition. This simulated vibration is then fed into a low-pass filter representing the damping effect caused by the engine block, so as to get the time domain signal $s_{e2}(t)$ which represents the mechanical vibration sound.

The exhaust system is fed with another low-pass filtered and band-pass filtered white noise, which represents the exhaust gas expelled into the environment. The energy of the exhaust gas is much higher than that of the intake gas flow, which results in higher amplitudes of the filtered white noise. The filtered white noise is then fed into a set of digital waveguides representing the pipe resonances and the muffler effects, with fixed feedback on the free end and variable feedback on the exhaust valve end according to the modulation caused by the corresponding motion of the exhaust valve. The exhaust noise is then represented by the time domain signal $s_{e3}(t)$ picked up from the free end.

The synthetic engine sound signal in the time domain can be obtained by the weighted sum of the simulated intake noise, mechanical vibration, and exhaust noise, expressed in equation:

$$s_e(t) = A_{e1}s_{e1}(t) + A_{e2}s_{e2}(t) + A_{e3}s_{e3}(t) \tag{6.6}$$

where $A_{e1}$, $A_{e2}$, and $A_{e3}$ are positive weighting parameters for the intake sound, mechanical sound, and the exhaust sound, respectively. These weighting parameters should be tuned according to the listening position and the specific structure and configuration of the engine. For example, for pass-by vehicles, exhaust gas plays the most critical role in the engine sound properties because it is expelled to the atmosphere at high pressures so $A_{e3}$ should be set with a relatively high value, while the mechanical vibration can be heard more obviously in-cabin as a relatively large portion of energy is transmitted via the chassis of the vehicle [277], so a greater $A_{e2}$ should be considered.

### 6.1.2 Tyre noise synthesis

In some vehicle pass-by auralisation studies, tyre noise is synthesised based on a physical model [118, 278] or a database of roadside recordings [279, 280]. These methods, however, are not suitable for the procedural audio implementation in the proposed auralisation framework because they are either too complicated and computationally intensive for real-time performance, or use recordings that are not congruent with the concept of procedural audio. In the study [119], an engineering model is developed for the European Harmonoise project (described as the 'Harmonoise tyre noise model') which can be used to calculate the time-averaged SPLs from road traffic noise. The SPLs of road/tyre noise close to the tyre are given as functions of frequency, vehicle speed, and vehicle category. As discussed in Chapter 3, there are five main categories of vehicles defined in the Harmonoise tyre noise model (light vehicles, medium-sized vehicles, heavy vehicles, other heavy vehicles, and two wheelers). Compared to some physical models of tyre noise (e.g.SPERoN model [118, 278]) which are computationally intensive in terms of calculating the contact patterns for the road-rubber system, the Harmonoise model can be implemented by a set of one-third octave-band filters to approximate the spectra of tyre noise. In this way, it is feasible to create an audible sense of tyre noise for vehicle pass-bys in real-time. As no recordings are involved during the whole process, this engineering model fulfils the requirement of procedural audio in terms of 'not recorded, not pre-sequenced sounds'. Therefore, the Harmonoise model is utilised for tyre noise synthesis in this thesis.

The frequency dependent relationship between the vehicle speed and the SPLs of the corresponding one-third octave bands can be established following Equation 3.27 introduced in Chapter 6. The regression parameters $a_t$ and $b_t$ presented in the Harmonoise model [119] are used, which are derived by the measured data from a series of vehicles. On top of the Equation 3.27, the following additional corrections are considered according to the methods proposed in [117, 119]:

- Road surface correction $\Delta L_{road}$: In the original Harmonoise model, $\Delta L_{road}$ is considered as a simple frequency independent correction item derived from two different surface materials [119]. A more recently published frequency dependent correction method in octave bands is proposed in the European CNOSSOS-EU project [281], in the form of:

$$\Delta L_{road}(v) = \alpha_{road} + \beta_{road} log \left( \frac{v}{v_{ref}} \right) \tag{6.7}$$

  where $\alpha_{road}$ and $\beta_{road}$ are regression parameters calculated from 15 different road surface types, $v$ is the constant vehicle speed, and $v_{ref}$ is the reference speed set as 70km/h. This more detailed road surface correction method has been implemented in the vehicle pass-by auralisation model proposed by Pieren et al. [117]. It can also be implemented in a procedural audio context for the proposed auralisation model in this thesis as this additional additive correction term in octave bands does not increase the overall complexity of the algorithm for tyre noise synthesis based on one-third octave bands [282]. In other words, this octave-band correction method has little impact on the real-time performance for the Harmonoise model for tyre noise synthesis.

- Directivity correction $\Delta L_D$: In the original Harmonoise model, directivity correction is considered for the vertical plane and the horizontal plane, respectively, as a function of frequency and angles, expressed as:

  $$\Delta L_D(f, \varphi, \psi) = \Delta L_H(f, \varphi) + \Delta L_V(f, \psi) \tag{6.8}$$

  where $f$ is frequency in Hz, $\varphi$ is the angle between the source and the receiver in the horizontal plane, $\psi$ is the angle between the source and the receiver in the vertical plane. The vertical directivity represents the sound insulation effect by the vehicle body, while the horizontal directivity simulates the horn effect of the rolling tyre [283].

  The vertical directivity correction is presented in a series of empirical equations in one-third octave bands, for the source heights 0.01m, 0.3m, and 0.75m, respectively. It is assumed that the vertical directivity is the same for all horizontal angles for a vehicle pass-by scene [119]. For horizontal directivity correction, some studies [119, 217] have revealed that the empirical equations for different vehicle categories, or for different brands/models in the same category, may be different from each other. As it is infeasible to measure every pass-by vehicle in terms of the horizontal directivity correction parameters, in this thesis, the following empirical equation can be considered for all types of pass-by vehicles if directivity patterns are required. This empirical equation is derived from a series of measurements for the pass-by sounds of a mid-sized vehicle [119]:

$$\Delta L_H(f,\varphi) = \begin{cases} (-2.5 + 4\,|cos(\varphi)|\,\sqrt{cos(\psi)}, & 800Hz < f < 6300Hz \\ 0, & otherwise \end{cases} \tag{6.9}$$

However, in this thesis, this directivity correction has not been implemented in practice in the plausibility evaluation stage, because it is claimed that the empirical equation still needs to be further validated with more measurements [119].

Some other correction terms, such as the wetness, age, and temperature of the road, are not considered as these are assumed not as important as the road surface patterns for the audible perception of the tyre noise [75]. Acceleration/deceleration is not considered either because the correction method in the Harmonoise model is relatively imprecise and is required to be further verified [119].

The final near field SPL of tyre rolling noise, with additional corrections for road surface and directivity patterns, used for tyre noise synthesis in this thesis can be expressed as:

$$L_{rolling} = L_t(f) + \Delta L_{road}(v) + L_D(f,\varphi,\psi) \tag{6.10}$$

$$= a_t(f) + b_t(f)log(\frac{v}{v_{ref}}) + \alpha_{road} + \beta_{road}log(\frac{v}{v_{ref}}) \quad + \Delta L_H(f,\varphi) + \Delta L_V(f,\psi) \tag{6.11}$$

The synthetic tyre rolling noise signal in the time domain can be obtained by summing the simulated signal in each one-third octave band, expressed as the equation:

$$s_t(t) = \sum_{k=1}^{N} p_{ref} 10^{\frac{L_{rolling}(v,\varphi,\psi)}{20}} \cdot s_{noise,k}(t) \tag{6.12}$$

where:

$N$ is the number of one-third octave bands.

$p_{ref} = 2\times\ 10^{-5}$Pa is the reference pressure

$s_{noise,k}(t)$ is the white noise filtered with the one-third octave bands filter in the $k^{th}$ frequency bin.

### 6.1.3 Sound source modelling for a vehicle pass-by

Following the steps in Section 6.1.1 and 6.1.2, it is possible to generate the synthetic engine sound signal and the tyre rolling noise signal in the time domain. No recordings are involved during these synthesis processes. According to the relevant studies in road traffic noise, [119, 257, 258], these two types of sounds contribute differently to the overall pass-by noise depending on the vehicle models, operating conditions, and road materials, etc.

The next step for creating the sound source for a vehicle pass-by is to combine the synthetic engine sound and the tyre rolling noise as a weighted sum. As in the Harmonoise model, the sound source of a vehicle pass-by is represented by two point sources at different heights, with an 80%/20% sound power distribution between them. The pass-by sound signals can be expressed as:

$$s_{1v}(t) = A_{1v}\cdot\frac{1}{\sqrt{5}}s_e(t) + B_{1v}\cdot\frac{2}{\sqrt{5}}s_t(t) \tag{6.13}$$

$$s_{2v}(t) = A_{2v}\cdot\frac{2}{\sqrt{5}}s_e(t) + B_{2v}\cdot\frac{1}{\sqrt{5}}s_t(t) \tag{6.14}$$

where $s_{1v}$ and $s_{2v}$ are sound signals for the lower and upper sources in the Harmonoise model, $\frac{1}{\sqrt{5}}$ is the normalisation factor coherent with the the 80%/20% energy distribution. $A_{1v}$, $A_{2v}$ and $B_{1v}$, $B_{2v}$ are weighting parameters for the engine sound and the tyre rolling noise, respectively. To get precise values for these weighting parameters, it is useful to take in-situ measurement of the vehicle pass-by sound, and then identify the sound energy emitted from the engine and the tyres separately for calculating the corresponding weighting parameters. Microphone array techniques as used for sound source localisation or identification can be used for this task, such as beamforming [284, 285], or nonlinear time mapping [286]. It is possible to get the spectra of noise emitted from the positions of the engine and the wheels for a vehicle pass-by with such measurements, and then use source-to-microphone transfer matrix methods to assign the corresponding strength at the heights of point sources in the Harmonoise model (0.01m, 0.3m/0.75m), expressed as the following equation:

$$E_m = C_{m,s} \cdot W_s \tag{6.15}$$

where $E_m$ is the SPL(dB) obtained at the $m^{th}$ microphone, $C_{m,s}$ is the transfer function which is determined by the layout of the transfer array, and $W_s$ is the sound power of the source.

However, it is still not an easy task to find appropriate values for these weighting parameters in this way. For instance, the weighing parameters $A_{1v}$, $A_{2v}$ and $B_{1v}$, $B_{2v}$ may be influenced by many factors, such as the vehicle speed, gear engagement status, the engine load, the acceleration/deceleration status, and the road surface, etc. It is difficult and too cumbersome to control all of these variables when taking acoustic measurements to obtain these weighting parameters. The weighting parameters acquired for one vehicle model may also not be suitable for another vehicle model, as they are designed by different manufacturers using different material, structures, and configurations for the engines and the tyres.

In fact, in most video games and VR applications, the engine/tyre sound balance is subjectively tuned by sound designers through formal or informal listening tests. The resulting realism is situated in a space between reality and fantasy: a ‘cinematic realism’ to fill the gap left by the lack of sensory information presented to the player [287]. Although results are not as physically accurate as those obtained from in-situ acoustic measurement, they can be both time-saving and convenient in implementation, and still achieve suitably plausible sounding results for game play or interactive applications in a virtual environment. Therefore, it is necessary to choose appropriate methods for balancing the engine sound and the tyre rolling noise for a vehicle pass-by according to the purpose of the required auralisation.

In this thesis, the level balance between engine and tyre rolling sound is conducted following the guidance of the Harmonoise model [119]. In this model, 80% of the sound power is assumed to emit from a source at a height 0.3/0.75m for light/heavy vehicles, respectively. 20% is assumed to be emitted from a low source, 0.01 m above the road surface. The sound power level of the engine sound is described by the empirical equation:

$$L_p(f) = a_p(f) + b_p(f) log(\frac{v - v_{ref}}{v_{ref}}) \tag{6.16}$$

where $f$ is frequency in Hz, $v$ is the speed of the vehicle, and the $v_{ref}$ is the reference speed set as 70km/h. The regression parameters $a_p$ and $b_p$ presented in the Harmonoise model [119] are used, which are derived by the measured data from a series of vehicles, presented in one-third octave bands. As vehicle speed is not a driving parameter for the procedural audio based engine sound synthesizer proposed in this thesis, the relationship between the engine speed and vehicle speed should be established. In this thesis, the relationship between the vehicle speed and the corresponding engine speed is calculated according to Equation 6.17 [268]:

$$n = 60 \cdot i_{gear} \cdot i_{axle} \cdot \frac{v}{3.6 \times 2\pi r_{tyre}} \tag{6.17}$$

in which:

$n$ is the engine speed in RPM.

$v$ is the speed of the vehicle in km/h.

$r_{tyre}$ is the tyre radius in m.

$i_{gear}$ is the gear ratio.

$i_{axle}$ is the axle ratio.

The synthetic engine noise signal in the time domain can be obtained by summing the simulated signal in each one-third octave band, expressed in the equation:

$$s_p(t) = \sum_{k=1}^{N} p_{ref} 10^{\frac{L_P(f)}{20}} \cdot s_{noise,k}(t) \tag{6.18}$$

where:

$N$ is the number of one-third octave bands.

$p_{ref} = 2\times10^{-5}$Pa is the reference pressure.

$s_{noise,k}(t)$ is the white noise filtered with the one-third octave bands filter in the $k^{th}$ frequency bin.

As s$_p(t)$ is a summation of band-pass filtered white noise signal, it is not perceived as an engine sound because of the lack of tonal part and rattling effects. However, the sound signal s$_p(t)$ created using the Harmonoise model provides useful cues for calibrating the sound levels of the synthetic engine sounds created by other approaches. To balance the sound levels between the an engine sound created by procedural audio and a tyre sound created by the Harmonoise model, the procedurally created engine sound can be normalised relative to the corresponding signal $s_p(t)$ generated by the Harmonoise model using power normalisation. This way, the power of the procedurally created engine sound is identical to the corresponding signal $s_p(t)$ generated by the Harmonoise model, which can then be further processed by the methodologies in the Harmonoise model, assigning the 80%/20% sound power distribution between the two point sources at different heights as defined in this model.

## 6.2 Microscopic Urban Sound Propagation Modelling

Following the methods introduced in Section 6.1, a sound synthesis model for vehicle pass-bys based on procedural audio can be built, which can be used to generate the required sound sources for auralisation of traffic noise. The next step for enhancing the realism of the auralisation is to simulate the sound propagation effects corresponding to the context of the specific acoustic environment, providing audible clues as to the nature of the space. This study focuses on the traffic noise caused by pass-by vehicles in a microscopic urban scene, which is one of the most common situations that pass-by sounds are perceived in our daily lives.

As discussed in Section 6.1, a 'microscopic' urban scene can be considered as the experience of being at a street or a square, where pass-by sounds from each individual vehicle can be identified and influences on every building surface should be considered. As these sources are moving at different speeds, the distance attenuation and Doppler shift must be considered separately for each source. Apart from distance attenuation and Doppler effects, sound propagation under a simulated urban scene can be influenced by the surrounding buildings, road surface, trees along the road, etc. If a sound source is relatively far from the listener, the air absorption effects might also be considered [75].

In this thesis, the distance attenuation, Doppler shift, and early reflections caused by ground and surrounding buildings are considered for simulation of sound propagation effects because it is assumed that they are important for the auralisation of vehicle pass-bys for a microscopic urban scene. This assumption is also congruent with some prior auralisation studies on traffic noise [117, 252, 260, 280]. Some other propagation effects, such as diffraction, late reverberation, vegetation effects, and air absorption are not considered because either they are too time-consuming or have relatively less impact on audible perception of the auralised sounds [75].

### 6.2.1 Distance Attenuation

According to the discussion of the inverse square law as presented in Chapter 2, the sound pressure $p$ is proportion to $1/r$ in the free field condition where a sound propagates equally in all directions. In the real world, reflective surfaces will modify the resulting sound field such that a direct application of the inverse square law no longer applies. In this thesis, distance attenuation is implemented based on Equation 6.19 which estimates the sound pressure field produced by a moving (mono) sound source at a constant speed $v$ [288]:

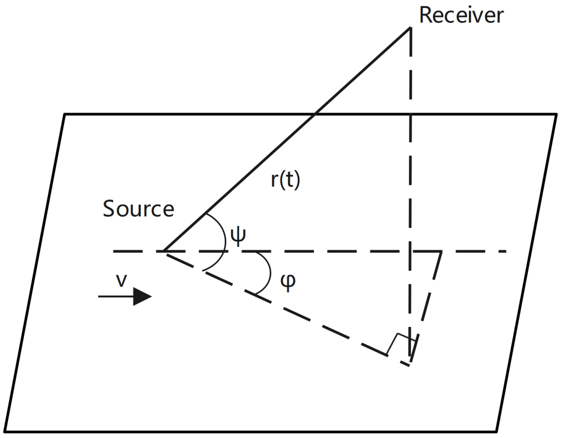


**Figure 6.8:** *llustration of the horizontal angle and vertical angle from a moving source to receiver. $\varphi$ is the angle between the source and the receiver in the horizontal plane; $\psi$ is the angle between the source and the receiver in the vertical plane; $v$ is the constant moving speed of a moving (mono) sound source.*

$$p(t) = \frac{1}{4\pi r(t)(1 - M_a \mathrm{cos}\varphi(t))^2} s(t - \frac{r(t)}{c}) \tag{6.19}$$

where:

$p(t)$ is the sound pressure at the listening position.

$r(t)$ is the distance between the source and the receiver

$M_a$ is the Mach number which is defined as the source moving speed over the speed of sound $M_a = v/c$.

$\varphi$ is the horizontal angle between the source moving direction and the source-receiver direction as illustrated in Figure 6.8.

$s(t)$ is the sound signal of the moving source, so that $s(t - r(t)/c)$ is the sound signal with $r(t)/c$ delay in the time domain.

### 6.2.2 Doppler Shift

As discussed in Chapter 2, when a sound source is moving relative to the listener, the perceived frequency will be shifted up or down compared to the emitted frequencies, according to the component of the source velocity relative to the listener. As the pitch shift in the frequency domain can be implemented by resampling the discretised sound signal with a variable delay line in the time domain [12], this technique can be used here to simulate the Doppler effect. Since the delay may not always be an integer value in samples, it is necessary to find a way to get the interpolated signal values corresponding to the fractional delays.

There are many schemes for signal interpolation. The most straightforward way is drawing a straight line between two neighbouring samples and returning the appropriate point along that line representing the interpolated value. This is called *linear interpolation*. For a sound signal travelling time $\Delta t$ from source to receiver, the fractional sample index of the signal $n_e$ of the signal $y[n_e]$ at the source time-axis can be expressed as:

$$n_e = n_r - \Delta t \cdot f_s \tag{6.20}$$

where $f_s$ is the sampling frequency in Hz, $n_r$ is the integer sample index of the signal at the receiver time-axis. Defining $\eta = n_e - \lfloor n_e \rfloor$, where $\lfloor n_e \rfloor$ is the floor function, then the interpolated value $y[n_e]$ can be calculated by:

$$y[n_e] = (1 - \eta) \cdot y[n_e] + \eta \cdot y[n_e + 1] \tag{6.21}$$

The most significant advantage of linear interpolation is the low computational complexity, and the simplicity for implementation. Nevertheless, linear interpolation is not suitable for all audio signals as it is not a bandlimited interpolation method. A non-bandlimited interpolation method will cause aliases and artefacts, which might be audible for some sounds [12]. Figure 6.9 illustrates an example of linear interpolation for the Doppler shift of a pure tone at 8kHz. As can be seen from the spectrogram, there is clear visible distortion due to aliasing, with relatively high energy compared to the pitch-shifted pure tone. This distortion is also audible as artefacts for the pure tone at 8kHz.

In order to reduce such artefacts, bandlimited interpolation is often used. According to Shannon's sampling theorem, a sinc filter is an ideal bandlimited interpolation kernel [12]. In practice, this filter should be windowed and truncated as an approximation of the ideal bandlimited interpolation. In [117], a hamming-windowed sinc interpolation is utilised. The values of the interpolation kernel is pre-calculated and stored in a look-up table to save computational effort at run time. In this thesis, the Lanczos-windowed sinc interpoation is studied in comparison with linear interpolation because it is claimed as the best compromise in terms of reduction of aliasing, sharpness, and minimal ringing compared to other windowed sinc filters [289]. The Lanczos kernel $K(z)$ is defined as:

$$K(z) = \begin{cases} \text{sinc}(z)\text{sin}(z/a), & -a < z < a \\ 0, & otherwise \end{cases} \tag{6.22}$$

where $z$ is a non-integer value representing the sample position, and $a$ is the truncated number of lobes representing the kernel size. The larger kernel size, the less distortion can be obtained with the compromise of more computational resources. For audio signal processing, $a = 2$ or $a = 3$ is often used, denoted as the *Lanczos2* or *Lanczos3* function. The interpolation is achieved by a convolution of the signal with the Lanczos kernel:

$$y(z_s) = \sum_{\lfloor z \rfloor - a + 1}^{\lfloor z \rfloor + a} x[n]K(z - n) \tag{6.23}$$

where $\lfloor z \rfloor$ means the floor function of $z$.

Figure 6.9 – Figure 6.11 demonstrate a comparison of Doppler shift implemented by linear interpolation and Lanczos-windowed sinc interpolation in

spectrograms. In this example, an 8kHz pure tone travelling at 70.0km/h along a straight line 100m away from the receiver is used to the test the Doppler shift implemented by different interpolation methods. Figure 6.9, Figure 6.10 and Figure 6.11 are the spectrogram of linear interpolation, Lanczos2-windowed sinc interpolation, and Lanczos3-windowed sinc interpolation, respectively. As can be seen from the figures, apart from the target pitch shift, there are other visible lines in these spectrograms. They are distortions caused by different interpolation methods, which may also result in audible artefacts. In general, a bandlimited interpolation like Lanczos-windowed sinc interpolation causes less distortion than linear distortion; the larger the kernel size, the fewer artefacts occur. However, the convolution process for the windowed sinc interpolation is inevitable, which makes it much more computationally expensive (complexity $O(n^2)$ without any fast convolution algorithms) even though the kernel can be substituted by a pre-calculated look-up table to save some computation time [117].

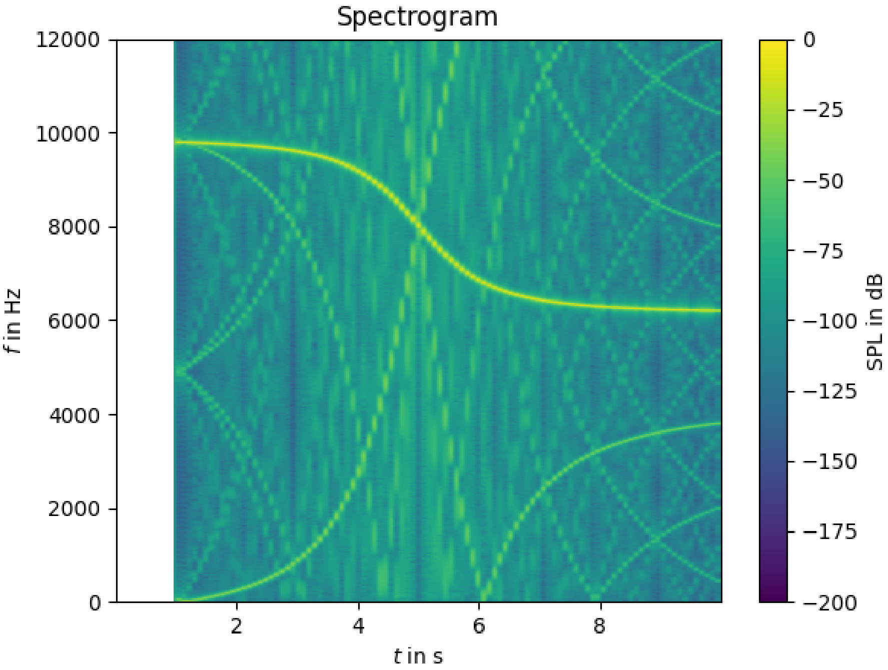


**Figure 6.9:** *Doppler shift implemented by a fractional delay line with linear interpolation. This figure shows a spectrogram of an 8kHz pure tone travelling at 70.0km/h along a straight line 100m away from the receiver.*

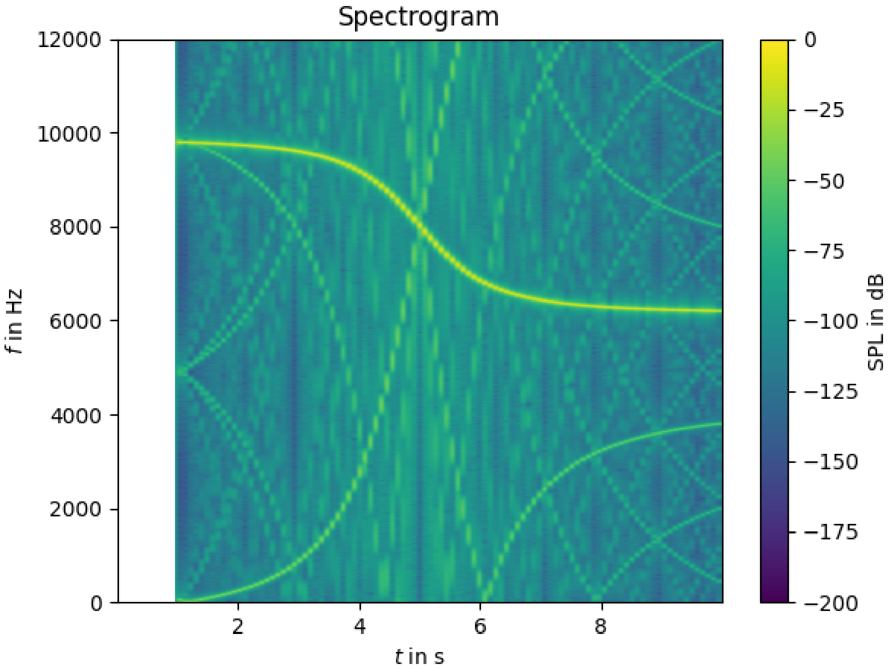


**Figure 6.10:** *Doppler shift implemented by a fractional delay line with Lanczos2 interpolation. This figure shows a spectrogram of an 8kHz pure tone travelling at 70.0km/h along a straight line 100m away from the receiver.*

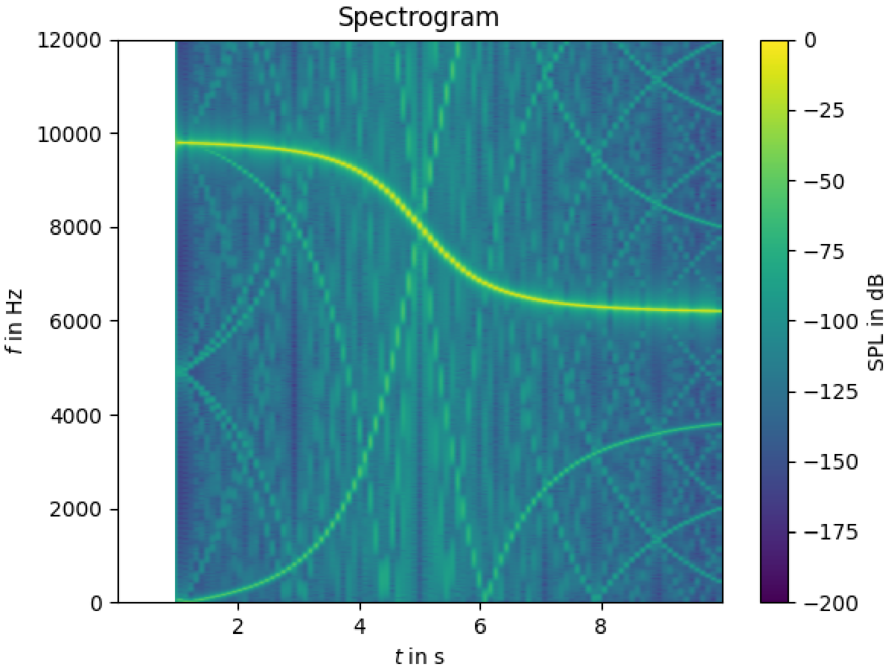


**Figure 6.11:** *Doppler shift implemented by a fractional delay line with Lanczos3 interpolation. This figure shows a spectrogram of an 8kHz pure tone travelling at 70.0km/h along a straight line 100m away from the receiver.*

It is noted that linear interpolation and sinc kernel bandlimited interpolation are two opposite 'extreme' interpolation methods for audio signal resampling in terms of quality and computational demand. There are some other interpolation methods standing at the 'middle' between these two 'extreme' methods. For example, cubic spline interpolation, which uses the information of four adjacent points to approximate the point to be interpolated, leading to a smoother transition between adjacent segments compared to linear interpolation, but being less computationally demanding compared to sinc kernel interpolation. A simulation of such a cubic spline interpolation algorithm is included in Appendix D. Although cubic spline interpolation has been reported to cause less distortion than linear interpolation [290], it is still not a bandlimited limited interpolation method and may lead to artefacts. On the other hand, cubic spline interpolation is more computational demanding compared to the linear interpolation method.

In this thesis, linear interpolation is the prioritised choice of Doppler shift implementation, although it causes more obvious artefacts. This is because of the simplicity and computational complexity ($O(n)$) of linear interpolation. As most of the pass-by sounds are perceived as broadband signals, potential artefacts are less pronounced when using linear interpolation [253, 291]. In case of some rare circumstances where the tonal part plays an important role (e.g. the engine roar for some 'sporty' cars), cubic spline interpolation or a windowed sinc interpolation with a small filter length such as the Lanczos2/Lanczos3 kernel can be considered with the potential compromise of non-real-time performance, if there are obvious audible artefacts caused by linear interpolation.

### 6.2.3 Early Reflections

An overview of sound propagation simulation techniques by geometrical acoustics and wave-based methods has been presented in Chapter 3. In a microscopic urban scene, there are multiple reflections for pass-by sounds, mainly from the ground and the facades of the surrounding buildings and walls. Compared to the acoustic environment in rooms, the sound field in an outdoor environment is more large-scaled, without much sense of reverberation, although strong reverberance may occur in some extreme cases, such as narrow street canyons [292], small squares or courtyards [123], etc. In this thesis, early reflections from the ground and surrounding buildings are of interest. This is because in reality, vehicle pass-by sounds are always coloured by the surrounding environment when they are heard. Even though the outdoor sound field is usually not as reverberant as the sound field of an indoor environment, it is assumed that early reflections still have impacts on the audible experience of such a sound scene.

To simulate the early reflections for a vehicle pass-by in a microscopic urban scene, an image source algorithm is implemented. As discussed in Chapter 3, image source methods are suitable for simulating early reflections within a space with simple geometry, and it is viable to implement the computation process running at interactive rates, allowing alteration in the acoustic model running out in real-time [72]. This potentiality fulfils the requirement of the real-time performance of the procedural audio model developed in this thesis. Two issues need to be considered here: 1) how many orders of image sources are appropriate for the perception of early reflections for the auralisation framework? and 2) is it possible to render such orders at interactive rates? In fact, how to define the transition of an impulse response from the early reflections to the late reverberation is still an open question. Different methodologies have been proposed to find the appropriate transition, depending on the different modelling methods used for the early and late reflections. For a micro-scopic urban environment, the maximum order of reflections simulated by image source methods used in different studies varies from $1^{st}$-order [72], $3^{rd}$-order [293], $5^{th}$-order [72], $8^{th}$-order [294] , to $10^{th}$-order [295], etc. As computation time goes up exponentially with the increase of image source order, after an overview of the computational time measurement data in [72], it is decided that in this study, only $1^{st}$-order reflections are considered with a calculation time of several milliseconds so as to meet the requirement for interactive rates more safely.

When both source and receiver are static, the $1^{st}$-order image sources have to be determined only once. If the source is moving, the source position at each frame is variable, as are the corresponding image source positions and the delay times. Thus, for auralisation of pass-by vehicles, each of the $1^{st}$-order image sources is implemented by a variable delay line, which is further scaled based on a time-varying factor according to the inverse square law, and filtered according to the sound absorption coefficients of the reflective surface. The delay is set according to the distance between source and receiver, and the speed of sound. In case a fractional delay is needed, a separate resampling process is performed. In order to simplify the model implementation, only reflections from near buildings and the ground are considered, and reflections from other components in the microscopic urban environment are ignored, e.g. trees, other vehicles, etc. Ground reflection is always audible, while the effectiveness of the image sources from building reflections should be judged by an audibility check [7].

Figure 6.12 illustrates an example of the audibility check for a moving source at different positions in the horizontal plane. In this figure, the road width is denoted as $d$. The lengths of the surrounding buildings (Building A, Building B) are denoted as $L_1$ and $L_2$, respectively. A source $S$ is moving from position $S_1$ to position $S_2$, and the receiver is marked as $R$. At the position $S_1$, the source is reflected in the facade of Building A, creating an image source $S_{1A}$. A line can be drawn between $S_{1A}$ and $R$. As the line intersects the facade of Building A, $S_{1A}$ should be considered as a valid image source, which means the reflection from this image source is audible. Repeat this process in terms of the reflection from Building B, and it is found that there is no intersection between the line and the facade of Building B. Therefore, $S_{1B}$ is not a valid image source. When the source moves to position $S_2$, then $S_{2A}$ becomes invalid

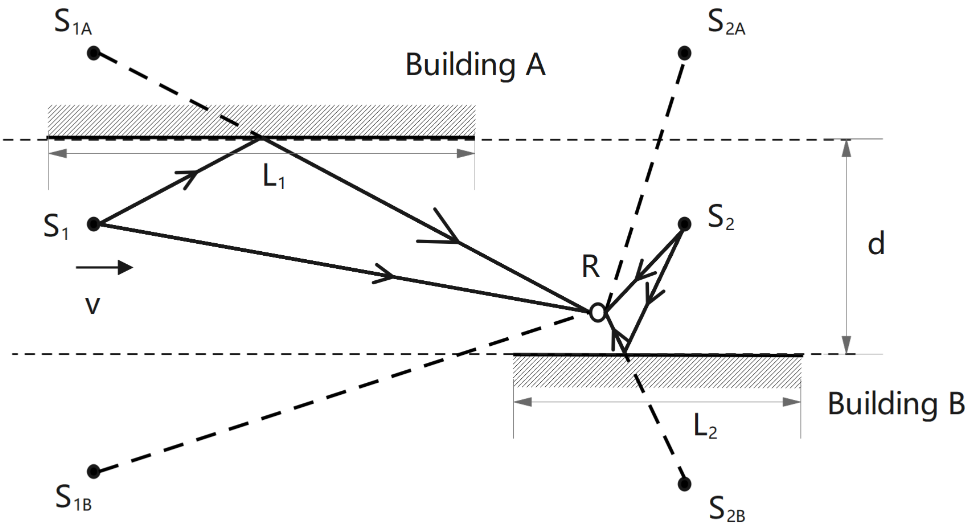


**Figure 6.12:** *Illustration of audibility check for a moving source in the horizontal plane. For the receiver position R, $S_{1A}$ is a valid image source for the source position $S_1$, but $S_{1B}$ is not a valid image source for the position. $S_{2A}$ is not a valid image source for the source position $S_2$, while $S_{2B}$ is a valid image source for the source position $S_2$.*

and $S_{2B}$ becomes valid.

So as to simulate the frequency-dependent sound absorption effect, a low order digital filter bank is designed to fit the absorption coefficient data [296]. The signal is converted into the frequency domain by Fourier transform, multiplied by the transfer functions of the filter bank, and then converted back into the time domain by inverse Fourier transform. The relationship between the sound absorption coefficient and the reflection coefficient is given by Equation 2.24. For the simulation of sound absorption, a $1^{st}$-order band-pass filter bank in two-octave-bands is applied to the signal in the frequency domain, for which the cutoff frequencies are set as 250Hz, 1000Hz, and 4000Hz, respectively.

## 6.3 Binaural Audio Rendering

In this thesis, the KEMAR HRTF database [92] is used for binaural audio rendering. This database consists of different HRIR data corresponding to azimuth angles in the range of 0 to 355 degrees and elevation angles in the range of -40 to 90 degrees, with a 5-degree spatial resolution. For each angle, two 512-sample impulses responses measured for the left and right ears are provided. As the sample rate for the HRIR measurement is 44.1kHz, a length of 512 samples corresponds to an audio file with 11.6ms duration. As discussed in Chapter 3, sound incident from a specific direction can be simulated by convolving a source $s(t)$ with a pair of HRIRs for that direction using Equation 3.26. In practice, this is implemented in the frequency domain as multiplication to save computation time.

Before conducting convolution, a low frequency correction for the KEMAR HRTF database is implemented, following the method in [297]. This is because the original HRTFs in the KEMAR database have been proved to be inaccurate below 400Hz [90]. According to relevant studies on vehicle pass-by noise [119, 286, 298, 299], low frequency traffic noise can be clearly perceived. Therefore, in order to avoid the potential loss of plausibility caused by the HRTF processing, low frequency correction as proposed in [297] is conducted. Specifically, the following steps have been taken in this thesis for the low frequency correction of the KEMAR data used for the HRTF processing of traffic noise auralisation:

1. Calculating the mean magnitude in the range of 100–345Hz in the original HRTF data (here 345Hz is the reference frequency used in [297], and can be changed to any frequency between 100–400Hz);

2. Replacing the magnitudes of 100–345Hz in the original HRTF data with the calculated mean value;

3. Calculating the phase information of the frequency at 345Hz;

4. Extrapolating the phase information in the range of 100–345Hz by linear interpolation according to the approximately linear relationship between the frequency and the phase within this frequency range.

The effectiveness of this low frequency correction method on the KEMAR

HRTF dataset in terms of the perception of low frequency sounds has been validated by Xie [297] using formal listening tests.

In order to create a smooth audible experience for a vehicle pass-by sound without a noticeable transition between different HRIRs, a crossfading approach between consecutive HRIRs can be taken, which has been used in many practical implementations of auditory virtual environments based on HRTFs [300].

## 6.4 Auralisation of Traffic Flow

A traffic flow consists of multiple vehicle pass-bys. In this thesis, *traffic flow* refers to the concept of microscopic traffic flow, which gives attention to the details of the driving behaviour and driving patterns of each single vehicle in a traffic flow, and the interactions taking place between different vehicles [301]. From the perspective of microscopic traffic flow, it is possible to analyse very small changes in a traffic stream over time and space. The properties of a microscopic traffic flow can be described by a series of parameters including speed ($v$), density ($k$), volume ($V$), and flow ($q$), etc. [301]. These parameters are defined as follows:

- Speed ($v$) is defined as the average speed of the traffic flow.
- Density ($k$) is defined as the number of vehicles per unit length of the road.
- Volume ($V$) is defined as the number of vehicles per unit time on the road.
- Flow ($q$) is defined as the number of vehicles passing a reference point per unit of time.

These parameters are used in this thesis to describe the traffic flow status to be auralised. Auralisation of traffic flow is achieved by summing all the auralised single vehicle pass-by sounds.

## 6.5 Implementation Platforms

A series of different platforms have been selected and combined for the implementation of the complete traffic flow auralisation framework proposed in this thesis, including Unity3D, Pure Data (PD), Wwise, and Matlab.

Unity3D is selected as the main platform for developing the auralisation framework. Unity3D [302] is a game engine which provides the main framework and functions for developing digital games. With the development of VR/AR, the use of this game engine has been extended into other industries such as architectural design [303], urban planning [304], and product design [305]. Virtual environments rendered by such a game engine are often highly interactive. This is congruent with the application of auralisation, which can be used for internal communication as part of the design process, as well as for demonstrating to, e.g. stakeholders and the wider public.

In order to render a sound scene (i.e. a vehicle pass-by) in an acoustic environment, it is necessary to render a virtual environment and assign acoustic properties to the objects within the environment, such as the sound absorption coefficients of each surface. A game object attached with 'Audiosource' features in Unity3D should be built, which represents the sound source emitting audio signals in the virtual environment. In order to hear the sound, the player should be assigned to the 'AudioListener' class in Unity3D.

Figure 6.13 shows a screenshot of the virtual environment developed for the vehicle pass-by auralisations in this thesis. The visual scene is rendered based on the open-source simulator AirSim [23]. In this figure, the acoustic environment of this microscopic urban scene is determined by the layout of the surrounding buildings and the acoustic properties of the building facades. The black car is a game object to which two audio sources are attached at heights of 0.01/0.3m above the ground according to the methodology of sound source modelling used in the Harmonoise model [119]. The player is in the first person view, registered as the 'AudioListener' in this virtual environment. Based on

**Figure 6.13:** *Visual rendering in Unity3D of a microscopic urban environment with vehicle pass-bys rendered, using the open-source simulator AirSim [23].*

the virtual environment rendered and game objects defined, the next step is to implement the procedural audio for the engine sound and the tyre sound, and assign them to the game object.

The procedural audio model for engine sound and tyre sound synthesis is implemented in Pure Data (PD) [306], which is a programming language developed for creating interactive audio content. The engine sound is synthesised according to the block diagram in Figure 6.3 by combining the intake sound, mechanical sound, and exhaust sound. Tyre sound is synthesised by implementing the Harmonoise model [119] based on Equation 6.10. Two-way parameter transmission between Unity3D and PD can be realised by using the open-source plug-in LibPd Unity Integration [307]. With this plug-in, on one hand, synthetic sounds from PD patches can be sent into Unity3D when the corresponding sound events are triggered; on the other hand, the data from Unity3D can be sent to PD , based on which the synthetic sounds may vary in real-time.

In terms of the simulation of sound propagation effects, the $1^{st}$-order image source is implemented in Wwise [308], an audio middleware for sound design, while the distance attenuation and Doppler shift are implemented in Matlab. There are several reasons for such a choice of different platforms for simulation of different sound propagation effects. Firstly, as Unity3D and Wwise are well integrated with each other, it is convenient to implement the $1^{st}$-order image source algorithm in Wwise for simulation of early reflections in a virtual environment in Unity3D running in real-time. This is congruent with the concept of procedural audio in terms of real-time performance. However, the approaches for distance attenuation and Doppler shift proposed for the auralisation framework cannot be fully implemented using either Unity3D or Wwise, because only some approximation methods, such as decay curves or real-time parameter control, can be used to provide an approximation of these propagation effects. Although there is a built-in function for Doppler shift within the Unity3D audio engine, the algorithm for its implementation is undocumented.

In order to fully implement the proposed sound propagation effects, including distance attenuation and Doppler shift, Matlab is used. The output sounds after being processed in Wwise for the implementation of a $1^{st}$-order image source algorithm are sent to Matlab for the calculation of distance attenuation based on Equation 6.19, and the Doppler shift using a fractional delay line with linear interpolation, as discussed in Section 6.2.

Note that the Matlab implementation will lead to an 'offline' process, which is different from the other steps which are 'online' in the auralisation framework. Here 'offline' means sounds cannot be created in real-time – there is significant latency between the input and the output, while 'online' means the algorithm has real-time performance. As plausibility is the main area to be investigated in this thesis rather than real-time performance, this 'offline' implementation is considered acceptable as long as all the proposed sound propagation simulation methods can be implemented appropriately.

The output sounds from Matlab after the process of distance attenuation and Doppler shift are sent back to Unity3D and assigned to the 'Audiosource' features for the game object. The final step is to binaurally render the auralised sounds using the HRTF processing discussed in Section 6.3. This step is implemented in Unity3D with the Audio Spatializer SDK [309], which is a built-in extension of the native audio engine for Unity3D.

In the Audio Spatializer SDK, the original KEMAR HRTF database [92] is utilised to spatialise each game object attached with the 'Audiosource' features in Unity3D. A built-in linear crossfading algorithm is implemented, which aims to create a smooth transition between the HRIRs in different angles for a moving source, although there might be 'zipper' artefacts if the source moves too fast. By conducting informal listening tests, the author of this thesis found that for a typical traffic flow scene (e.g. a listener stands several metres away from a road, and pass-by vehicles at 30–80km/h), there are no noticeable artefacts in transition between different HRIRs for the moving sources when using the built-in linear crossfading algorithm for the Audio Spatializer SDK. Therefore, this linear crossfading algorithm is implemented for the auralisation of vehicle pass-bys in the rest of this thesis. Some other crossfading methods may be considered if some noticeable 'zipper' artefacts occur in some special cases.

The original KEMAR HRTF database is substituted by a modified version with low frequency correction, for which the magnitudes and the phase information are rebuilt using the correction method proposed by Xie [297] as discussed in Section 6.3. The impact of this low frequency correction has been evaluated by the author of this thesis using informal listening tests, where it was found that the loss of low frequencies in the auralised vehicle pass-by sounds is less noticeable when this correction method applied. Therefore, this low frequency correction method may be considered as being suitable for the auralisation of traffic noise in which low frequency sounds should be clearly perceivable.

## 6.6 Summary

This chapter started with the development of an auralisation model for a single vehicle pass-by sound using procedural audio. After an overview of the relevant studies in road traffic noise auralisation and sound design using procedural audio approaches, it is found that engine sound synthesis can be achieved by using a physical-based synthesis model which is a modified version of the model proposed by Farnell [13] and the model proposed by Baldan [264, 274]. An extra consideration on the VVT effects has been implemented, which is considered as a more accurate approximation of recent gasoline engines. It is found that tyre noise synthesis can be implemented by an engineering model derived from the Harmonoise project [119]. Both the engine sound synthesis and the tyre noise synthesis are congruent with the idea of procedural audio, without any recordings used. A synthetic single vehicle pass-by sound is realised by adding the synthetic engine sound and synthetic tyre noise according to the descriptive parameters of the driving status of the vehicle. In this study, these two types of sounds are combined as a weighted sum of two point sources at different heights with different power distribution between them, following the method in the Harmonoise model [119].

In order to improve the plausibility of the auralisation following the principles of auralisation [7], some sound propagation effects have been implemented in the context of a micro-scopic urban environment. These effects include distance attenuation, the Doppler shift, and early reflections. It is found that distance attenuation can be calculated according to the inverse square law of a moving source. After analysing the distortion caused by linear interpolation and windowed-sinc interpolation, it is decided that although a sinc interpolation method is ideal for audio signals, a linear interpolation is the prioritised choice of Doppler shift implementation in this study because it is less computationally demanding and the potential artefacts are less pronounced for pass-by sounds. As computation time goes up exponentially with the increase of image source order, only the $1^{st}$-order image source algorithm has been implemented for rendering early reflections in the micro-scopic urban scene.

This auralisation is presented binaurally with HRTF processing based on the KEMAR HRTF database [92]. Considering the relatively obvious perception of low frequency sounds, a low frequency correction for the KEMAR HRTF database is implemented in this study, based on the method proposed by Xie [297]. Traffic flow auralisation is achieved by summing the auralised single vehicles according to the descriptive parameters of traffic flow dynamics.

The auralisation framework is built in Unity3D as the hosting platform. The sound source modelling is developed in Pure Data, and the sound propagation effects are implemented in Wwise and Matlab. The parameter transmission between different platforms is realised using open-source middleware or plug-ins working across Wwise, PD, and Unity3D. It is found that a highly flexible and interactive auralisation of traffic flow noise using procedural audio methods can be achieved and implemented with these tools by combining the algorithms developed in this study. The plausibility of such a procedural audio based auralisation approach will be evaluated by subjective listening tests in the next chapter.

# Chapter 7

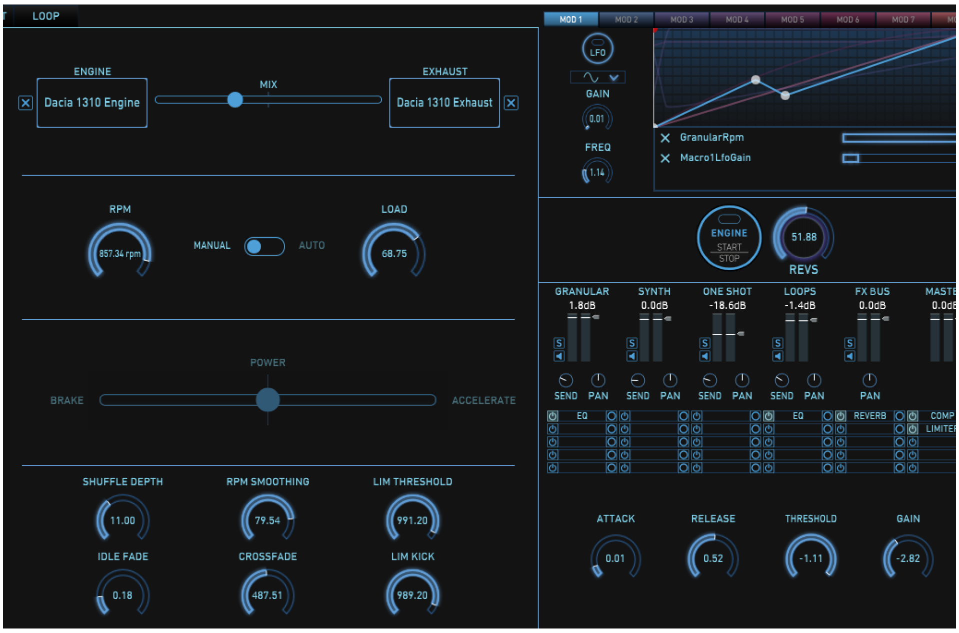
# Evaluation of the Auralisation Framework

In the previous chapter, an auralisation framework for traffic flow noise based on single vehicle pass-bys has been presented. It is built based on procedural audio methods, in which sound source models for engine noise and tyre noise synthesis have been developed, as well as a model for the main sound propagation effects and HRTF processing for spatial audio rendering of vehicle pass-bys at a microscopic level for a given urban scene. As discussed in Chapter 4, when using procedural audio, the plausibility of the result should be seriously considered. Therefore the next step is to evaluate whether the procedural audio based auralisation of traffic noise can actually sound plausible. If so, under what conditions and to what extent can plausibility be achieved.

Before this evaluation, it is important to clarify two terms *plausibility* and *authenticity*. For sound reproduction, plausibility refers to the agreement of the auditioned scene with an expected inner reference (expectations) [63], while authenticity is usually used to describe whether a sound scene is perceptually identical to an external reference [64]. It is considered that plausibility is more important for VR applications where a real reference is not available for subjects, and authenticity is more useful for AR applications where the coherence of virtual sound with environmental sound can be easily noticeable [310]. In practice, these two terms are often used without noticeable differences for the evaluation of auralisation studies since most auralisation models are designed for demonstration purposes rather than running on a specific VR/AR platform. This can be demonstrated in either a virtual environment or based on a real situation, according to the specific requirement and purpose of the auralisation. Therefore, in this thesis, a wider concept of plausibility is considered, which can be expressed as the agreement of the auditioned scene with expectations, or compatibility with an external reference (e.g. a playback recording or a real sound). A 'plausible' auralisation means that subjects should consider the sound scene credible for a real-world circumstance when it happens.

Following the concept of plausibility above, the purpose of this chapter is to check whether the auralised traffic noise, when using the proposed procedural audio methods, actually sounds like a vehicle, a car pass-by, or the traffic flow event that is simulated. A recording-based sound synthesis technique – granular synthesis is used for creating the counterparts for comparison. The granular synthesis is implemented in a commercial sound design tool named *Igniter*, which is developed based on the granular synthesis algorithm proposed by Jagla and Maillard [262]. Figure 7.1 shows the graphical user interface of this granular synthesis tool. For each specific vehicle model, a dataset of engine sounds and exhaust sounds at different driving patterns is required. The main driving parameters to run the model are the engine speed and engine load. Some other parameter, such as EQs, gains, and low frequency oscillation (LFO), etc. are used to give extra control of the granular synthesis model to fine-tune the sound for different applications.

As discussed in Chapter 4, if all the properties of the sound can be inherently captured and replayed in recordings, there should be fewer plausibility issues for recording-based audio reproduction than for procedural audio. Moreover, the perceived plausibility of the granular synthesis model for vehicle sound synthesis has also been quantitatively validated by Maillard for a microscopic urban scene through a subjective listening test [311]. It is hypoth-



ENGINE ☒ Dacia 1310 Engine
MIX
EXHAUST Dacia 1310 Exhaust ☒
RPM 857.34 rpm
LOAD 68.75
SHUFFLE DEPTH 11.00
RPM SMOOTHING 79.54
LIM THRESHOLD 991.20
IDLE FADE 0.18
CROSSFADE 487.51
LIM KICK 989.20
GAIN 0.01
FREQ 1.14
× GranularRpm
× Macro1LfoGain
REVS 51.88
GRANULAR 1.8dB
SYNTH 0.0dB
ONE SHOT -18.6dB
LOOPS -1.4dB
FX BUS 0.0dB
MASTE 0.0dE
ATTACK 0.01
RELEASE 0.52
THRESHOLD -1.11
GAIN -2.82

**Figure 7.1:** *Graphical user interface of Igniter, a commercial sound design tool developed based on granular synthesis methods.*

esised that if there is no perceived difference between a sound scene auralised by procedural audio and recording-based audio, the procedural audio model can be considered plausible for rendering such a sound scene.

This chapter is organised as follows: firstly it presents the specific hypotheses which correspond to the evaluation of the plausibility of the proposed auralisation framework for different acoustic scenes; secondly, it documents the design, implementation and results of two listening tests for exploring these hypotheses; a discussion on the results is then presented based on a statistical analysis of the listening test results.

## 7.1 Introduction

### 7.1.1 Hypotheses for the listening tests

Before designing a listening test, it is crucial to clarify the specific research questions and the associated hypotheses to be verified for making any conclusions. As the developed auralisation framework is used for creating pass-by noise in a microscopic urban scene, either for a single vehicle or a traffic flow, the following hypotheses are proposed for evaluating its plausibility corresponding to different sound scenes:

- **Hypothesis 1**: The perceived plausibility of the auralised single vehicle pass-by noise via procedural audio is comparable to auralisation via recording-based audio;

- **Hypothesis 2**: The perceived plausibility of the auralised traffic flow noise via procedural audio is comparable to auralisation via recording-based audio;

As discussed in Chapter 6, there are multiple components and steps for auralisation of vehicle pass-by noise, such as the engine sound, tyre noise, Doppler effects, etc. It is assumed that each component or step contributes to the output sounds, modifying the reality perception of the auralised scene to some extent. The perceived differences between a procedural audio model and a granular synthesis model in a specific component/step may be weakened or strengthened when other components or effects are added in the steps that follow. In the following sections, two listening tests pertaining to the plausibility of a single vehicle pass-by sound and traffic flow noise are conducted. The perceived differences between the procedural audio model and the granular synthesis model in different steps during the modelling process are checked by subjective evaluation. The purpose of these tests is to explore to what extent plausibility can be confirmed for such an auralisation framework, even though the individual plausibility for some specific components might not be considered as being high. The result of this subjective evaluation might also be useful when tuning the parameters of the model in terms of plausibility – it is assumed that an auralisation model can be adjusted from a holistic or global perspective rather than focusing purely on some single components.

### 7.1.2 Related works on plausibility evaluation of prior auralisation models for a pass-by vehicle

As discussed in Chapter 3, there is still a lack of guidance on how to evaluate the plausibility of auralised sounds. Some prior auralisation models for vehicle noise have been validated by subjective listening tests in terms of plausibility, for which the methodologies and the conclusions can be taken for reference here. Apart from the listening test conducted by Maillard [311] as discussed previously, Pandharkar [260] implemented an A/B listening test for analysis of the use of spectral modelling synthesis (SMS) in auralisation of a single vehicle pass-by sounds. The perceptual similarity between the original sound and the synthesised sound is compared in terms of realism, annoyance, and speed. Hoffman [312] implemented a perceptual validation of a tyre noise prediction model combined with the auralisation tool developed in the LISTEN project [252]. A semantic differential (SD) method was used with a categorical scale to compare the recorded and simulated signals. Southern and Murphy [313] implemented a MUSHRA test to evaluate the plausibility of a recording-based auralisation framework of vehicle pass-bys. They focused on comparing the plausibility of pure tyre noise synthesised by a recording-based method and an engineering method which is a modified version of the Harmonoise model in terms of tyre noise synthesis [119, 313].

As can be seen from the relevant literature, none of these listening tests is designed for evaluating the plausibility of a complete auralisation framework for vehicle pass-bys including sound source, propagation effects and spatial audio rendering. Furthermore, only single vehicle pass-bys are considered in these previous listening tests, and there has been no published perceptual validation for traffic flow auralisation so far. This thesis aims to bridge these gaps to some extent through the subjective evaluation of the plausibility of the complete auralisation framework developed for single vehicle pass-bys and traffic flow scenes.

### 7.1.3 Test overview

Two listening tests are designed for evaluating the plausibility of the developed auralisation framework. The first listening test is designed as an ABX test for evaluating a single vehicle pass-by noise to prove or disprove Hypothesis 1, and the second listening test is designed as a MUSHRA test for evaluating the auralisation of multiple vehicle pass-bys under different traffic flow circumstances to prove or disprove Hypothesis 2.

The first listening test is designed as an ABX test because it is the most straightforward way to explore whether subjects can determine the difference between two given samples. A series of sounds extracted from the different steps of the modelling process are set as test cases. The purpose is to check to what extent the added components or effects following the steps in the auralisation framework influence the perceived plausibility of the final sounds. Some questions can be raised as part of this process, such as, 'what about the plausibility of a synthetic engine sound', 'how will the plausibility change when tyre noise is added', and 'to what extent can plausibility be achieved when sound propagation effects are added', etc. If there is no perceived difference between the two stimuli created by two completely different methods for a given step, it means that the sound synthesised using these two methods can be considered as interchangeable from that step in the process until a perceived difference is found in a step that follows. The main benefit of such interchangeability is that the advantages or disadvantages of both methods might be considered. For example, it is feasible to take advantage of procedural audio in terms of its improved flexibility and variable cost if the sounds created by this model and the granular synthesis model are interchangeable without noticeable plausibility issues.

The second listening test is designed as a MUSHRA test because it is convenient for testing the perceived plausibility of intermediate cases between two extreme examples. For traffic flow auralisation, the extreme examples can be considered as, 'full recording-based audio' in which all single vehicle pass-bys are synthesised using recording-based methods, and 'full procedural audio' in which all single vehicle pass-bys are synthesised using procedural audio approaches. Then the intermediate cases can be achieved by changing the number of pass-by vehicles synthesised using the granular synthesis model and the procedural audio model. In this way, it is convenient to evaluate whether the plausibility of the procedural audio model is comparable to the recording-based granular synthesis model for traffic flow auralisation, and under what circumstances the sounds generated by the two methods are interchangeable.

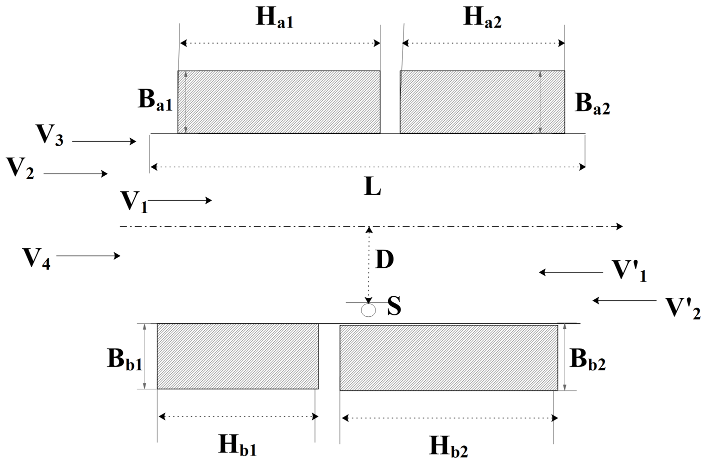


**Figure 7.2:** *The microscopic urban scene designed for the MUSHRA listening tests of traffic flow auralisation. Vehicle speeds are denoted by ($V_n$) and ($V'_n$). The length of the road is Lm, and the listening position S facing the road is set D at a distance away from the middle of the road. There are buildings facades set separately on each side of the road, of finite length $H_{a1}$, $H_{a2}$, $H_{b1}$, and $H_{b2}$, respectively. The corresponding widths for these buildingds are $B_{a1}$, $B_{a2}$, $B_{b1}$, and $B_{b2}$, respectively. All dimensions are given in metres.*

Figure 7.2 shows an example of a traffic flow scenario in a microscopic urban environment, which is simulated in this thesis. In this figure, vehicle speeds are denoted by ($V_n$) and ($V'_n$). The length of the road is $L$, and the listening position S facing the road is set $D$ at a distance away from the middle of the road. There are buildings facades setting separately on each side of the road, of finite length $H_{a1}$, $H_{a2}$, $H_{b1}$, and $H_{b2}$, respectively. The corresponding widths for these buildingds are $B_{a1}$, $B_{a2}$, $B_{b1}$, and $B_{b2}$, respectively. All dimensions are given in metres. The visual simulation is implemented within the game engine Unity3D using an open-source simulator AirSim [23], and a screenshot of the visual rendering is shown in Figure 6.13. Although visual aspects are not the main concern of this thesis, these scenes provide opportunities for interactive VR applications pertaining to vehicle pass-bys in an urban environment.

As introduced in Chapter 6, a vehicle pass-by sound is mainly composed of the engine sound and the tyre sound of that vehicle. Both of these two types of sounds are synthesised by different procedural audio methods in these tests. To compare the perceived plausibility, at least one of these two types of dominant sound sources should be substituted by the counterpart stimuli generated by the recording-based granular synthesis model. In these tests, engine sound synthesis is selected for the sound source to be substituted in each case.

There are several reasons for choosing the engine sound rather than the tyre sound or both of these sounds. Firstly, as discussed previously in this chapter, the plausibility of engine sound synthesis via granular synthesis has been validated by subjective evaluation [262, 311], so it is reasonable to use it as a 'ground truth' for comparison with other auralisation methods. Secondly, to the best of the author's knowledge, there has been very little published study on tyre noise synthesis by granular synthesis methods so far. Although Maillard and Jagla [112] have proposed a method based on Close-Proximity Method (CPX) measurements for granular synthesis of tyre noise, it is too difficult and cumbersome to take controlled measurements of a variety of tyre sounds emitted by different vehicles driving at different conditions in this way. Thirdly, in theory it is possible to take a series of in-situ recordings of pure tyre noise associated with the driving status of the vehicle together with the ground conditions, and use these recordings as the 'ground truth'. However, in practice, it is not cost-effective or even feasible to take recordings of all the required pure tyre sounds as too many samples are needed in terms of different vehicle types, vehicle speeds, and driving patterns, etc. Therefore, a granular synthesis model for the engine sound synthesis is used to create the counterpart stimuli for each example in the listening tests that follow. Tyre sounds in all the stimuli, when required, are synthesised using the procedural method based on the Harmonoise model [119] developed in Chapter 6. Directivity patterns are not implemented for the listening tests in these tests because the method based on the empirical Equation 6.9 still needs to be further validated with more measurement data [119], as discussed in Chapter 6. Hence, both engine sounds and tyre sounds assigned to the sound sources are considered as being omnidirectional for the listening tests here.

The listening tests are designed and conducted as online tests. Compared to on-site listening tests in a laboratory, there are both advantages and disadvantages for an online listening test. The greatest advantage is the convenience of recruiting participants if a test requires more subjects than locally available, particularly, if participants from different cultures speaking different languages are required. For example, Cartwright et al. [314] conducted a MUSHRA online listening test, in which they collected data from 540 participants all over the world within only 8.2 hours. Compared to laboratory tests, the most obvious disadvantage of online listening tests is the lack of control of the test environment, including listening environments, listening devices, etc. However, it is still an open question that how, and to what extent, such a lack of test environment control may impact the validity of an auditory experiment implemented online.

There are several studies on the correlation between the results of laboratory- and web-based listening tests. Schoeffler et al. [315] conducted an experiment to compare the results of musical instrument perception between laboratory- and web-based listening tests (62 vs 1168 subjects). They found very little significant difference between the results obtained from these two scenarios. Cartwright et al. [314] compared their data collected in the laboratory- and web-based MUSHRA listening tests (20 vs 540 subjects). They found that the resulting perceptual evaluation scores in the online listening test are comparable to those obtained in the controlled lab environment. Pedersen et al. [316] conducted an online listening test to rate the annoyance potential of neighbours' activities heard through different simulated walls. By taking a feasibility test with four stimuli and six walls, they found that the online test methodology is suitable for subjective evaluation of airborne sound insulation of walls in terms of rating the annoyance potential.

Conducting web-based auditory experiments is now becoming more popular [314, 317–320], especially for the validation of VR auditory environments [321]. According to relevant studies on the feasibility of online listening tests [314–316], it is reasonable to consider that if the experiment is well designed, and uncontrolled variables can be considered negligible for the research question, the results of web-based experiments should be reliable. In practice, auralisation is often utilised for communication as part of the design process, or for demonstrating to, e.g. stakeholders and the wider public. For these purposes, the plausibility should not vary too much with different listening environments or listening devices. Therefore, it is feasible to use online listening tests, without strict control of the listening environments and listening devices, so as to evaluate the plausibility of an auralisation in a wider context. On the introduction page for the two listening tests, subjects are recommended to be in a listening environment where they feel quiet and comfortable, and are asked to provide some information about their listening devices, such as model or brand, if possible.

## 7.2 Evaluation of auralised single vehicle sound

The goal of the first listening test is to verify to what extent the perceived plausibility of the auralised single vehicle noise via procedural audio is comparable to auralisation via recording-based audio. An ABX test is designed and implemented to explore answers to the following sub-questions:

- $Q_{s1}$: For a pure engine sound, to what extent is the perceived plausibility of the procedural audio model comparable to the granular synthesis model?

- $Q_{s2}$: For engine sound plus tyre noise, to what extent is the perceived plausibility of the procedural audio model comparable to the granular synthesis model for engine sound?

- $Q_{s3}$: For a single vehicle pass-by sound, to what extent is the perceived plausibility of the developed auralisation framework comparable to that based on the granular synthesis model for engine sound?

These sub-questions are considered from different perspectives of the level of details implemented for the auralisation of a single vehicle pass-by. $Q_{s1}$ is raised from the perspective of low-level details, focusing on the perceived plausibility of the pure engine sound, which is a single component within the auralisation framework. $Q_{s2}$ is raised from the perspective of a higher level, which concentrates on the sound source modelling of a pass-by vehicle consisting of engine sound and tyre noise. $Q_{s3}$ is raised from a holistic perspective, considering the sound scene of the single vehicle pass-by as a whole. The plausibility evaluation is then decomposed into three test cases, Case 1: Engine Sound, Case 2: Engine + Tyre Sound, and Case 3: Pass-by Sound, regarding $Q_{s1}$, $Q_{s2}$, and $Q_{s3}$, respectively. The aim of such a decomposition is to evaluate the plausibility from different perspectives, so as to find the appropriate context for the application of different auralisation methods.

### 7.2.1 Method – ABX Listening Test

In an ABX test, the sample 'X' is a random choice between stimulus 'A' or stimulus 'B'. A listener's task is to identify whether 'X' is identical to 'A' or 'B' for a number of trials. As introduced above, for evaluating the plausibility of the developed auralisation framework, three test cases are designed. A series of stimuli can be created by varying the high-level driving parameters of the procedural audio model and the granular synthesis model, such as the engine type, vehicle speed, and engine load, etc. The time length of each stimulus is 4–7s, according to its suitability for different test cases. All the stimuli are processed with a short (0.4s) fade-in and fade-out to sound more natural and avoid onset/offset discontinuity.

For each test case, sounds from a vehicle model with a flat-plane four-cylinder engine (V4) and a vehicle model with a cross-plane eight-cylinder engine (V8) are synthesised using the proposed procedural audio model and the granular synthesis model, respectively, and are used as pairs of stimuli for comparison.

For simplicity of the listening test, it is assumed that all the pass-by vehicles move at constant vehicle speeds between a range of 40–70km/h. The engine load is also considered as being constant, setting as 50% of the maximum value. For the V4 model, the gear ratio is set as 1.00, which corresponds to the $4^{th}$ gear for a specific vehicle model with a flat-plane four-cylinder engine. For the V8 model, the gear ratio is set as 1.67, which corresponds to the $3^{rd}$ gear for a specific vehicle model with a cross-plane eight-cylinder engine. For the V4 model, the axle ratio is set as 3.46, which corresponds to the axle ratio of a specific vehicle model with a flat-plane four-cylinder engine. For the V8 model, the axle ratio is set as 2.66, which corresponds to the the axle ratio of a specific vehicle model with a cross-plane eight-cylinder engine. It is considered the tyre radius is $r_{tyre} = 0.29$m for the listening tests. With these parameters defined, the corresponding engine speeds can be calculated by Equation 6.17. The driving parameters to run the procedural audio model and the granular synthesis model are summarised in Table 7.1.

For Case 1: Engine Sound test, ten stimuli are created for each vehicle model by the procedural audio model, with a 4s time length. Five of these stimuli are created by running the flat-plane four-cylinder model with constant engine speeds as inputs, and the other five stimuli are created by running the cross-plane eight-cylinder model with the same input parameters. The

Table 7.1: The driving parameters for creating the corresponding stimuli in the listening tests

| | Vehicle Speed (km/h) | Engine speed (RPM) | Gear ratio | Axle ratio |
|---|---|---|---|---|
| **Four-cylinders** | 40 | 1250 | 1.00 | 3.46 |
| | 45 | 1400 | 1.00 | 3.46 |
| | 50 | 1600 | 1.00 | 3.46 |
| | 60 | 1900 | 1.00 | 3.46 |
| | 70 | 2200 | 1.00 | 3.46 |
| **Eight-cylinders** | 40 | 1600 | 1.67 | 2.66 |
| | 45 | 1850 | 1.67 | 2.66 |
| | 50 | 2050 | 1.67 | 2.66 |
| | 60 | 2450 | 1.67 | 2.66 |
| | 70 | 2850 | 1.67 | 2.66 |

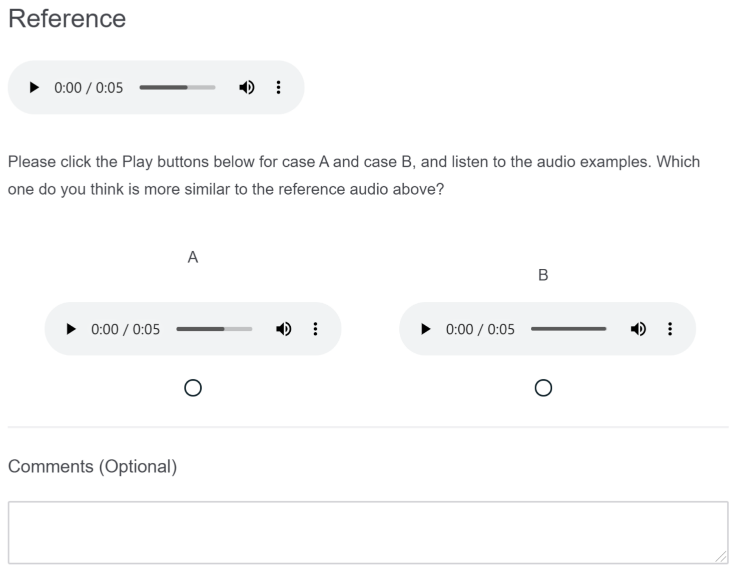
counterparts of these ten procedural audio stimuli are then created by the granular synthesis model with the same inputs for the engine speed and vehicle speed.

For Case 2: Engine + Tyre Sound test, all the engine sounds synthesised in the pure engine sound case are used, which are then combined with synthetic tyre noise. The tyre noise sounds are synthesised using the procedural audio model developed on the basis of the Harmonoise model [119] introduced in Chapter 6, with different vehicle speeds as the input parameters.

For Case 3: Pass-by Sound test, a series of auralised single vehicle pass-by sounds via the procedural audio model are compared with their counterparts synthesised using the granular synthesis method. All the sound propagation effects are simulated according to the microscopic urban scene in Figure 7.2 based on the methodology discussed in Chapter 6. The HRTF processing for binaural audio rendering are the same to the procedural audio model and the granular synthesis model.

Each participant is first presented with an experiment statement and consent statement, shown in Appendix B.1, followed by a calibration session in which the subject is asked to set appropriate sound volume for this test by listening to a pure 1kHz sound. Then a demographic questionnaire and a short training session are presented to let the participant familiarise themselves with the test procedure and user interface.

For each test case, ten pairs of A/B stimuli are selected in random order. In each pair, the sound synthesised using the procedural audio model is set as ‘A’ or ‘B’ randomly. Their counterparts, which are synthesised using the granular synthesis model, are set as the other choice in ‘A’ or ‘B’. ‘X’ is then selected randomly from ‘A’ or ‘B’. In total, there are 10 trials for each test case in a complete round. There is a comment box on the test page of each trial, which is optional to be filled-in if a participant is willing to leave any comments regarding the trial. For each participant, the test round repeats after completion of all these three test cases. All the questions in each test round are randomly presented to each subjects, so there is a total of 20 trials to be finished in each test case. All the questions are randomly presented to each subject. Considering all the 60 trials in these three test cases, it takes about 45 minutes for each participant to finish the complete listening test. Figure 7.3 demonstrates the graphical user interface of a typical test case in the ABX test.



Reference

Please click the Play buttons below for case A and case B, and listen to the audio examples. Which one do you think is more similar to the reference audio above?

Comments (Optional)

**Figure 7.3:** *An example of graphical user interface of the ABX test.*

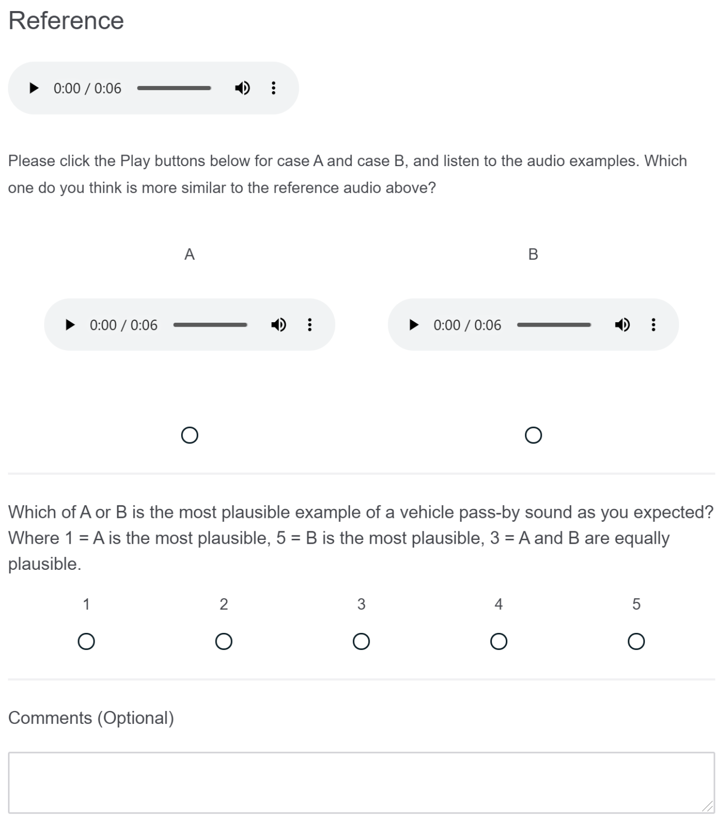
On top of the traditional ABX test, there is an additional question for the Case 3: Pass-by Sound test, stated as, ‘which of A or B is the most plausible example of a vehicle pass-by sound? where 1 = A is the most plausible, 5 = B is the most plausible, 3 = A and B are equally plausible’. This question is to evaluate the agreement of the auditioned scene with expected inner references (expectations) of the participant [63], rather than an external reference. In other words, it can be seen as the 'plausibility to expectation', which is assumed to be useful for sound evaluation in VR applications because it is usually difficult or impossible to find a 'ground truth' in a virtual world [63]. Participants should choose a value $k$ between 1–5 to evaluate whether 'A' or 'B' is more likely to be a pass-by sound as according to their personal experience and expectation. If $k > 3$, then $k$ points are added to the corresponding model for the 'plausibility to expectation', and $(6 - k)$ points is added to the other model. In contrast, if $k \leqslant 3$, $(6 - k)$ is to the corresponding model and $k$ points is added to the other model for the 'plausibility to expectation'.

In fact, this can be considered as a modified version of Alternative Forced Choice (AFC) test, which has been widely used in psychophysical studies [322, 323]. As no reference stimulus 'X' appears, the data collected for this question can be considered as the general preference on the plausibility, or 'plausibility to expectation' to some extent. Figure 7.4 demonstrates the graphical user interface of such an ABX test with an additional question on preference for plausibility.

### 7.2.2 Results of ABX Listening Test

These ABX listening tests were conducted following ethical approval from the University of York Physical Sciences Ethics Committee (PSEC) with a reference code 'Fu200519'.

There are a total of 27 participants in this ABX listening test. All the data are collected and stored anonymously. The demographic data collected in the test are shown in Figure 7.5. As can be seen from the histograms, there is a biased distribution in age, gender, and occupation. In fact, most of the participants are young students with some study or working experience in acoustics. Although there is no research on the effect of gender and age specif-



Reference

Please click the Play buttons below for case A and case B, and listen to the audio examples. Which one do you think is more similar to the reference audio above?

| A | B |
|---|---|
| 0:00 / 0:06 | 0:00 / 0:06 |
| ○ | ○ |

Which of A or B is the most plausible example of a vehicle pass-by sound as you expected? Where 1 = A is the most plausible, 5 = B is the most plausible, 3 = A and B are equally plausible.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

Comments (Optional)

**Figure 7.4:** *An example of graphical user interface of the ABX test with an extra question on preference for plausibility.*

ically for evaluation of auralisation studies, it has been claimed that there is no significant correlation between age/gender and noise sensitivity or annoyance in an urban environment [324–326]. Therefore, the statistical analysis of the collected data can be extrapolated into a wider population with other distribution patterns of gender and age. The study/work experience related to acoustics means that these participants are more reliable and discriminating than untrained listeners [327, 328], which enhances the robustness of the test results.

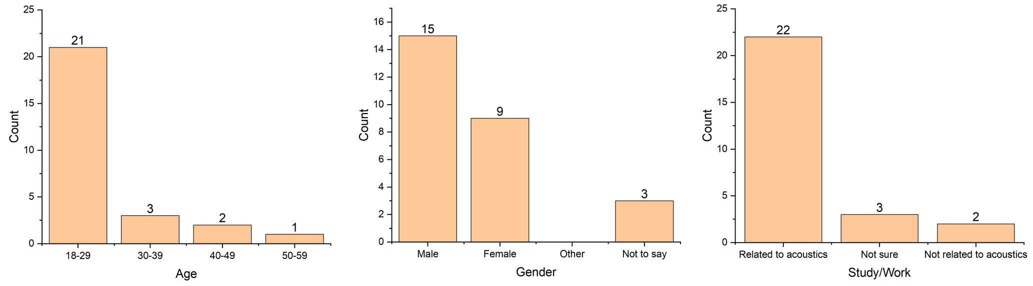


**Figure 7.5:** *Demographic data collected in the ABX listening test.*

The purpose of the ABX test is to find out whether subjects can reliably perceive the difference between sounds synthesised using the procedural audio model and the granular synthesis model or instead identify ‘X’ by guessing. The probability of successfully identifying ‘X’ in $n$ trials by guessing can be calculated by the *binomial distribution*:

$$P(X = x) = \begin{pmatrix} n \\ x \end{pmatrix} p^x \cdot (1-p)^{n-x} \tag{7.1}$$

where $x$ is the correct number of ‘X’ identified, $p = 50\%$ is the correct rate for guessing for each selection, and $n = 20$ is the number of trials for each test case. The null hypothesis $H_0$ and the corresponding alternative hypothesis $H_1$ for each test case is defined as follows:

- $H_0$: Subject cannot hear a difference between the procedural audio model and the granular synthesis model.
- $H_1$: Subject can hear a difference between the procedural audio model and the granular synthesis model.

The probability that each subject is guessing in each test case can be calculated using Equation 7.1, and the results in C.2 in Appendix C can be obtained. If the critical level is set as $\alpha = 5\%$, which means the $H_0$ should be rejected when the calculated probability is below 5%. According to Equation 7.1, this corresponds to a threshold integer count number 14. In other words, if a subject can correctly identify ‘X’ 14 times or more out of 20 trials, it is considered that that subject can perceive a difference between ‘A’ and ‘B’, at the 95% confidence level. The number of participants that cannot perceive a difference at the 95% confidence level in each test is shown as Figure 7.6.

As introduced at the beginning of this section, there is an extra question pertaining to the ‘plausibility to expectation’ for the pass-by scene. For each pair of pass-by sounds A/B, a subject gives $k$ points to one stimuli and $(6-k)$ points to the other stimuli, using a discrete 1–5 integer scale. For every subject,

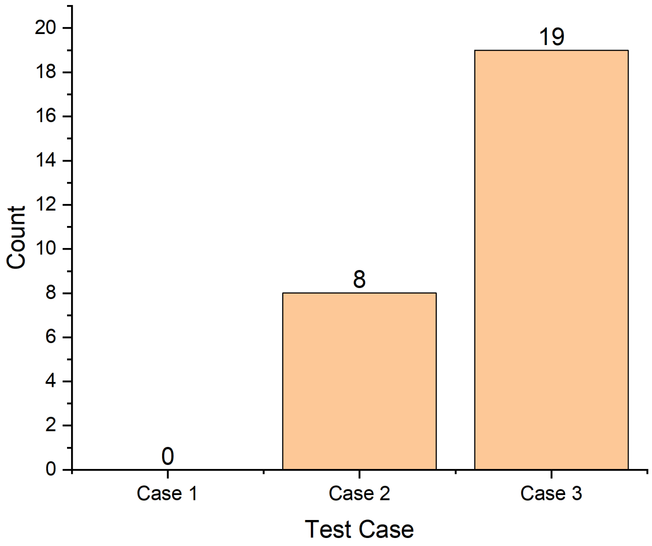


**Figure 7.6:** *The count number of participants that cannot perceive a difference at the 95% confidence level.*

points given to sounds synthesised using the same model (procedural audio or granular synthesis) will be summed together. The total points that every subject gave to the stimuli synthesised by each model are shown in Figure 7.10 based on the data obtained in Table C.3 in Appendix C.

### Discussion

As can be seen from Figure 7.6, at the 95% confidence level, the count number of retaining $H_0$ is 0 in test Case 1, which means that all of the 27 participants can perceive the differences between the sounds synthesised using the procedural audio model and the counterparts synthesised using the granular synthesis model, at the 5% level of significance. The mean rate of 'X' being identified correctly is 19.81/20=99.1%. The correct rate of guessing is assumed as $p = 50\%$. Thus, regarding $Q_{s1}$, it can be concluded that for a pure engine sound, 100% of the participants consider that the perceived plausibility of the procedural audio model is different from the granular synthesis model, at the 95% confidence level.

In fact, the differences between the pure engine sounds synthesised using the two methods can be obviously found by comparing the power spectral density of the synthetic sounds. Figure 7.7 shows an example of the power spectral density of two engine sounds synthesised by the granular synthesis model and the procedural audio model, respectively, corresponding to the status of a vehicle with an eight-cylinder engine driving at 40km/h. The energy higher than 12kHz is not displayed as the these synthetic engine sounds are calibrated regarding to the Harmonoise model, in which the empirical equation is valid in 1/3 octave bands from the frequency bin $f_c = 25$Hz to the frequency bin $f_c = 10$kHz. As can be seen from this figure, the energy of the granular synthesis engine sound has is higher than that of the procedural audio sound across a wide range of frequencies, particularly within 500–2000Hz. According to the results from the ABX test Case 1, these differences can be audibly perceived by the participants.

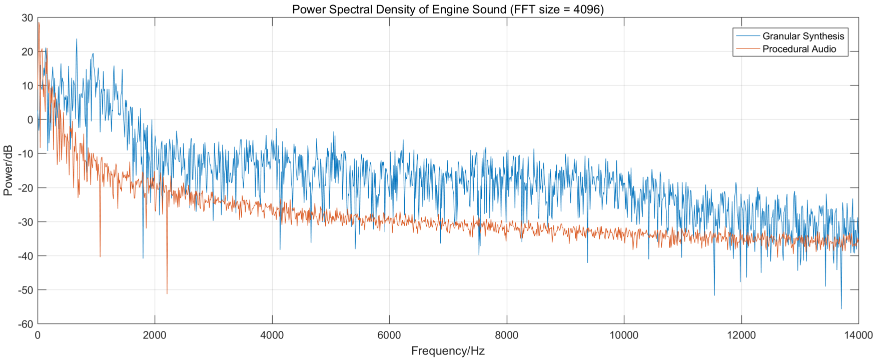


**Figure 7.7:** *Power spectral density (FFT size = 4096) of two engine sounds synthesised by the granular synthesis model and the procedural audio model, respectively, corresponding to the status of a vehicle with an eight-cylinder engine driving at 40km/h.*

This is mainly because of the different nature of these two sound synthesis methods. In granular synthesis, a dataset of recordings is used, in which most of the features in the original sounds can be captured and retained. In contrast, there is no recording involved when building up a procedural audio model. A series of pre-assumptions and simplifications are required to implement a model in practice, so a lot of details are neglected which may contribute to the sound quality or the plausibility of the sound, more or less. Therefore, it is easy to distinguish a recording-based synthetic engine sound and a procedural audio engine sound by hearing them. This is also reflected in the comments left by some participants, saying that 'they are totally different sounds' or 'It is too easy (to discriminate the sounds)'.

In test Case 2, the count number of retaining $H_0$ is 8, at the 95% confidence level. The mean rate of 'X' being identified correctly is 15.04/20=75.2%. When analysing the performance of each subject in test Case 1 and Case 2, it can be found that all participants did fewer correct identifications of 'X' in test Case 2 than Case 1, as shown in Figure 7.9. It reveals that when tyre noise is added on top of the engine sound, it is more difficult for all the participants to distinguish the sounds synthesised using the procedural audio model and the granular synthesis model. Thus, regarding $Q_{s2}$, it can be concluded that 8/27=29.6% of the participants consider the perceived plausibility of the procedural audio model is comparable to the granular synthesis model when tyre noise is added, and all of the participants found it is more difficult to distinguish the sounds synthesised using the two different methods when tyre noise is added on top of engine sound, at the 95% confidence level.

This result can be partly explained by the auditory masking theory introduced in Chapter 2. As there is a large portion of energy in low frequencies for tyre noise, which can be seen as a good masker for high frequencies in the synthetic engine sounds [35]. According to this theory, engine sound is partly masked by tyre noise when played simultaneously, so some of the differences between the procedural audio sound and the granular synthesis sound become inaudible when they are heard as a whole.

Figure 7.8 shows an example of the power spectral density of two engine+tyre sounds synthesised by the granular synthesis model and the procedural audio model, respectively, corresponding to the status of a vehicle with an eight-cylinder engine driving at 40km/h. The energy higher than 12kHz is not displayed as these sounds are synthesised based on the Harmonoise model, in which the empirical equation is valid in 1/3 octave bands from the frequency bin $f_c = 25$Hz to the frequency bin $f_c = 10$kHz. As can be seen from this figure, the energy distribution is similar for these two sounds across all the frequency range. According to the results from the ABX test Case 2, only 29.6% of the participants consider the perceived plausibility of the procedural audio model is comparable to the granular synthesis model when tyre noise is added. Therefore, it is not sufficient to use the power spectral density to describe the perceived differences between plausibility of the procedural audio model and the granular synthesis model when tyre sound is added to the engine sound.

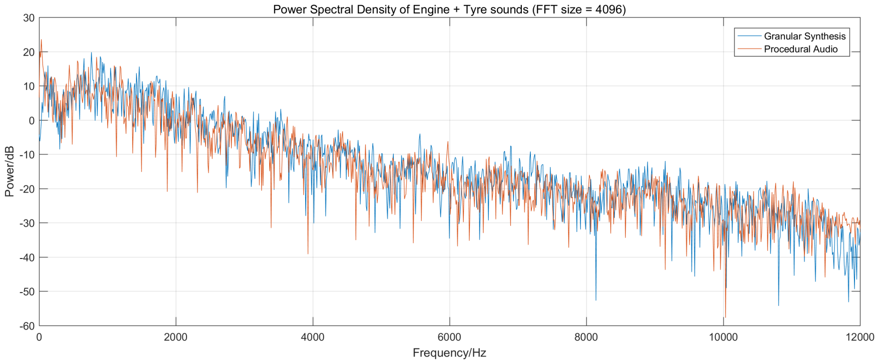


**Figure 7.8:** *Power spectral density (FFT size = 4096) of two engine+tyre sounds synthesised by the granular synthesis model and the procedural audio model, respectively, corresponding to the status of a vehicle with an eight-cylinder engine driving at 40km/h.*

In test Case 3, a total of 19 participants cannot hear a difference when the sound propagation effects and the spatial audio effects are added on top of engine sound and tyre noise, at the 95% confidence level.

The mean rate of ‘X’ identified correctly is 12.48/20=62.4%. It can be claimed that compared to test Case 2, it is more difficult to distinguish the sounds synthesised using the different two methods in Case 3. This can be explained from two perspectives. On one hand, the count number of retaining $H_0$ increases by 11, from 8 to 19, as shown in Figure 7.6. On the other hand, when analysing the performance of each subject in test Case 2 and Case 3, it can be found that 23 participants did fewer correct identifications of ‘X’ in test Case 3 than Case 2, and none of them can identified more correct ‘X’ in Case 3 than that in Case 2, which is shown in Figure 7.9. It means that the discrimination task becomes harder for at least 23/27=85.2% of the participants in Case 3 than Case 2. The difficulty of making such a distinction can also be reflected in the 'plausibility to expectation' result. The mean scores of 'plausibility to expectation' of the pass-by sounds synthesised by procedural audio and granular synthesis are 58.93 and 61.07, respectively, as shown in Figure 7.10. The standard deviation is 3.18 for the scores of both synthesis models, as listed in Table C.3. Considering the relatively small value of difference in the mean score and the standard deviation, it is reasonable to claim that for auralisation of a single vehicle pass-by, a similar level of 'plausibility to expectation' can be achieved by either the procedural audio model and the granular synthesis model. Thus, regarding $Q_{s3}$, it can be concluded that 19/27=70.1% of the participants consider that the perceived plausibility of the developed auralisation framework is comparable to that based on the granular synthesis model for engine sound, and at least 23/27=85.2% of the participants found that it is more difficult to distinguish the sounds synthesised using the two different methods when sound propagation effects and spatial audio effects are added, at the 95% confidence level. A similar level of 'plausibility to expectation' can be achieved by either the procedural audio model and the granular synthesis model for engine sound synthesis. In other words, the sounds auralised by the granular synthesis model and the procedural model are interchangeable in the case of a single vehicle pass-by for at least 70.1% people without plausibility issues. Although some people can perceive a difference between the granular synthesis model and the procedural audio model, a similar level of 'plausibility to expectation' is achieved by using either of these two methods.

Recalling these three sub-questions raised at the beginning of this section, the following answers can be summarised according to the ABX listening test conducted:

- $Q_{s1}$: For a pure engine sound, 100% of the participants consider that the perceived plausibility of the procedural audio model is different from the

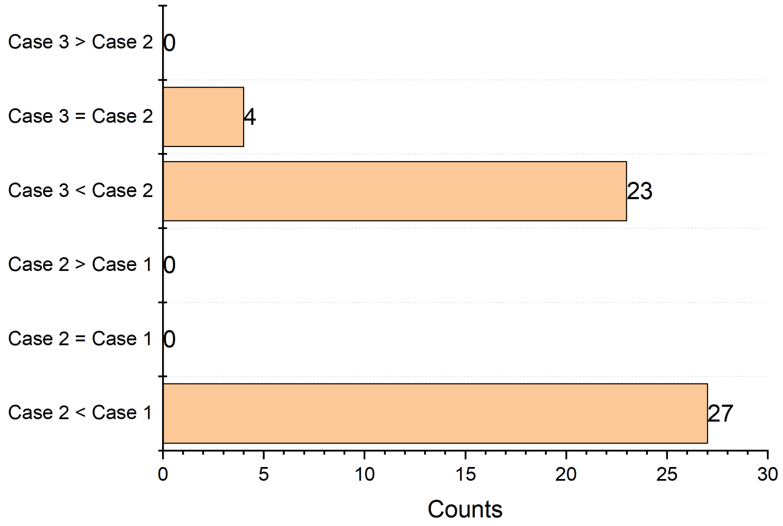


**Figure 7.9:** *The count number of subjects who make more/equal/fewer correct identified between different test cases.*

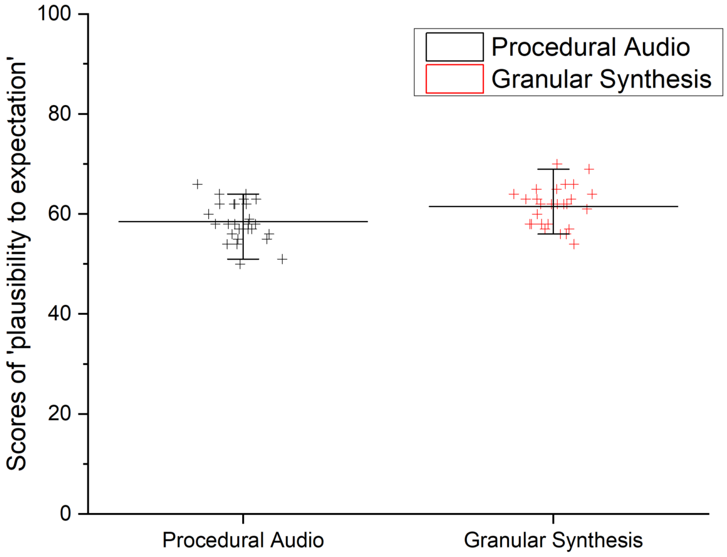


**Figure 7.10:** *The mean scores of 'plausibility to expectation' of the pass-by sounds synthesised by procedural audio and granular synthesis.*

granular synthesis model, at the 95% confidence level.

- $Q_{s2}$: For the engine sound plus tyre noise, 29.6% of the participants consider that the perceived plausibility of the procedural audio model is comparable to the granular synthesis model, and 100% of the participants consider that it is more difficult to distinguish the sounds synthesised using the two different methods when tyre noise is added on top of engine sound compared with the pure engine sound, at the 95% confidence level.
- $Q_{s3}$: For a single vehicle pass-by sound, 70.1% of the participants consider the perceived plausibility of the developed auralisation framework as being comparable to that based on the granular synthesis model for the engine sound, and at least 85.2% of the participants found it becomes more difficult to distinguish the sounds synthesised using the two different methods when sound propagation effects and spatial audio effects are added compared with engine sound plus tyre noise, at the 95% confidence level. A similar level of 'plausibility to expectation' can be achieved by either the procedural audio model and the granular synthesis model for engine sound synthesis.

In this section, the perceived plausibility of the auralisation framework developed in terms of a single vehicle pass-by has been validated using a ABX listening test. Based on this, the plausibility of traffic flow auralisation will be evaluated by a MUSHRA listening test that follows.

## 7.3 Evaluation of auralised traffic flow noise

With the validation of the plausibility of a single vehicle pass-by auralised using the proposed framework, the next step is to explore how plausibility may vary under different traffic flow conditions when multiple pass-by vehicles move with different driving patterns. According to the discussion in Chapter 6, the properties of a microscopic traffic flow can be described by a series of parameters including the vehicle velocity ($v$), vehicle density ($k$), and flow speed ($q$), etc. [301]. The second listening test is then designed to test different cases by changing these descriptive parameters to explore how the plausibility of the auralisation varies under different traffic flow conditions.

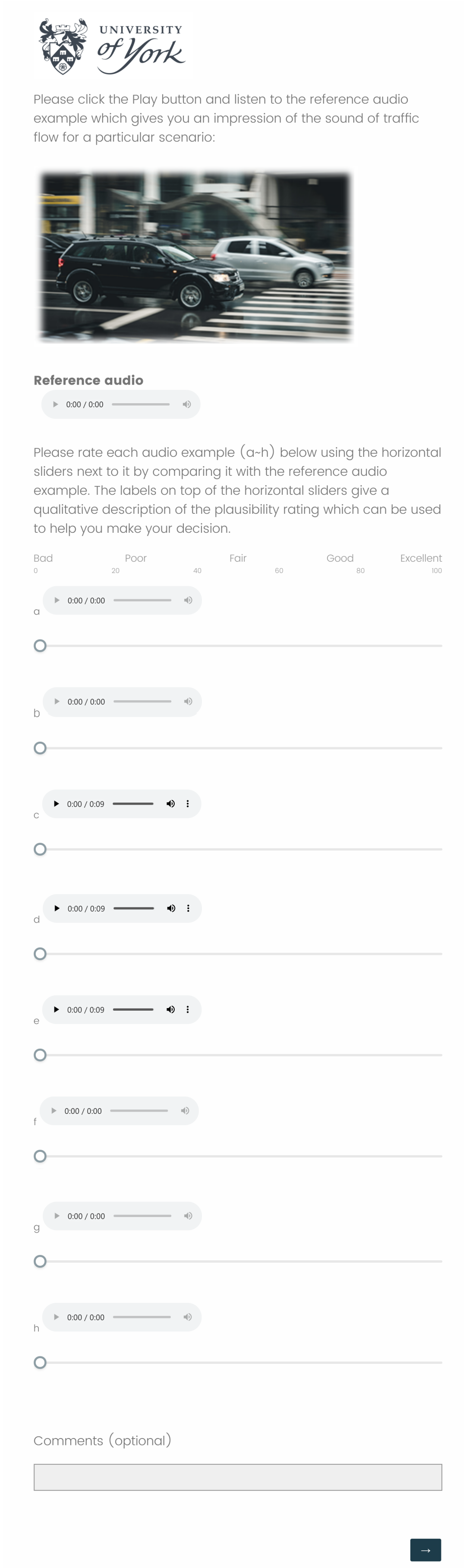
### 7.3.1 Method – MUSHRA Listening Test

This listening test is designed as a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) test. This method has been widely used for exploring perceived audio quality processed by different signal processing methods. According to the state-of-art guidance ITU-R BS.1534-3 [329] for this test method, a MUSHRA test case consists of a series of alternative stimuli treated in different ways that need to be rated at once using a scale of 0 to 100. An external reference is assumed to be of the best quality, which is also used as a hidden reference stimulus that is expected to get the highest rating. On the other hand, an anchor (usually corresponding to a low-pass filtered sound) is also included which represents a low-quality sound and is expected to be rated as the worst audio quality. There are often five additional text labels at intervals of 20 from each other on top of the 0 to 100 scale, marked as Bad, Poor, Fair, Good, and Excellent. The order of the stimuli is randomised for each listener.

As the aim of this listening test is to explore to what extent the plausibility of the proposed auralisation framework using procedural audio is comparable to the counterparts created by the granular synthesis method and whether sounds synthesised using these two methods are interchangeable without plausibility issues in different traffic flow conditions, here the stimuli 'treated by different methods' in the MUSHRA test are considered as the different mix of vehicle pass-by sounds synthesised by granular synthesis and procedural audio. The variation of traffic flow conditions is achieved by changing the descriptive parameters of traffic flow including the vehicle speeds ($v$), density ($k$), and flow speed ($q$), etc. For each traffic flow condition, a test case is created. In each test case, six pass-by vehicles are auralised according to the descriptive parameters set for that specific traffic flow. There are seven stimuli for each test case in addition to an anchor which is a low-pass filtered sound of one of these stimuli. The cutoff frequency for the low-pass filter is set as $f_c = 3.5\text{kHz}$ according to the guidance of ITU-R BS.1534-3 [329]. Hence, there are eight stimuli in total for each test case. Figure 7.11 demonstrates the graphical user interface of a typical test case in the MUSHRA test.

Among the seven stimuli except the Anchor, the 'pure procedural audio' stimulus (P6) is auralised by adding six single vehicle pass-by sounds synthesised using the procedural audio model, while the 'pure granular synthesis' stimulus (G6) is auralised by adding the counterparts created using granular synthesis. The remaining stimuli are intermediate cases which consist of a mix of granular synthesis sounds and procedural audio sounds varied by changing the number of pass-by vehicles auralised using these two methods (G1P5: 5 procedural audio + 1 granular synthesis; G2P4: 4 procedural audio + 2 granular synthesis; G3P3: 3 procedural audio + 3 granular synthesis; G4P2: 2 procedural audio + 4 granular synthesis; and G5P1: 1 procedural audio + 5 granular synthesis). The time length of each stimulus is 8–10s, according to the entering time of each vehicle and vehicle speeds. All the stimuli are processed with a short fade-in and fade-out to avoid particularly distinguishable onset/offset cues. In this way, it is feasible to evaluate whether the plausibility of the procedural audio model can be comparable to the recording-based granular synthesis model for a specific traffic flow condition, and to what extent the traffic flow sound generated by these two methods are interchangeable.

There are 7 test cases for the complete MUSHRA test, corresponding to 5 different typical traffic flow conditions in total. The first three test cases (Case 1 – Case 3) are designed based on the same traffic flow condition but with different sounds for the external reference (and also the hidden reference): pure granular synthesis, pure procedural audio synthesis, and 3 granular synthesis



Please click the Play button and listen to the reference audio example which gives you an impression of the sound of traffic flow for a particular scenario:

**Reference audio**

Please rate each audio example (a–h) below using the horizontal sliders next to it by comparing it with the reference audio example. The labels on top of the horizontal sliders give a qualitative description of the plausibility rating which can be used to help you make your decision.

Comments (optional)

**Figure 7.11:** *The graphical user interface of the MUSHRA test.*

pass-bys + 3 procedural audio pass-bys. This is different from the standard MUSHRA test, in which the external reference always corresponds to the sound with the highest quality. It is because of the nature of the listening test designed here – which is to compare the plausibility of traffic flow auralisation implemented by the procedural audio model and the granular synthesis model. The external reference is designed to bring an impression of the sound corresponding to a specific traffic flow condition, rather than presenting 'the best plausibility' of the example traffic flow sound. In fact, which stimulus has 'the best plausibility' is still an unknown question for each test case. Therefore, it is necessary to explore how the evaluation of plausibility may change when different external references are used. The other test cases (Case 4 – Case 7) correspond to 4 different traffic flow conditions, pertaining to the variation of vehicle directions, vehicle types, vehicle speeds, and flow speed compared to the first three test cases. Case 4 is designed by changing the vehicle direction in Case 1 from a mix of left to right and right to left into all left to right. Case 5 is designed by changing the vehicle type in Case 1 from three eight-cylinder vehicles + three four-cylinder vehicles into six four-cylinder vehicles. Case 6 is designed by changing the vehicle speeds in Case 1 from a mix of low-speed vehicles (40km/h) + mid-speed vehicles (50km/h) + high-speed vehicles (60km/h) into a mix of low-speed vehicles (40km/h, 45km/h) + high-speed vehicles (60km/h). Case 7 is designed by changing the flow speed from an 'averaged' entering-leaving time distribution into a 'congested' entering-leaving time distribution. The test cases are randomised for each listener. The driving parameters for all the 7 test cases are summarised as below:

- Case 1 – Case 3: 3 cross-plane eight-cylinder engine vehicles, 3 flat-plane four-cylinder engine vehicles; vehicle speeds: two 40km/h vehicles, two 50km/h vehicles, two 60km/h vehicles; driving directions: 3 left to right vehicles, 3 right to left vehicles; flow speed: two vehicles randomly enter between 0–1.5s, two vehicles randomly enter between 1.5–3s, two vehicles randomly enter between 3–4.5s. The engine speeds and engine loads are set according to Table 7.1. Stimuli synthesised using a pure granular synthesis for the engine sound, a pure procedural model, and a mix of 3 granular synthesis engine sound pass-bys + 3 procedural audio pass-bys are used as external references, respectively.

- Case 4: 3 cross-plane eight-cylinder engine vehicles, 3 flat-plane four-cylinder engine vehicles; vehicle speeds: two 40km/h vehicles, two 50km/h vehicles, two 60km/h vehicles; driving directions: six left to right vehicles; flow speed: two vehicles randomly enter between 0–1.5s, two vehicles randomly enter between 1.5–3s, two vehicles randomly enter between 3–4.5s. The engine speeds and engine loads are set according to Table 7.1.

- Case 5: 6 flat-plane four-cylinder engine vehicles; vehicle speeds: two 40km/h vehicles, two 50km/h vehicles, two 60km/h vehicles; driving directions: 3 left to right vehicles, 3 right to left vehicles; flow speed: two vehicles randomly enter between 0–1.5s, two vehicles randomly enter between 1.5–3s, two vehicles randomly enter between 3–4.5s. The engine speeds and engine loads are set according to Table 7.1.

- Case 6: 3 cross-plane eight-cylinder engine vehicles, 3 flat-plane four-cylinder engine vehicles; vehicle speeds: two 40km/h vehicles, two 45km/h vehicles, two 70km/h vehicles; driving directions: 3 left to right vehicles, 3 right to left vehicles; flow speed: two vehicles randomly enter between 0–1.5s, two vehicles randomly enter between 1.5–3s, two vehicles randomly enter between 3–4.5s. The engine speeds and engine loads are set according to Table 7.1.

- Case 7: 3 cross-plane eight-cylinder engine vehicles, 3 flat-plane four-cylinder engine vehicles; vehicle speeds: two 40km/h vehicles, two 50km/h vehicles, two 60km/h vehicles; driving directions: 3 left to right vehicles, 3 right to left vehicles; flow speed: 3 vehicles randomly enter between 0–1s, 3 vehicles randomly enter between 1–2s. The engine speeds and engine loads are set according to Table 7.1.

Figure 7.12 – Figure 7.15 shows some examples of the spectrograms of the synthetic traffic flow sounds to be test in test Case 1, including the stimuli synthesised using a pure granular synthesis for the engine sounds, a mix of 3 granular synthesis for engine sounds + 3 procedurally created engine sounds, a pure granular synthesis for the engine sounds, and the low-pass filtered anchor stimulus for the MUSHRA test, respectively. As can be seen from these spectrograms, the difference between the anchor and the other stimuli is relatively obvious. There is little visible difference between the stimuli synthesised by pure granular synthesis, a mix of 3 granular synthesis + 3 procedural audio, and a pure granular synthesis for the engine sounds. However, according to the results from Case 2 in the ABX test, the auditory perception of these sound signals may still be very different in terms of plausibility although they look similar from the perspective of sound levels.

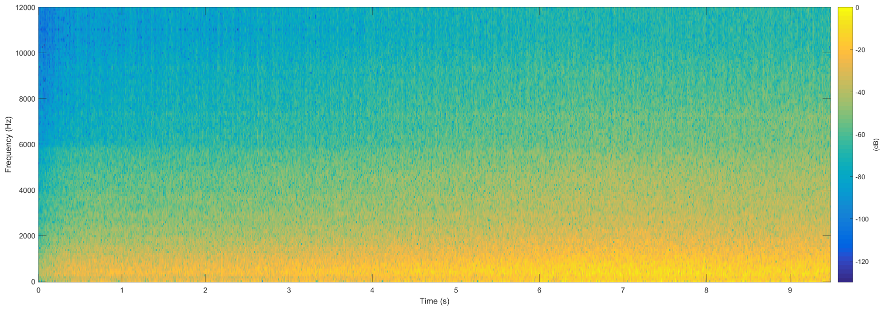


**Figure 7.12:** *The spectrogram of the auralised traffic flow in Case 1: Pure granular synthesis for the engine sounds (G6).*

### 7.3.2 Results of MUSHRA Listening Test

These MUSHRA listening test was conducted following ethical approval from the University of York Physical Sciences Ethics Committee (PSEC) with a reference code ‘Fu200519’.

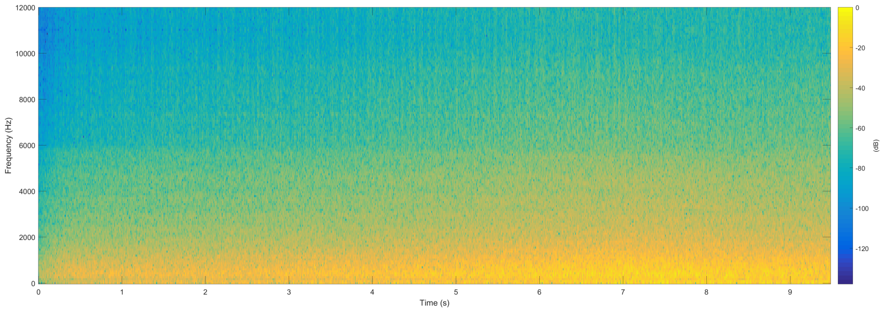


**Figure 7.13:** *The spectrogram of the auralised traffic flow in Case 1: Mix of granular synthesis and procedural audio for engine sounds synthesis (G3P3).*

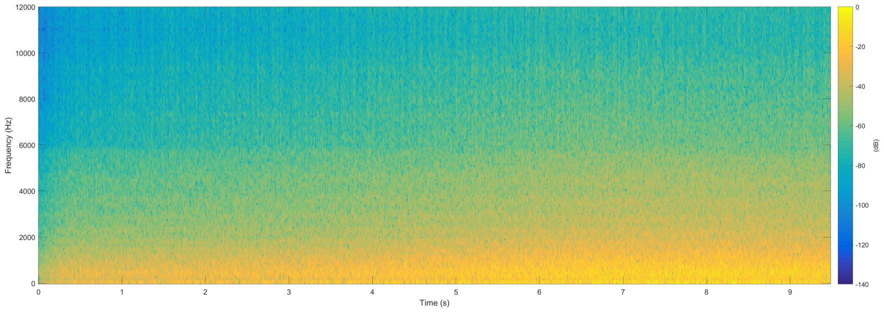


**Figure 7.14:** *The spectrogram of the auralised traffic flow in Case 1: Pure procedural audio for the engine sounds (P6).*

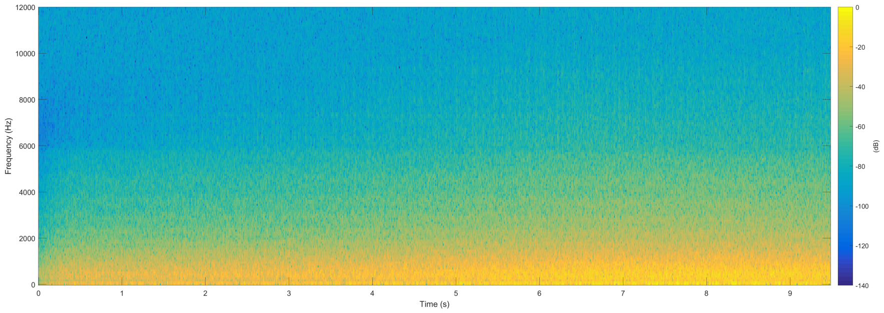


**Figure 7.15:** *The spectrogram of the auralised traffic flow in Case 1: the low-pass filtered Anchor.*

Each participant is first presented with an experiment statement and consent statement, shown in Appendix B.2, followed by a calibration session in which the subject is asked to set appropriate sound volume for this test by listening to a pure 1kHz sound. Then a demographic questionnaire and a short training session are presented to let the participant familiarise themselves with the test procedure and user interface.

There are a total of 21 participants in this listening test. All the data are collected and stored anonymously. By post-screening of the test data, one response has been rejected because the rating of the hidden reference is not conspicuously lower than other examples, which does not fulfil the requirement for statistical analysis of the MUSHRA test. Therefore, there are a total of 20 effective responses. The demographic data collected from the effective responses are shown in Figure 7.16. As can be seen from the histograms, there is also a biased distribution in age, gender, and occupation, which is similar to that of the first ABX listening test. Most of the participants are young students with some study or working experience in acoustics, which is again similar to the first ABX listening test. As discussed previously, it can be considered that these participants are more reliable and discriminating than untrained listeners [327, 328], which enhances the robustness of the results of the MUSHRA test.

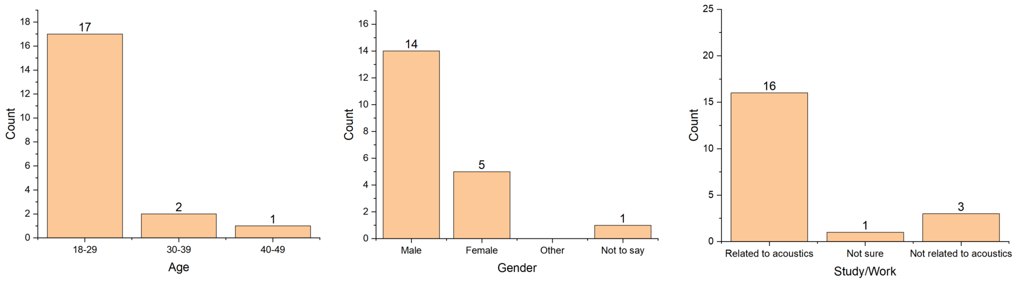


**Figure 7.16:** *Demographic data collected in the MUSHRA listening test.*

Figure 7.17 demonstrates the MUSHRA result for Case 1 in a boxplot, which is created according to the rating data in Appendix C. The stimulus consisting of six pass-by vehicles for which engine sounds are synthesised using granular synthesis is the external reference. Medians of ratings are shown as black horizontal bars. According to the guidance in [330], data obtained from a small number of listeners participating in MUSHRA tests are usually not normally distributed, nor uncorrelated to each other, because subjects usually take direct comparisons of multiple stimuli within a test page rather than rate using the full scale to give an absolute rating value. Therefore, parametric statistics such as ANOVA test are not suitable for the statistical analysis here. It is recommended that the non-parametric Friedman test should be used as the alternative to the parametric ANOVA test for MUSHRA data [330, 331]. As the Friedman test is sensitive to unequal distributions, the direct comparison between any stimulus with the hidden reference should be avoided when using this test because the hidden reference always has a larger rank value [330]. As can be seen from Figure 7.17, the rating of the anchor is obviously lower than the other stimuli, while the rating of the hidden reference is obviously higher than the other stimuli. The main purpose of the non-parametric Friedman test is to investigate whether there are significant differences between other stimuli apart from the anchor and the hidden reference.

By taking a Friedman test for the MUSHRA data from stimuli P6, G1P5, G2P4, G3P3, G4P2, and G5P1, it can be found that there is no significant difference between these stimuli in terms of plausibility at the 95% confidence level ($p = 0.091 > 0.05, \chi^2(5) = 4.03$). It reveals that the plausibility of these stimuli, either synthesised using the procedural audio model or a mix of granular synthesis for the engine sounds and the procedural audio model, are comparable to each other under a traffic flow condition with various types of multiple pass-by vehicles driving bidirectionally at a mix of low, mid and high vehicle speeds, under a smooth flow rate.

Figure 7.18 demonstrates the MUSHRA result for Case 2 in a boxplot, which is created according to the rating data in Appendix C. The stimulus consisting of three pass-by vehicles for which engine sounds are synthesised using granular synthesis and three pass-by vehicles for which engine sounds are

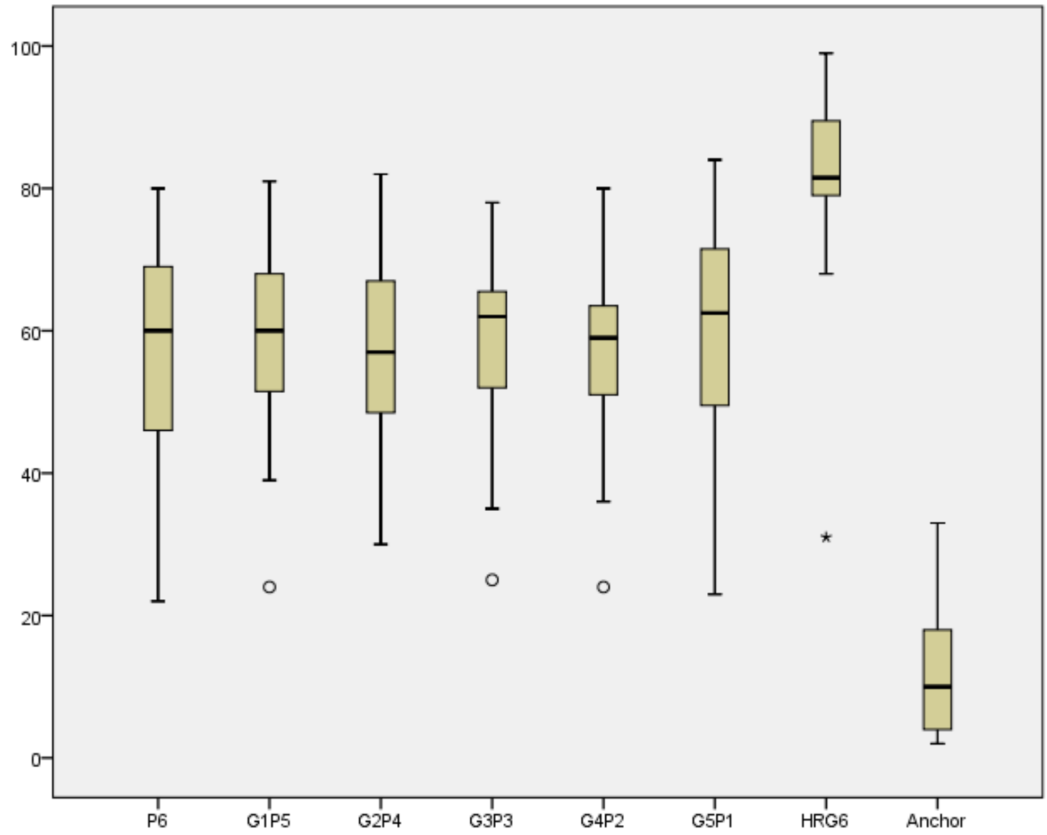


**Figure 7.17:** *Result of Case 1 in the MUSHRA listening test. The reference case used is G6 which is 'pure granular synthesis' sounds of six vehicle pass-bys.*

synthesised using the procedural audio model is used as the external reference. As can be seen from Figure 7.18, the rating of the anchor is obviously lower than other stimuli, while the rating of the hidden reference is obviously higher than the other stimuli. By taking a Friedman test for the MUSHRA data from stimuli P6, G1P5, G2P4, G4P2, G5P1, and G6, it can be found that there is no significant difference between these stimuli in terms of plausibility at the 95% confidence level ($p = 0.06 > 0.05, \chi^2(5) = 10.41$). As the traffic flow condition in Case 2 is identical to that in Case 1, it reveals that the plausibility of the pure procedural audio model, a mix of granular synthesis for the engine sounds and the procedural audio model, and pure granular synthesis for the engine sound are comparable to each other under a traffic flow condition with various types of multiple pass-by vehicles driving bidirectionally at a mix of low, mid and high vehicle speeds, under a smooth flow rate.

Figure 7.19 demonstrates the MUSHRA result for Case 3 in a boxplot,

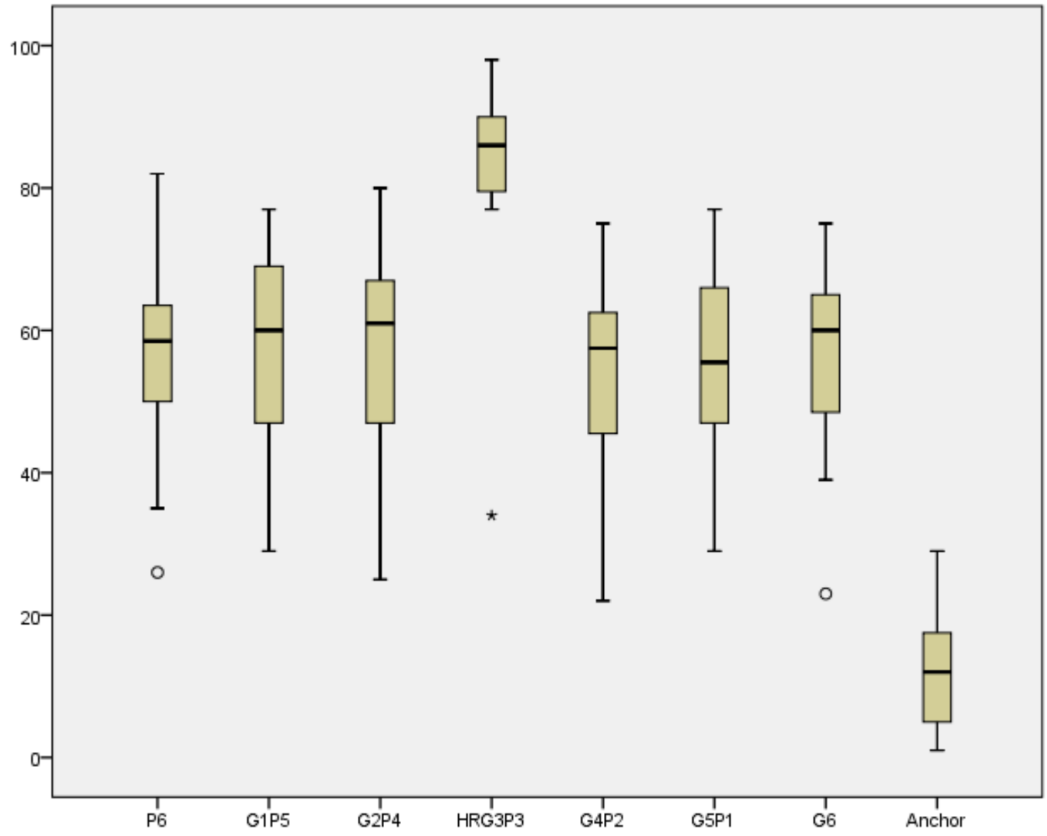


**Figure 7.18:** *Result of Case 2 in the MUSHRA listening test. The reference case used is G3P3 which is a mix of 3 vehicle pass-bys sounds synthesised using granular synthesis and 3 vehicle pass-bys sounds synthesised using procedural audio.*

which is created according to the rating data in Appendix C. The stimulus consisting of six pass-by vehicles for which engine sounds are synthesised using the procedural audio model is used as the external reference. As can be seen from Figure 7.19, the rating of the anchor is obviously lower than other stimuli, while the rating of the hidden reference is obviously higher than the other stimuli. By taking a Friedman test for the MUSHRA data from stimuli G1P5, G2P4, G3P3, G4P2, G5P1, and G6, it can be found that there is no significant difference between these stimuli in terms of plausibility at the 95% confidence level ($p = 0.39 > 0.05, \chi^2(5) = 5.24$). As the traffic flow condition in Case 3 is identical to that in Case 1, it reveals that the plausibility of a mix of granular synthesis for the engine sounds together with the procedural audio model,is comparable to using only pure granular synthesis for the engine sound under a traffic flow condition with various types of multiple pass-by vehicles drive bidirectionally at a mix of low, mid and high vehicle speeds, under a smooth flow rate.

When combining the results from Case 1, Case 2, and Case 3, it can be seen that the plausibility of the pure procedural audio model, a mix of granular synthesis for the engine sounds and the procedural audio model, and pure granular synthesis for the engine sound, are comparable to each other whatever the external reference is used. This can be also verified using the non-parametric Friedman test. Figure 7.20 shows a summary of ratings for different stimuli acquired in Case 1, Case 2, and Case 3. Ratings for the hidden references are excluded because they are not valid for these statistical analysis. By taking a Friedman test for all data collected in these three test cases, it can be found that there is no significant difference between these stimuli in terms of plausibility at the 95% confidence level ($p = 0.15 > 0.05, \chi^2(5) = 23.12$). In other words, the external reference is interchangeable without plausibility issues for the auralised traffic flow condition. In Case 4 to Case 7, as the same auralisation models and driving parameters are utilised, it is assumed that the interchangeability of the external reference still holds for these cases. As the

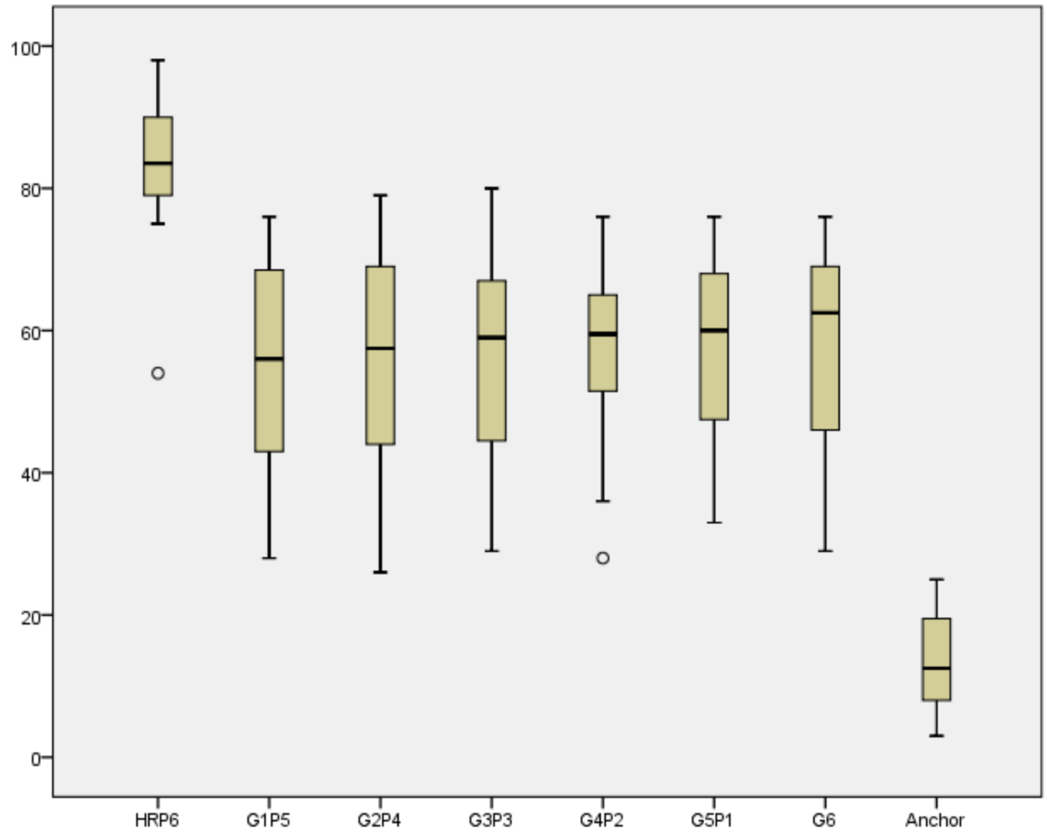


**Figure 7.19:** *Result of Case 3 in the MUSHRA listening test. The reference case used is P6 which is 'pure procedural audio' sounds of six vehicle pass-bys.*

absolute ratings of the external reference are invalid for the Friedman test, a mix of granular synthesis for the engine sounds and the procedural audio model (G3P3) is used as the external reference for the remaining test cases. This is for obtaining valid ratings on the extreme stimuli synthesised by pure procedural audio and pure granular synthesis for the engine sound for statistical analysis. Even though the ratings for G3P3 stimuli are no longer valid, there are other mixed stimuli (e.g. G2P2, G5P1) which provide a reasonable context for the mix of procedural audio model and granular synthesis.

Figure 7.21 demonstrates the MUSHRA result for Case 4 in a boxplot, which is created according to the rating data in Appendix C. As can be seen from Figure 7.21, the rating of the anchor is obviously lower than other stimuli, while the rating of the hidden reference is obviously higher than the other stimuli. By taking a Friedman test for the MUSHRA data from stimuli P6, G1P5, G2P4, G4P2, G5P1, and G6, it can be found that there is a significant

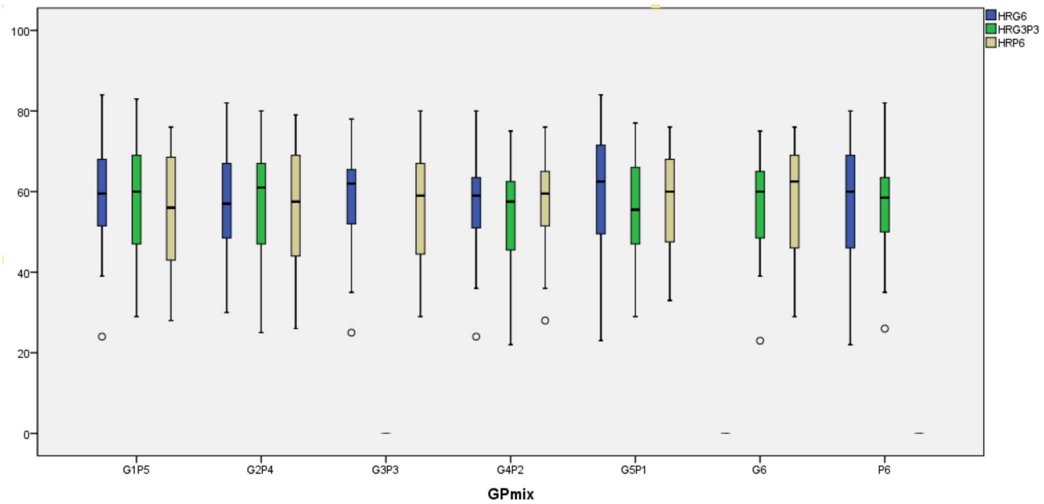


**Figure 7.20:** *Summary of results in Case 1, Case 2, and Case 3 without the ratings of hidden references and anchors. This is for comparison of the plausibility ratings when different external reference stimuli are used.*

difference between these stimuli in terms of plausibility at the 95% confidence level ($p = 0.001 < 0.05, \chi^2(5) = 20.59$). Post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied [332], resulting in a significance level set at $p < 0.004$. Median (IQR) perceived effort levels for P6, G1P5, G2P4, G4P2, G5P1, and G6 are 54.5 (48.25 to 71), 58 (46.25 to 68.75), 58 (49 to 69), 64 (58.25 to 72), 66.5 (58.25 to 74.75), and 66 (58.5 to 71), respectively. Using Wilcoxon signed-rank tests for each pair of stimuli, it is found that there is a statistically significant difference in perceived plausibility for pairs G5P1-P6 ($Z = -2.92, p = 0.003$), G5P1-G2P4 ($Z = -2.84, p = 0.04$), and G6-G2P4 ($Z = -2.88, p = 0.04$). There are no statistically significant differences between other pairs. By comparing the medians and the quantiles of these pairs, it can be concluded that the G5P1 sound is more plausible than the P6 sound, the G5P1 sound is more plausible than the G2P4 sound, and the G6 sound is more plausible than the G2P4 sound, at the 95% confidence level.

These perceived differences in terms of plausibility can also be reflected in the homogeneous subsets created with a stepwise step-down method [333], shown in Figure 7.22. It demonstrates the clustering of homogeneous groups together in the same column of the resulting table with differences marked with different colour codes. As can be seen from this figure, the output of the average rank of the six stimuli are clustered into two subsets: the subset composed of traffic flow consisting of 0–2 granular synthesis pass-by sounds and 4–6 procedural audio pass-by sounds, and the other subset composed of traffic flow consisting of 4–6 granular synthesis pass-by sounds and 0–2 procedural audio pass-by sounds, at the 95% confidence level. It reveals that there are perceived differences between the traffic flow sounds synthesised using the procedural audio model, a mix of granular synthesis for engine sounds and the procedural audio model, and pure granular synthesis for engine sounds under a traffic flow condition with various types of multiple pass-by vehicles drive one-way at a mix of low-, mid- and high speed vehicles, under a smooth flow rate. A traffic flow sound that consists of more recording-based engine sounds tends to be rated higher in terms of plausibility under such a traffic flow condition.

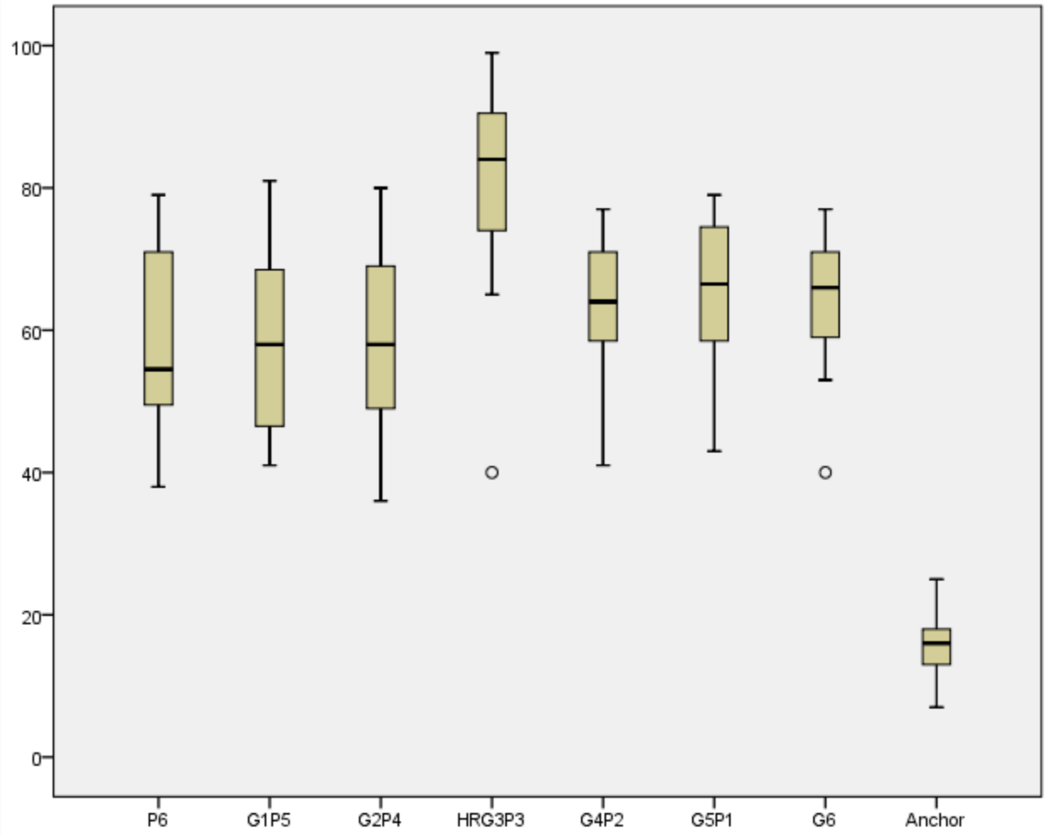
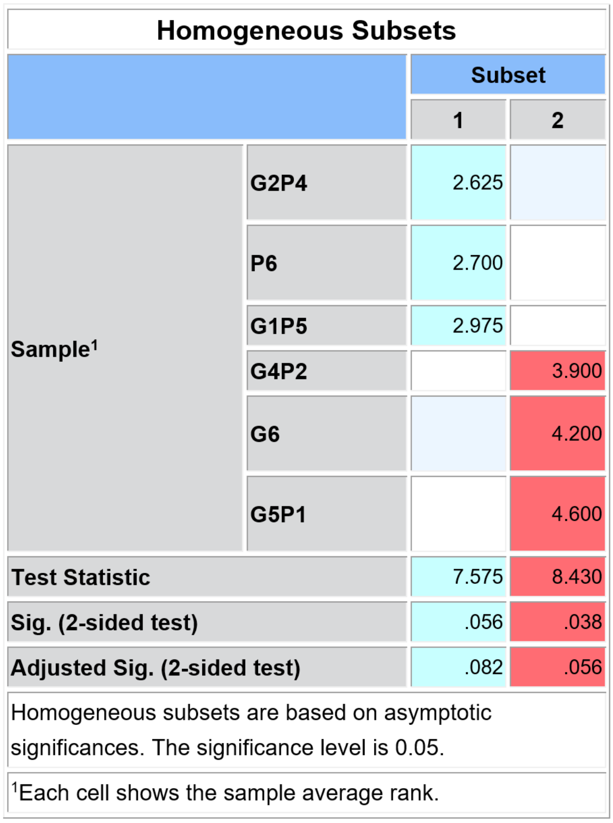


**Figure 7.21:** *Result of Case 4 in the MUSHRA listening test. The reference case used is G3P3 which is a mix of 3 vehicle pass-bys sounds synthesised using granular synthesis and 3 vehicle pass-bys sounds synthesised using procedural audio.*



**Homogeneous Subsets**

| | | Subset | |
|---|---|---|---|
| | | 1 | 2 |
| Sample[1] | G2P4 | 2.625 | |
| | P6 | 2.700 | |
| | G1P5 | 2.975 | |
| | G4P2 | | 3.900 |
| | G6 | | 4.200 |
| | G5P1 | | 4.600 |
| Test Statistic | | 7.575 | 8.430 |
| Sig. (2-sided test) | | .056 | .038 |
| Adjusted Sig. (2-sided test) | | .082 | .056 |

Homogeneous subsets are based on asymptotic significances. The significance level is 0.05.

[1]Each cell shows the sample average rank.

**Figure 7.22:** *Homogeneous Subsets of results of Case 4 in the MUSHRA listening test.*

Figure 7.23 demonstrates the MUSHRA result for Case 5 in a boxplot, which is created according to the rating data in Appendix C. As can be seen from Figure 7.23, the rating of the anchor is obviously lower than other stimuli, while the rating of the hidden reference is obviously higher than the other stimuli. By taking a Friedman test for the MUSHRA data from stimuli P6, G1P5, G2P4, G4P2, G5P1, and G6, it can be found that there is no significant difference between these stimuli in terms of plausibility at the 95% confidence level ($p = 0.97 > 0.05, \chi^2(5) = 0.86$). It reveals that the plausibility of the pure procedural audio model, a mix of granular synthesis for the engine sounds and the procedural audio model, and pure granular synthesis for the engine sound are comparable to each other under a traffic flow condition with a reduced type of pass-by vehicles driving bidirectionally at a mix of low, mid and high vehicle speeds, under a smooth flow rate.

Figure 7.24 demonstrates the MUSHRA result for Case 6 in a boxplot,

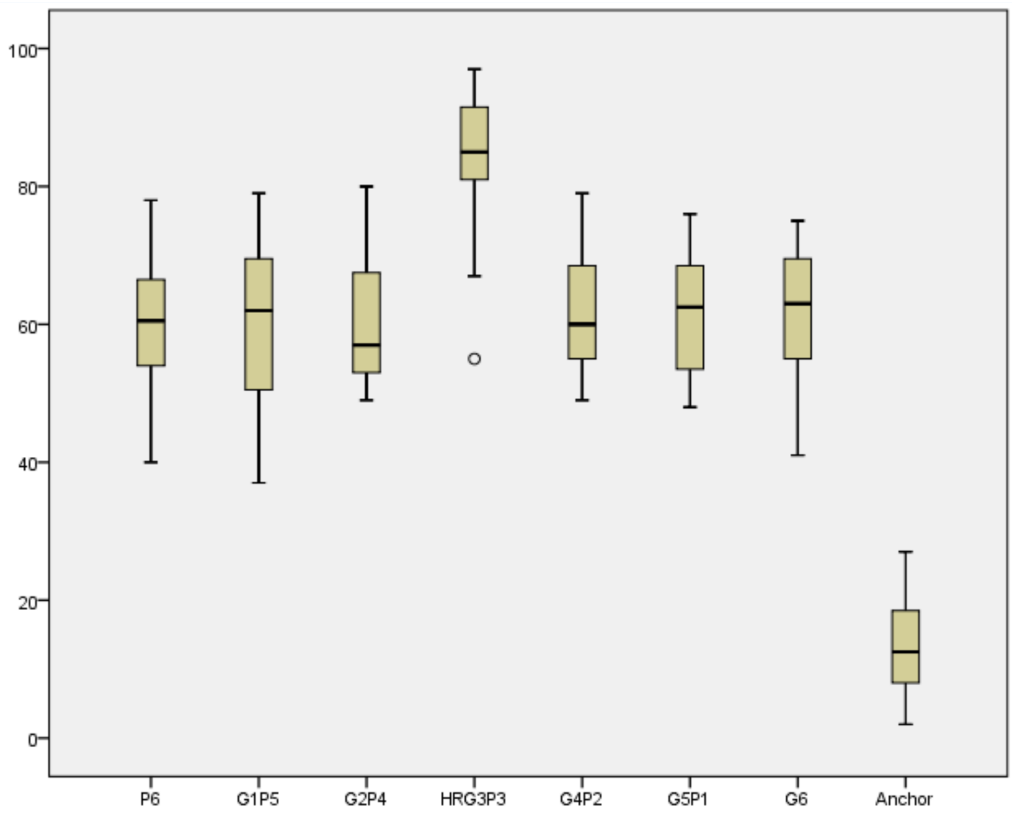


**Figure 7.23:** *Result of Case 5 in the MUSHRA listening test. The reference case used is G3P3 which is a mix of 3 vehicle pass-bys sounds synthesised using granular synthesis and 3 vehicle pass-bys sounds synthesised using procedural audio.*

which is created according to the rating data in Appendix C. As can be seen from Figure 7.24, the rating of the anchor is obviously lower than other stimuli, while the rating of the hidden reference is obviously higher than the other stimuli. By taking a Friedman test for the MUSHRA data from stimuli P6, G1P5, G2P4, G4P2, G5P1, and G6, it can be found that there is no significant difference between these stimuli in terms of plausibility at the 95% confidence level ($p = 0.88 > 0.05, \chi^2(5) = 1.76$). It reveals that the plausibility of the pure procedural audio model, a mix of granular synthesis for the engine sounds and the procedural audio model, and pure granular synthesis for the engine sound are comparable to each other under a traffic flow condition with various types of pass-by vehicles driving bidirectionally at a mix of low and high vehicle speeds, under a smooth flow rate.

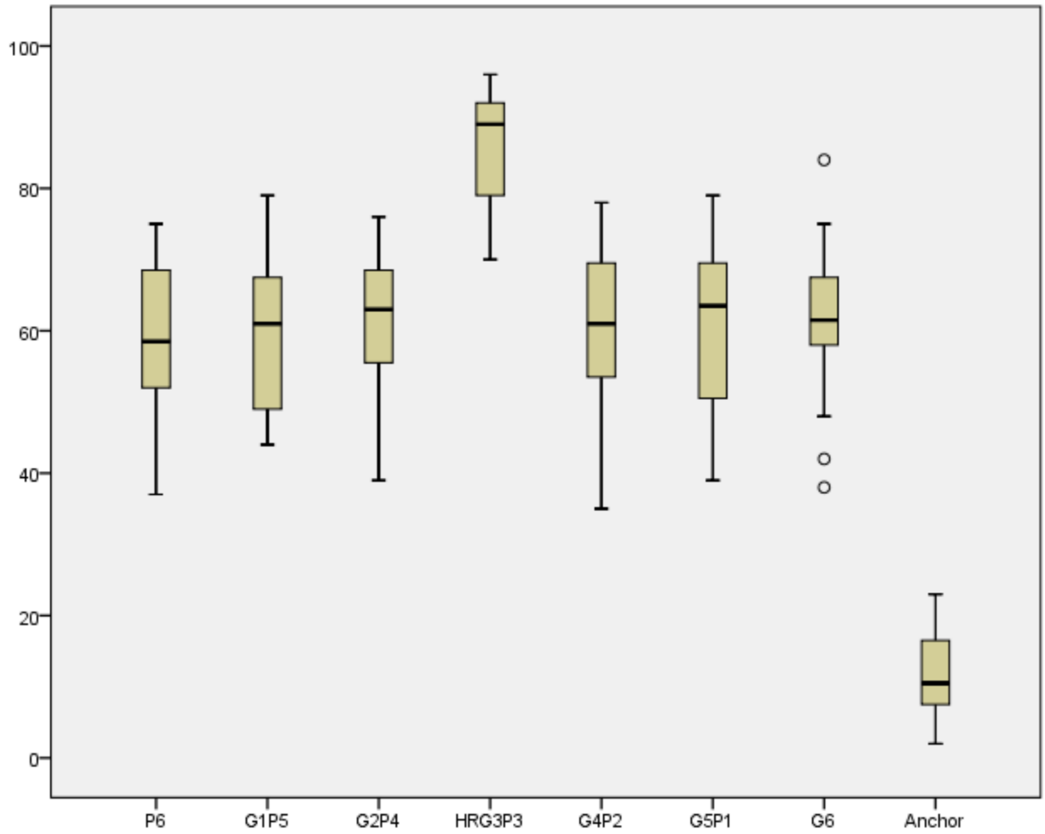


**Figure 7.24:** *Result of Case 6 in the MUSHRA listening test. The reference case used is G3P3 which is a mix of 3 vehicle pass-bys sounds synthesised using granular synthesis and 3 vehicle pass-bys sounds synthesised using procedural audio.*

Figure 7.25 demonstrates the MUSHRA result of Case 7 in a boxplot, which is created according to the rating data in Appendix C. As can be seen from Figure 7.25, the rating of the anchor is obviously lower than other stimuli, while the rating of the hidden reference is obviously higher than the other stimuli. By taking a Friedman test for the MUSHRA data from stimuli P6, G1P5, G2P4, G4P2, G5P1, and G6, it can be found that there is no significant difference between these stimuli in terms of plausibility at the 95% confidence level ($p = 0.72 > 0.05, \chi^2(5) = 2.86$). It reveals that the plausibility of the pure procedural audio model, a mix of granular synthesis for the engine sounds and the procedural audio model, and pure granular synthesis for the engine sound are comparable to each other under a traffic flow condition with various types of pass-by vehicles drive bidirectionally at a mix of low, mid and high vehicle speeds, under a mass flow rate which represents a street with heavy traffic.

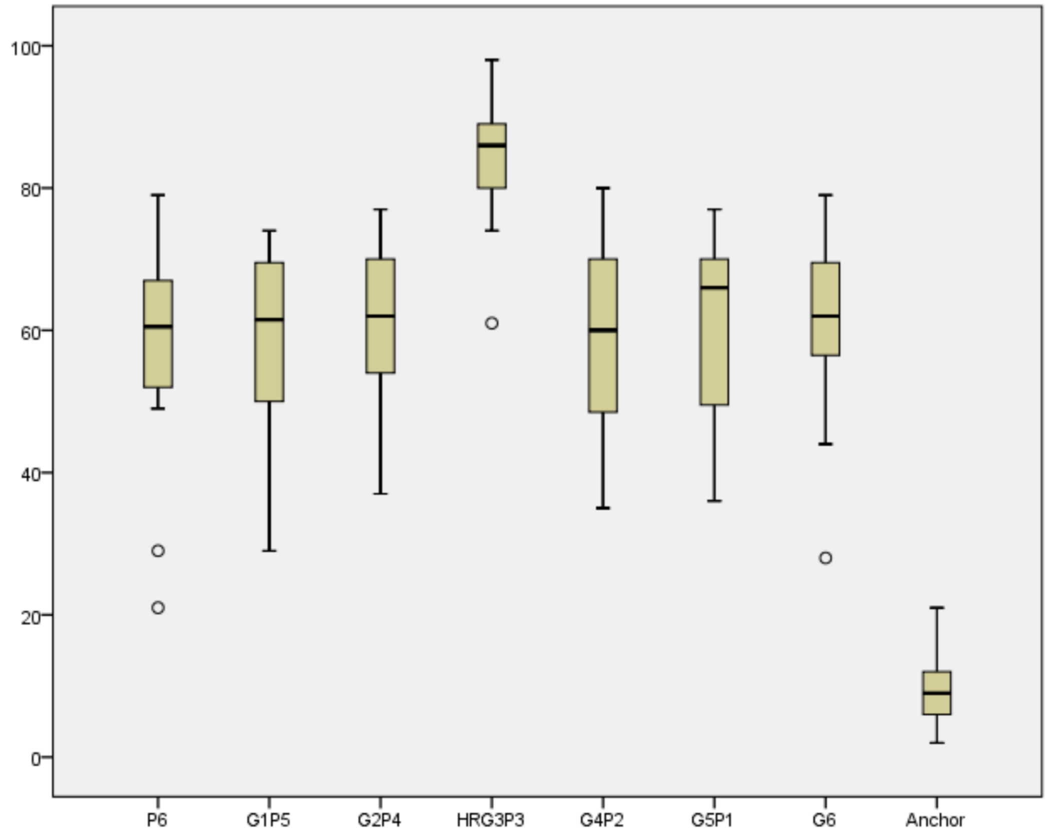


**Figure 7.25:** *Result of Case 7 in the MUSHRA listening test. The reference case used is G3P3 which is a mix of 3 vehicle pass-bys sounds synthesised using granular synthesis and 3 vehicle pass-bys sounds synthesised using procedural audio.*

### Discussion

The aim of this test was to evaluate the plausibility of traffic flow auralised by a proposed auralisation framework based on the use of a procedural audio method. This is conducted by comparing the plausibility with related examples synthesised using granular synthesis for the engine sound under a variety of traffic flow conditions in terms of vehicle type, vehicle speeds, driving directions, and flow rate, etc. The MUSHRA test method has been chosen to investigate to what extent that traffic flow sounds synthesised using the procedural audio model and the granular synthesis model are interchangeable without causing plausibility issues.

According to the statistical analysis of the different test cases corresponding to different traffic flow conditions presented here, it is found that there are no perceived differences in terms of plausibility when there is a rational variation on vehicle type, vehicle speeds, and flow rates for the traffic flow sounds. Here 'rational' means the driving parameters for every single vehicle pass-by sound should be within the requirements of the procedural audio model used here, e.g. a vehicle speed of 180km/h would violate the limitation of the Harmonoise model for tyre noise synthesis, which would be expected to decrease the plausibility of the auralised traffic flow sound dramatically.

Based on the statistical analysis on the results for all cases, it is found that subjective evaluation of plausibility is relatively sensitive to the driving directions for multiple pass-by vehicles. Statistically, significant differences emerge when all the vehicles drive in one direction, and the auralisations that rely on the use of more recording-based granular synthesis methods tends to be rated as more plausible.

To the best of the author's knowledge, there is no existing theory that provides a full explanation for this phenomenon. Based on the study presented in [334] showing that peripheral attention and central attention can be converted to some extent when auditory stimuli from different directions are perceived, one potential assumption could be that a person's central and peripheral attention might change in some way when listening to sounds moving along one direction when compared to bidirectional sounds, which leads to more attention paid to the specific timbre of the sounds rather than localisation cues. This assumption needs to be further proved/disproved using better designed experiments with more measurement data.

Another potential explanation worth considering is related to the variation of the ability to detect tones under different binaural listening conditions. In Chapter 2, the concept of the BMLD has been discussed, which might be related to the findings in this listening test. However, most prior studies used only static signal and masker sources to explore the threshold of signal detection at different localisation patterns. In this study, both the engine sound and the tyre noise come from the same direction simultaneously, but are moving according to the driving patterns of the vehicle. As the rich tonal components in the engine sound are masked by the noise components in both the procedural audio model and the granular synthesis model, but to a different extent, it can be assumed that there might be a variation in the thresholds of tone-in-noise detection for the different MUSHRA test cases, which leads to the differences in the perceived plausibility. Moreover, a moving source consisting of engine sound and tyre noise might be masked by another moving source with some variations in timbre. These assumptions related to the BMLD need to be further investigated with more measurement data to draw robust conclusions.

## 7.4 Summary

In this chapter, the plausibility of the auralisation framework developed in this thesis has been evaluated by two listening tests. First, the plausibility of a single vehicle pass-by has been validated by an ABX listening test. The results have indicated that a similar level of 'plausibility to expectation' can be achieved by using either the procedural audio model or the granular synthesis model for engine sound synthesis in the proposed auralisation framework, noting that there are significant differences in the first steps of building up the complete auralisation framework, such as engine sound synthesis and combining the engine sound and tyre noise. This indicates that an auralisation framework can be adjusted systematically and globally to achieve plausible sounds rather than focusing purely on a specific component for which the plausibility might be considered limited.

Based on the validation of the proposed auralisation framework for single vehicle pass-bys, a MUSHRA test has been implemented to evaluate the plausibility of traffic flow auralised using the proposed auralisation framework. As it is not easy or realistic to obtain real recordings for the external references corresponding to the specific traffic flow conditions to be tested, an investigation has been conducted to find the impact of plausibility evaluation when different stimuli are used as the external references. The results have indicated that the MUSHRA ratings remain stable no matter which external reference is used. The plausibility has then been tested in a variety of traffic flow conditions, with the variation of vehicle type, vehicle speeds, driving directions, and flow rates, etc. The results have verified that the plausibility of the auralisation framework using procedural audio is comparable to a granular synthesis model for the engine sound under a wide range of traffic flow conditions in terms of vehicle type, vehicle speeds, and flow rates, etc. However, the plausibility of the pure procedural audio model tends to be weaker than that using a recording-based synthesis model for traffic flow examples based on only a single one-way driving direction.

# Chapter 8

# Conclusion

This thesis has presented a portfolio of research into the auralisation of traffic noise using procedural audio methods, with the goal of developing a framework that can generate traffic flow auralisation for a microscopic urban scene with a suitable level of plausibility and flexibility. As auralisation has been widely used for demonstration purposes and VR applications in the context of room acoustics, the premise behind this thesis is that the application of this technique can be extended into outdoor environments, which would be useful for designers, urban planners, transportation planners, and the public who are interested in understanding how sounds are perceived, and what the associated impacts might be for their interactions with their sound environment. Before drawing the main conclusions from this study (presented in Section 8.3), a summary of the thesis work (Section 8.1) and re-statement of the hypothesis (Section 8.2) as introduced in Chapter 1 will be presented. The main conclusions for the thesis and contributions to the relevant fields will be summarised, followed by the consideration of potential research directions and topics in the future (Section 8.4).

## 8.1 Summary

The thesis begins with Chapter 2 covering the fundamentals of acoustic theory and providing a basic understanding of the relevant aspects for the development and evaluation of an auralisation framework for traffic noise. This includes sound generation mechanisms, sound propagation effects in an outdoor environment, and the relevant acoustic and psychoacoustic theory of sound perception and evaluation. Following the basic acoustic theory, the concepts of auralisation, procedural audio, and environmental sounds are presented in Chapters 3, 4, and 5, respectively, alongside an overview of the main methodologies used in each field. The purposes of such a wide range of overview on these related fields include:

- Investigating the requirement and context of developing such a traffic noise auralisation framework using procedural audio methods.
- Exploring potential solutions to integrate the concepts and methodologies in each field in order to create an auralisation framework.
- Estimating what can be potentially achieved by combining techniques from these relevant fields.

Specifically, Chapter 3 covered the fundamentals of auralisation, including the background theory on treating airborne sound auralisation problems as LTI systems, three essential elements within an auralisation framework consisting of sound source modelling, sound propagation modelling and spatial audio reproduction, in addition to considerations on the applications of auralisation in the field of room acoustics, road traffic noise, and VR/AR systems. Special attention has been given to prior studies on the auralisation of road traffic noise, and in particular to different methodologies for sound source modelling and the simulation of sound propagation effects. A categorisation method consisting of micro-, meso-, and macro-scopic models has been introduced, which clusters the auralisation models for road traffic noise according to the levels of detail required to build up a suitable framework for the problem of this nature. It is useful to develop a holistic understanding of how to establish such an appropriate auralisation framework, and what should be involved for the auralisation of road traffic problems.

Chapter 4 introduced the concepts and characteristics of procedural audio. The advantages and disadvantages of procedural audio have been discussed and compared to recording-based audio techniques. As algorithms are the core assets for procedural audio models, an overview of some typical sound synthesis methods is then presented, including additive synthesis, subtractive synthesis, spectral modelling synthesis, physical modelling synthesis, wavetable synthesis, and FM synthesis. Each synthesis method has its pros and cons, so it is necessary to choose appropriate sound synthesis methods and integrate them accordingly when implementing procedural audio algorithms for different purposes. Although procedural audio has been utilised in some video games and soundscape studies for the creation of sound effects and ambient sounds for interactive applications, it is worth exploring its potential for auralisation purposes in a broader context. Its flexibility and interactivity can be effectively used for the auralisation of other applications such as urban planning or to enhance the listeners' perception of a specific acoustic environment through the use of audio demonstrations.

Chapter 5 has covered the topic of environmental sounds, including the underlying aspects of the definition, categorisation, impacts, evaluation and prediction methods, and the use of soundscape concepts and methodologies in public's engagement. The potential for auralisation in supporting conventional noise evaluation/prediction methods and soundscape studies has been discussed, which provides a perspective for the requirement and application context for the development of a traffic flow auralisation framework that is both flexible and interactive.

Based on the overview of the relevant fields in Chapters 3, 4, and 5, it is reasonable and meaningful to develop an auralisation framework for environmental sounds using procedural audio approaches, such as traffic noise auralisation, which offers a rational compromise between plausibility, flexibility and interactivity.

The methodologies for developing such a framework and the key factors to be considered for sound synthesis and modelling methods have been documented in Chapter 6. This chapter started with the development of an auralisation model for a single vehicle pass-by sound using procedural audio methods. Engine sound synthesis is realised using a physical-based synthesis model, and tyre noise synthesis is implemented using an engineering model derived from the Harmonoise project [119]. Sound propagation effects including distance attenuation, Doppler shift, and early reflections using $1^{st}$-order image sources have been introduced, which are considered to have an impact on the plausibility of the auralisation for a microscopic urban scene. This auralisation is presented binaurally with HRTF processing. Traffic flow auralisation is achieved by summing the auralised single vehicles according to the descriptive parameters of traffic flow dynamics. The auralisation framework is built in Unity3D as the hosting platform. The sound source modelling is developed in Pure Data, and the sound propagation effects are implemented in Wwise with some simulations in Matlab. The parameter transmission between different platforms is realised using open-source middleware or plug-ins working across Unity3D, PD, and Wwise.

Chapter 7 has presented the evaluation of the plausibility achieved by the auralisation framework developed in Chapter 6. Two listening tests have been conducted to investigate to what extent the auralised vehicle pass-by and traffic flow sounds might be considered as being plausible. Granular synthesis has been used as a recording-based audio synthesis technique for creating the counterparts for comparison. It is assumed that if there is no perceived difference between a sound scene auralised using procedural audio and recording-based audio, the procedural audio model can be considered sufficiently plausible for rendering the results. By conducting an ABX test, it is found that the perceived plausibility of the procedural audio model is comparable to the granular synthesis approach when reproduced binaurally with sound propagation effects added, although the pure engine sounds synthesised using these two methods are perceived as being different. With this validation of single vehicle pass-by sounds, a MUSHRA test has been conducted to evaluate the plausibility of traffic flow auralised using the proposed auralisation framework. It is found that the plausibility of traffic flow auralisation using a procedural audio model is comparable to counterparts auralised using a granular synthesis model under a wide range of traffic flow conditions in terms of vehicle type, vehicle speeds, and flow rates. However, for the case of all traffic flow driving in one direction, the plausibility of the procedural audio model tends to be weaker than the recording-based synthesis model.

## 8.2 Restatement of Hypothesis

The hypothesis that has informed this research as summarised in Chapter 1, is revisited here again as follows:

> *The auralisation of traffic flow sounds using a procedural audio approach is comparable to methods based on recorded or sampled audio when considering the plausibility of the results obtained.*

In order to prove or disprove this hypothesis, two steps should be taken: 1) Finding a way to develop a traffic flow auralisation model using procedural audio approaches, and 2) Evaluating the plausibility of the auralisation model using appropriate methods. In this thesis, these two steps have been taken as follows:

- A traffic flow auralisation model has been designed and implemented after in-depth overview of the relevant fields. The engine sound is synthesised by a physical-based synthesis model, while tyre noise is synthesised by an engineering model based on empirical equations, both based on procedural audio methods. Following the principles of auralisation, a sound propagation model for a micro-scopic urban environment has been developed, and the auralised traffic flow scenes are processed via HRTF processing for spatial audio reproduction. A variety of traffic flow sounds corresponding to different traffic flow conditions can be auralised in this way. The design and development of such a traffic flow auralisation model can be viewed as strong evidence supporting the first step to prove the research hypothesis.

- The proposed auralisation framework has been validated by a series of listening tests, in which the plausibility of the results have been compared to an approach that replaces the procedural audio engine sound model with one based on recording-based granular synthesis. The results from such a subjective evaluation process demonstrate that the plausibility of the developed procedural audio model is comparable to the granular synthesis model and verified that: a) a similar level of 'plausibility to expectation' can be achieved by using either of these two models for engine sound synthesis, and b) there is no significant difference between the audible perception of the traffic flow auralised by these two methods for a variety of traffic flow conditions. The results from these listening tests provide solid evidence in favour of the second step to prove the research hypothesis

These two points demonstrate the main value of this study, and clearly confirm the research hypothesis proposed in this thesis. The novelty and main contributions to the relevant fields throughout this process will now be considered.

## 8.3 Main Contributions to the Field

The research that has been completed in the presentation of this thesis has resulted in the following contributions to the relevant fields as summarised below:

- **Development of a procedural audio model for auralisation purposes**. Prior to this study, procedural audio has mainly been used by sound designers in video games and animations, while auralisation has been mainly implemented by acoustic consultants and researchers in room acoustics projects for the purpose of demonstrating and disseminating relevant results. The rapid development of VR technology and Human-Computer Interaction (HCI) in more recent years has resulted in a wider range of potential application areas and contexts for developing more flexible auralisation tools for different types of acoustic environments. The idea of using procedural audio methods for auralisation purposes as proposed in this thesis tries to bridge the gap to some extent. An auralisation framework for traffic noise via procedural audio methods has been developed. The plausibility has been evaluated by comparing the proposed model with a more established and acceptable recording-based method. On one hand, this multi-disciplinary approach may extend the application of procedural audio into a broader context moving away from game design and animation films. On the other hand, it provides a feasible way to improve both flexibility and interactivity in auralisation projects implemented by acoustic consultants and researchers.

  Such a flexible and interactive auralisation tool can be useful in many ways. As discussed in Chapter 5, the most popular use of such an interactive auralisation is demonstrating some large-scale construction projects to enhance public engagement, such as the projects reported in [249] and [218]. In these existing projects, recordings were used for sound source modelling of vehicle pass-by noise, which have very limited flexibility. It is worth trying to use such an auralisation model based on procedural audio approaches to enhance the flexibility and interactivity for these demonstrating tools. Apart from demonstration purposes, the proposed traffic flow auralisation method can also be useful for internal communication as part of the design process. Plausible acoustic scenes corresponding to a variety of traffic flow conditions can be created by tuning the procedural audio model. It is useful for internal communication with other team members and stakeholders of the project, in which multiple rounds of modifications are always needed. With such a flexible auralisation tool, it is possible to create the corresponding acoustic scenes when the design of the roads changes, by tuning the auralisation model according to the modified traffic flow conditions. This will be helpful for urban planning and traffic management from an acoustic perspective.

- **Development and application of auralisation for traffic flow scenes**. Auralisation is typically used for room acoustics to demonstrate the perceived audible experience of a space. Although there have been several studies on the application of auralisation for outdoor sound events and outdoor environments, such as wind turbine noise [291], aircraft noise [110, 253], railway noise [111, 256], and pass-by vehicles [112, 117, 126, 252, 262, 278–280, 288], there has been little literature on traffic flow auralisation. This is mainly because a large dataset consisting of audio samples from a wide range of vehicle types, vehicle speeds, and driving patterns is required for creating a variety of traffic flow scenes. This kind of dataset is not always accessible, nor easy to use, particularly in applications where the computational and storage resources for audio signal processing are quite limited (e.g. on a mobile platform or for VR applications). By taking advantage of procedural audio that demonstrates both flexibility and variable computational cost, it is convenient to obtain various synthetic pass-by sounds representing different vehicle types, vehicle speeds, and driving patterns by appropriately tuning a procedural audio model without the additional issues that arise through using recording-based methods. The novelty of this work contributes to the field of traffic noise auralisation by providing a new method for the auralisation of complex traffic flow scenes in a concise and efficient way. The plausibility of the proposed method has been validated by listening tests under a variety of vehicle pass-by scenarios and traffic flow scenes.

- **Development of a complete auralisation framework for vehicle pass-by sounds**. Prior to this thesis, there has been very little study on the development of a complete auralisation framework for pass-by vehicles, including the aspects of sound source modelling, sound propagation modelling, and spatial audio reproduction. Instead, most related traffic noise auralisation studies focus on a single component within such a framework (e.g. road-tyre noise [312], engine sound [266, 335], etc.), and use quantitative or qualitative studies for that specific component. Notwithstanding, it can be extremely complicated and time-consuming to find 'perfect' solutions for every component in a complete auralisation framework. In this thesis, the novelty leans on developing a complete auralisation framework for pass-by vehicles from a more holistic perspective, which focuses on the plausibility of traffic flow for a 'micro-scopic' urban scene as a whole, rather than the plausibility of a specific sound component for a single vehicle. The results from the listening tests in Chapter 7 reveal that the auralisation of traffic flow can be still plausible even when there are significant perceived differences to certain aspects or sub-components developed within the auralisation framework.

- **Validation of the plausibility of the proposed auralisation framework for traffic flow by subjective evaluation**. Prior to this thesis, there had been no published study on the subjective evaluation of plausibility for traffic flow auralisation. Some relevant studies on the evaluation of auralised single vehicle pass-bys [311, 313] or sub-components (e.g. road-tyre noise [312], engine sound [266, 335], etc.) in a moving car can be found, but their results and conclusions cannot be directly applied into traffic flow scenes because sounds emitted from multiple vehicles interact with each other, and with the surroundings, which may have an impact on the audible perception of both the sound event, and the sound environment. In this thesis, an ABX listening test has been designed and implemented for validating the plausibility of auralised single vehicle pass-bys, and a MUSHRA listening test has been designed and implemented for evaluating the plausibility of the proposed traffic flow auralisation model. The methodologies used in the experimental design and the results obtained from these subjective evaluation data can be taken as reference for other studies on evaluating or validating other auralisation work for traffic flow scenes.

- **Combination of different platforms for the implementation of auralisation with flexible and interactive features**. This brings together some existing techniques as used in different platforms and takes advantage of them selectively. This combination includes using Unity3D as the hosting platform for the sound event rendering, Pure Data as the procedural audio tool for source modelling, and Wwise and Matlab as the sound propagation effects simulation. The parameter transmission between different platforms is achieved by open-source middleware or plug-ins. Prior to this thesis, the combination of these audio-related tools has not yet been attempted in this application area, and what could be achieved was as yet unknown. The procedural audio auralisation tools and methods developed in this thesis for the first time show that these platforms can be combined for multi-disciplinary and complex tasks, such as rendering plausible auralisation of traffic flow in a micro-scopic urban scene with flexible and interactive features.

## 8.4 Future Work

There are many areas for future research on the basis of the findings and contributions in this thesis, particularly in the fields of auralisation, environmental sound evaluation, and procedural audio. The following suggestions on future steps are quite broad, depending on the purpose of any follow on study and how it relates to different areas.

**Evaluating the auralisation model by comparison with in-situ recordings**

In this thesis, the plausibility of the developed traffic flow auralisation framework has been evaluated by comparison with a granular synthesis model. Although granular synthesis is a recording-based audio technique and has been widely used in the game audio and music production industries, it is still a sound synthesis technique that is not physically identical to in-situ recordings. A possible future step could be conducting subjective listening tests to compare the developed auralisation model to in-situ recordings in terms of plausibility. An open question would be how to acquire appropriate data for in-situ recordings. On one hand, it might be not easy or possible to record pure traffic flow sounds without contamination by other sound events (e.g. natural sounds, human sounds) in an outdoor environment. For instance, it is difficult to record pure traffic flow sounds along a busy street in a city centre because sounds emitted from human activities may always be detectable in the background. On the other hand, there might be limitations to building up and running a procedural audio model for every single vehicle within a traffic flow scenario due to the lack of data related to the structure, configurations, and driving patterns for all of the vehicle models. Therefore, it is worth exploring how to take appropriate traffic flow recordings and collect data for developing better defined procedural audio models for auralising such sound events for different purposes.

### Evaluating traffic flow auralisation with visual cues

Audiovisual cross-modal perception has been proved to be crucial for an understanding of an environment's soundscape according to a series of relevant studies [140, 141, 336–339], and it is reported that when audio-visual perceptions are coupled, attention paid to the visual cues may change the conscious perception of sound, and vice versa [174]. For example, the impact of the perception of a sound event may be influenced by visibly presenting vegetation (e.g. trees, grass) although they may have very little influence on the acoustic properties of the sound (e.g. SPL) [338, 340], or even by changing the colour of a visual component, with no associated change in any related acoustic property [141]. As discussed in Chapter 6, a visual scene has been implemented in Unity3D in this study but has not been used for listening tests in the evaluation stage. In future research, it is worth exploring to what extent the presence of visual features can affect the perception of the auralised traffic flow, and how to integrate the audiovisual tools for rendering more realistic traffic flow scenes.

Based on the validation of plausibility of the proposed traffic flow auralisation framework, this auralisation method may be used to render acoustic scenes for lab-based audio-visual perception experiments of soundscape studies regarding road traffic noise. It would be convenient to create a series of traffic flow sounds according to a variety of vehicle driving conditions by tuning such a procedural audio model. As the visual scene is modelled via the game engine Unity3D, it is also convenient to modify the landscape components and features (e.g. naturalness, brightness, tranquility, etc.) by changing the game objects and graphical rendering methods. By combining the visual modelling techniques in a game engine and the traffic flow auralisation based on procedural audio approaches, a highly flexible and interactive platform for audio-visual perception of traffic noise can be established, which provides more flexible experimental settings for soundscape studies. Such a flexible platform would contribute to soundscape studies in many ways. For instance, it has been reported that VR with 360degree field-of-view photographic images are more apt for evaluating existing environments, whereas VR models are more suitable for evaluating modified environments [341]. With such a flexible audio-visual rendering tool, it would be feasible to design qualitative and quantitative experiments to explore the differences between the recording-based methods (VR photographs for visual rendering with in-situ recordings for acoustic rendering) and the model-based methods (VR models in a game engine for visual rendering with procedural audio approaches for acoustic rendering), in terms of road traffic noise in soundscape studies.

#### Investigating the real-time performance of the auralisation model

As discussed in Chapter 4, a procedural audio model should have the characteristic of real-time performance, which supports its suitability for interactive usage, e.g. rendering acoustic scenes for VR/AR applications. Real-time performance has not been investigated in this thesis, as plausibility was the main area to be investigated rather than latency. Potential future research would be investigating the computation time for rendering more complex traffic flow scenes, e.g. a larger number of pass-by vehicles in a traffic flow scene, or testing the latency caused by different sound synthesis or sound propagation algorithms used for the auralisation framework. For example, in this study, only $1^{st}$-order image sources have been implemented for the early reflections at the sound propagation modelling stage. It is worth exploring some higher orders of reflections in terms of the trade-off between the real-time performance and the plausibility of the auralised sounds. Some algorithms and data structures might be considered to improve the efficiency of the image source algorithm, such as the binary space partitioning method proposed in [72].

### Exploring objective metrics for the evaluation of plausibility in traffic flow soundscapes

In this study, the plausibility of the developed auralisation model has been evaluated using subjective listening tests. Although subjective evaluation has been widely used in most auralisation studies, it is worth exploring objective metrics representing the plausibility of auralised sounds to some extent. On one hand, it would be time-saving and cost-effective to evaluate the plausibility of auralisation by calculating metrics, without resorting the use of listening tests. On the other hand, it would be convenient to compare the plausibility of the output sounds rendered by different auralisation models. Some studies on soundscape indices [234, 342, 343] and sound quality metrics [29, 152, 233] can be taken for reference as a start for future research in this area.

### Integrating the auralisation model into other areas for acoustic rendering

There are many potential application scenes for traffic flow auralisation with procedural audio methods. The auralisation model should be further tailored based on the purposes of auralisation in different applications, with a variety of improvements and modifications to be conducted. For example, when used for urban planning, the sound propagation model may be improved to be more appropriate for rendering complex urban scenes for urban design. For traffic planning, it is worth considering a more effective way for rendering far-field traffic noise in addition to the developed auralisation model for single vehicle pass-bys because when a vast number of vehicles are involved, the far-field pass-by sounds may be perceived as background noise to some extent [122]. For sound design in video games, it is important to take advantage of the flexibility and variable computational cost of procedural audio methods in this auralisation framework to create a variety of traffic flow scenes corresponding to the game context with as little storage space as possible.

In this study, traffic noise has been chosen as the sound scene to be auralised using procedural audio methods. As discussed in Chapter 6, this is mainly because it is not feasible to get a large number of required recordings corresponding to different vehicle driving patterns using conventional recording-based auralisation methods. An interesting topic worth exploring is whether such an auralisation framework based on procedural audio approaches is suitable for other environmental sound sources, such as bird sounds, insect sounds, etc. In theory, every sound can be auralised following the principles of auralisation by doing sound source modelling, sound propagation modelling, and spatial audio reproduction. However, there are two important issues to be considered: 1) whether procedural audio is suitable for modelling the target sound source, and 2) how accurate the auralisation should be for the target sound scene. As discussed in Chapter 4, the advantages of procedural audio are its variety and flexibility, while aesthetic issues might be an inherent drawback that should be treated carefully. Therefore, when modelling a sound source, it is crucial to explore what the extent of the sound source model's variety and flexibility should be. If a high level of flexibility is required, it is then worth considering if recording-based approaches are suitable for the context to be auralised. For example, for bird sounds, when a variety of different types of bird sounds are required and needed to be controlled for interactive auralisation, procedural audio approaches, such as the physical-based models proposed in [13], can be viewed as an alternative because it is difficult to take all the required recordings to reach a high level of variety and flexibility. As there are numerous kinds of environmental sound sources and sound scenes to be auralised, the suitability and feasibility of procedural audio should be judged case by case. In terms of the accuracy of the auralisation, there is still a lack of research on the relationship between the physical accuracy of acoustics and the perceived realism. As human perception is multisensory, the perceived realism of an auralisation may vary according to the changes of attention on the audible cues, varying from peripheral attention to central attention [24, 142, 334]. Therefore, subjective listening tests are indispensable for evaluating the perceived accuracy of the auralisation, by which useful cues may be obtained to simplify the models in the auralisation framework.

The procedural audio based auralisation approach proposed in this thesis can be potentially tailored for some interactive VR/AR applications. As the available computation time is quite limited for VR/AR demonstrating (the latency between visual and audio cues should be typically below 60ms [7]), some improvements in the complexity of the algorithms used in this auralisation framework are worth considering, so as to achieve a ‘smooth’ user experience. It is also necessary to have a more holistic understanding of what might be considered as an acceptable compromise on the plausibility of the final results, according to the specific context of the VR/AR demonstration.

# Bibliography

[1] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of acoustics.* John Wiley & Sons, 1999.

[2] D. I. für Normung, "Measurement and assessment of low-frequency noise immissions," *German Standard*, 1997.

[3] S. Harriet and D. Murphy, "Auralisation of an urban soundscape," *Acta Acustica united with Acustica*, vol. 101, no. 4, pp. 798–810, 2015.

[4] S. Oxnard, "Efficient hybrid virtual room acoustic modelling," Ph.D. dissertation, University of York, 2016.

[5] S. Harriet, "Application of auralisation and soundscape methodologies to environmental noise," Ph.D. dissertation, University of York, 2013.

[6] C. Peckens and J. P. Lynch, "Utilizing the cochlea as a bio-inspired compressive sensing technique," *Smart Materials and Structures*, vol. 22, no. 10, pp. 105–027, 2013.

[7] M. Vorländer, *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality.* Springer Science & Business Media, 2007.

[8] F. Stevens, "Strategies for environmental sound measurement, modelling, and evaluation," Ph.D. dissertation, University of York, 2018.

[9] L. M. Wang and C. B. Burroughs, "Directivity patterns of acoustic radiation from bowed violins," *Catgut Acoustical Society Journal*, vol. 3, no. 7, pp. 7–15, 1999.

[10] B. Pueo, J. V. Rico, and J. J. Lopez, "Vibration analysis of edge and middle exciters in multiactuator panels," in *Audio Engineering Society Convention 131.* Audio Engineering Society, 2011.

[11] G. Zalles, Y. Kamel, I. Anderson, M. Y. Lee, C. Neil, M. Henry, S. Cappiello, C. Mydlarz, M. Baglione, and A. Roginska, "A low-cost, high-quality mems ambisonic microphone," in *Audio Engineering Society Convention 143.* Audio Engineering Society, 2017.

[12] J. O. Smith, *Physical Audio Signal Processing.* http://ccrma.stanford.edu/~jos/pasp/, 2010, online Book.

[13] A. Farnell, *Designing sound.* Mit Press, 2010.

[14] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[15] W. W. Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[16] I. ISO, "Acoustics — soundscape — part 1: Definition and conceptual framework," *International Organization for Standardization, Geneva, Switzerland*, 2014.

[17] J. Kang and M. Zhang, "Semantic differential analysis of the soundscape in urban open public spaces," *Building and environment*, vol. 45, no. 1, pp. 150–157, 2010.

[18] M. Fellendorf and P. Vortisch, "Microscopic traffic flow simulator vissim," in *Fundamentals of traffic simulation.* Springer, 2010, pp. 63–93.

[19] udaix, *Engine Four Stroke Cycle infographic diagram including stages of intake compression power and exhaust showing parts and valves open and closed for mechanical physics science education.* Shutterstock, 2007. [Online]. Available: https://www.shutterstock.com/zh/image-vector/engine-four-stroke-cycle-infographic-diagram-707664295

[20] www.howacarworks.com, *engine-with-overhead-camshaft.* www.howacarworks.com, 2020. [Online]. Available: https://www.howacarworks.com/illustrations/engine-with-overhead-camshaft

[21] Sashkinw, *V4 engine pistons and crankshaft on white background.* www.dreamstime.com, 2017. [Online]. Available: http://www.dreamstime.com/royalty-free-stock-photography-\v-engine-pistons-crankshaft-white-background-d-render-image73970307

[22] A. Mitiuc, *V4 pistons and cog isolated on white.* www.dreamstime.com, 2016. [Online]. Available: https://www.dreamstime.com/stock-illustration-v-pistons-cog-isolated-white-image53758941

[23] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and service robotics.* Springer, 2018, pp. 621–635.

[24] T. W. Picton, S. A. Hillyard, R. Galambos, and M. Schiff, "Human auditory attention: a central or peripheral process?" *Science*, vol. 173, no. 3994, pp. 351–353, 1971.

[25] K. Bijsterveld *et al.*, *Mechanical sound: Technology, culture, and public problems of noise in the twentieth century.* MIT press, 2008.

[26] E. Directive, "Directive 2002/49/ec of the european parliament and the council of 25 june 2002 relating to the assessment and management of environmental noise," *Official Journal of the European Communities, L*, vol. 189, no. 18.07, p. 2002, 2002.

[27] D. M. Howard and J. Angus, *Acoustics and psychoacoustics.* Taylor & Francis, 2017.

[28] S. Bech and N. Zacharov, *Perceptual audio evaluation-Theory, method and application.* John Wiley & Sons, 2007.

[29] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models.* Springer Science & Business Media, 2013.

[30] B. C. Moore, *An introduction to the psychology of hearing.* Brill, 2012.

[31] J. Blauert, "Sound localization in the median plane," *Acta Acustica united with Acustica*, vol. 22, no. 4, pp. 205–213, 1969.

[32] B. C. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The journal of the acoustical society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[33] I. E. Commission *et al.*, "Electroacoustics—sound level meters—part 1: Specifications (iec 61672-1)," *Geneva, Switzerland*, 2013.

[34] D. Huron, *Voice leading: The science behind a musical art.* MIT Press, 2016.

[35] A. M. Mayer, "Researches in acoustics," *American Journal of Science*, no. 277, pp. 1–28, 1894.

[36] B. C. Moore, J. I. Alcántara, and T. Dau, "Masking patterns for sinusoidal and narrow-band noise maskers," *The Journal of the Acoustical Society of America*, vol. 104, no. 2, pp. 1023–1038, 1998.

[37] T. Kim, S. Lee, and H. Lee, "Characterization and quantification of luxury sound quality in premium-class passenger cars," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 223, no. 3, pp. 343–353, 2009.

[38] D. Moore, R. Currano, and D. Sirkin, "Sound decisions: How synthetic motor sounds improve autonomous vehicle-pedestrian interactions," in *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2020, pp. 94–103.

[39] E. A. Clendinning, "Driving future sounds: imagination, identity and safety in electric vehicle noise design," *Sound Studies*, vol. 4, no. 1, pp. 61–76, 2018.

[40] S. A. Gelfand, *Hearing: An introduction to psychological and physiological acoustics.* CRC Press, 2016.

[41] K. Iida, M. Yairi, and M. Morimoto, "Role of pinna cavities in median plane localization," *Proc. 16th Int'l Cong. on Acoust*, pp. 845–846, 1998.

[42] J. F. Culling and H. S. Colburn, "Binaural sluggishness in the perception of tone sequences and speech in noise," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 517–527, 2000.

[43] L. A. Jeffress, H. C. Blodgett, T. T. Sandel, and C. L. Wood III, "Masking of tonal signals," *The Journal of the Acoustical Society of America*, vol. 28, no. 3, pp. 416–426, 1956.

[44] M. van der Heijden and P. X. Joris, "Interaural correlation fails to account for detection in a classic binaural task: Dynamic itds dominate n0s$\pi$ detection," *Journal of the Association for Research in Otolaryngology*, vol. 11, no. 1, pp. 113–131, 2010.

[45] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[46] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, "Auralization-an overview," *Journal of the Audio Engineering Society*, vol. 41, no. 11, pp. 861–875, 1993.

[47] N. Xiang and J. Blauert, "Binaural scale modelling for auralisation and prediction of acoustics in auditoria," *Applied Acoustics*, vol. 38, no. 2-4, pp. 267–290, 1993.

[48] M. Barron and C. Chinoy, "1: 50 scale acoustic models for objective testing of auditoria," *Applied Acoustics*, vol. 12, no. 5, pp. 361–375, 1979.

[49] J.-D. Polack, X. Meynial, and V. Grillon, "Auralization in scale models: Processing of impulse response," *Journal of the Audio Engineering Society*, vol. 41, no. 11, pp. 939–945, 1993.

[50] J. H. Rindel, "Modelling in auditorium acoustics. from ripple tank and scale models to computer simulations," *Revista de Acústica*, vol. 33, no. 3-4, pp. 31–35, 2002.

[51] M. Schroeder, B. Atal, and C. Bird, "Digital computers in room acoustics," *Proc. 4th ICA, Copenhagen M*, vol. 21, 1962.

[52] M. R. Schroeder, "Computer models for concert hall acoustics," *American Journal of Physics*, vol. 41, no. 4, pp. 461–471, 1973.

[53] C. Pösselt, J. Schroeter, H. Opitz, P. Divenyi, and J. Blauert, "Generation of binaural signals and home entertainment," *Proc. 12th ICA, Toronto*, 1986.

[54] G. M. Naylor, "Odeon—another hybrid room acoustical model," *Applied Acoustics*, vol. 38, no. 2-4, pp. 131–143, 1993.

[55] B. Dalenbäck, "Catt-acoustic," 2002.

[56] W. Ahnert and R. Feistel, "Ears auralization software," in *Audio Engineering Society Convention 93*. Audio Engineering Society, 1992.

[57] M. Noisternig, B. F. Katz, S. Siltanen, and L. Savioja, "Framework for real-time auralization in architectural acoustics," *Acta Acustica United with Acustica*, vol. 94, no. 6, pp. 1000–1015, 2008.

[58] D. Schröder and M. Vorländer, "Raven: A real-time framework for the auralization of interactive virtual environments," in *Forum Acusticum.* Aalborg Denmark, 2011, pp. 1541–1546.

[59] A. Oliveira, G. Campos, P. Dias, D. T. Murphy, J. Viera, C. Mendonça, and J. Santos, "Real-time dynamic image-source implementation for auralisation," in *Proceedings of the 16th International Conference on Digital Audio Effects.* York, 2013, pp. 368–372.

[60] R. Gupta, B. Lam, J. Hong, Z. Ong, W. Gan, S. H. Chong, and J. Feng, "3d audio ar/vr capture and reproduction setup for auralization of soundscapes," in *Proceedings of the 24th Intl. Congress on Sound and Vibration, ICSV24*, 2017.

[61] J. R. Higgins *et al.*, *Sampling theory in Fourier and signal analysis: foundations.* Oxford University Press on Demand, 1996.

[62] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 856–865, 2008.

[63] A. Lindau and S. Weinzierl, "Assessing the plausibility of virtual acoustic environments," *Acta Acustica united with Acustica*, vol. 98, no. 5, pp. 804–810, 2012.

[64] J. Blauert, *Communication acoustics.* Springer, 2005, vol. 2.

[65] J. Y. Hong, J. He, B. Lam, R. Gupta, and W.-S. Gan, "Spatial audio for soundscape design: Recording and reproduction," *Applied sciences*, vol. 7, no. 6, p. 627, 2017.

[66] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, 2015.

[67] J. Kirszenstein, "An image source computer model for room acoustics analysis and electroacoustic simulation," *Applied Acoustics*, vol. 17, no. 4, pp. 275–290, 1984.

[68] A. Krokstad, S. Strom, and S. Sørsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *Journal of Sound and Vibration*, vol. 8, no. 1, pp. 118–125, 1968.

[69] T. Funkhouser, N. Tsingos, I. Carlbom, G. Elko, M. Sondhi, J. E. West, G. Pingali, P. Min, and A. Ngan, "A beam tracing method for interactive architectural acoustics," *The Journal of the acoustical society of America*, vol. 115, no. 2, pp. 739–756, 2004.

[70] M. Hodgson and E.-M. Nosal, "Experimental evaluation of radiosity for room sound-field prediction," *The Journal of the Acoustical Society of America*, vol. 120, no. 2, pp. 808–819, 2006.

[71] S. Siltanen, T. Lokki, and L. Savioja, "Frequency domain acoustic radiance transfer for real-time auralization," *Acta Acustica united with Acustica*, vol. 95, no. 1, pp. 106–117, 2009.

[72] D. Schröder and T. Lentz, "Real-time processing of image sources using binary space partitioning," *Journal of the Audio Engineering Society*, vol. 54, no. 7/8, pp. 604–619, 2006.

[73] J. H. Rindel, "The use of computer modeling in room acoustics," *Journal of vibroengineering*, vol. 3, no. 4, pp. 219–224, 2000.

[74] T. Sakuma, S. Sakamoto, and T. Otsuru, *Computational simulation in architectural and environmental acoustics.* Springer, 2014.

[75] M. Hornikx, "Ten questions concerning computational urban acoustics," *Building and Environment*, vol. 106, pp. 409–421, 2016.

[76] S. Brenner and R. Scott, *The mathematical theory of finite element methods.* Springer Science & Business Media, 2007.

[77] A. Pietrzyk and M. Kleiner, "The application of the finite-element method to the prediction of soundfields of small rooms at low frequencies," in *Audio Engineering Society Convention 102.* Audio Engineering Society, 1997.

[78] Q. Li, J. Xing, R. Tang, and Y. Zhang, "Finite-element method for calculating the sound field in a tank with impedance boundaries," *Mathematical Problems in Engineering*, vol. 2020, 2020.

[79] S. Kirkup, "The boundary element method in acoustics: A survey," *Applied Sciences*, vol. 9, no. 8, p. 1642, 2019.

[80] A. D. Pierce, *Acoustics: an introduction to its physical principles and applications.* Springer, 2019.

[81] J. T. Katsikadelis, *Boundary elements: theory and applications.* Elsevier, 2002.

[82] S. Marburg, "Six boundary elements per wavelength: Is that enough?" *Journal of computational acoustics*, vol. 10, no. 01, pp. 25–51, 2002.

[83] B. Hamilton and S. Bilbao, "Fdtd methods for 3-d room acoustics simulation with high-order accuracy in space and time," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2112–2124, 2017.

[84] M. Hornikx, R. Waxler, and J. Forssén, "The extended fourier pseudospectral time-domain method for atmospheric sound propagation," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 1632–1646, 2010.

[85] D. Murphy, S. Shelley, A. Foteinou, J. Brereton, and H. Daffern, "Acoustic heritage and audio creativity: the creative application of sound in the representation, understanding and experience of past environments," *Internet Archaeology*, 2017.

[86] S. Petrausch and R. Rabenstein, "Highly efficient simulation and visualization of acoustic wave fields with the functional transformation method," *Simulation and Visualization*, pp. 279–290, 2005.

[87] R. Mehra, N. Raghuvanshi, L. Savioja, M. C. Lin, and D. Manocha, "An efficient gpu-based time domain solver for the acoustic wave equation," *Applied Acoustics*, vol. 73, no. 2, pp. 83–94, 2012.

[88] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, 1993.

[89] N. Hahn and S. Spors, "Sound field synthesis of virtual cylindrical waves using circular and spherical loudspeaker arrays," in *Audio Engineering Society Convention 138.* Audio Engineering Society, 2015.

[90] B. Xie, *Head-related transfer function and virtual auditory display.* J. Ross Publishing, 2013.

[91] M. Otani and H. Shigetani, "Reproduction accuracy of higher-order ambisonics with max-re and/or least norm solution in decoding," *Acoustical Science and Technology*, vol. 40, no. 1, pp. 23–28, 2019.

[92] W. G. Gardner and K. D. Martin, "Hrtf measurements of a kemar," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.

[93] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575).* IEEE, 2001, pp. 99–102.

[94] F. Brinkmann, A. Lindau, S. Weinzierl, G. Geissler, and S. van de Par, "A high resolution head-related transfer function database including different orientations of head above the torso," in *Proceedings of AIA-DAGA 2013, Merano*, 2013, pp. 596–599.

[95] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database," *Applied Sciences*, vol. 8, no. 11, p. 2029, 2018.

[96] P. C. Lee, C. W. Senders, B. J. Gantz, and S. R. Otto, "Transient sensorineural hearing loss after overuse of portable headphone cassette radios," *Otolaryngology—Head and Neck Surgery*, vol. 93, no. 5, pp. 622–625, 1985.

[97] Z. Schärer and A. Lindau, "Evaluation of equalization methods for binaural signals," in *Audio Engineering Society Convention 126.* Audio Engineering Society, 2009.

[98] S. Pelzer, L. Aspöck, D. Schröder, and M. Vorländer, "Interactive real-time simulation and auralization for modifiable rooms," *Building Acoustics*, vol. 21, no. 1, pp. 65–73, 2014.

[99] J. Mori, S. Yokoyama, F. Satoh, and H. Tachibana, "Auralization of municipal public address announcements by applying geometrical sound simulation and multi-channel reproduction techniques," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. Acoustical Society of America, 2013, pp. 105–134.

[100] R. Suárez, A. Alonso, and J. J. Sendra, "Archaeoacoustics of intangible cultural heritage: The sound of the maior ecclesia of cluny," *Journal of cultural heritage*, vol. 19, pp. 567–572, 2016.

[101] M. R. Thomas, "Wayverb: A graphical tool for hybrid room acoustics simulation," Ph.D. dissertation, University of Huddersfield, 2017.

[102] H. Bai, "Hybrid models for acoustic reverberation," Ph.D. dissertation, Laboratoire Traitement et Communication de l'Information (LTCI), 2016.

[103] A. Southern, D. T. Murphy, and L. Savioja, "Boundary absorption approximation in the spatial high-frequency extrapolation method for parametric room impulse response synthesis," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2770–2782, 2019.

[104] J. A. Hargreaves, L. R. Rendell, and Y. W. Lam, "A framework for auralization of boundary element method simulations including source and receiver directivity," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2625–2637, 2019.

[105] D. Takeuchi, K. Yatabe, and Y. Oikawa, "Source directivity approximation for finite-difference time-domain simulation by estimating initial value," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2638–2649, 2019.

[106] S. Bilbao, J. Ahrens, and B. Hamilton, "Incorporating source directivity in wave-based virtual acoustics: Time-domain models and fitting to measured data," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. 2692–2703, 2019.

[107] F. Georgiou and M. Hornikx, "Incorporating directivity in the fourier pseudospectral time-domain method using spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 140, no. 2, pp. 855–865, 2016.

[108] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A round robin on room acoustical simulation and auralization," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2746–2760, 2019.

[109] B. N. J. Postma and B. F. G. Katz, “Perceptive and objective evaluation of calibrated room acoustic simulation auralizations,” *The Journal of the Acoustical Society of America*, vol. 140, no. 6, pp. 4326–4337, 2016. [Online]. Available: https://doi.org/10.1121/1.4971422

[110] F. Rietdijk, K. Heutschi, C. Zellmann, and M. June, “Determining an empirical emission model for the auralization of jet aircraft,” in *Proceedings of the 10th European Conference on Noise Control, Maastricht, The Netherlands*, vol. 31, 2015, pp. 781–784.

[111] R. Pieren, A. Zemp, S. Sohr, and K. Heutschi, “Auralisation of railway noise: a concept for the emission synthesis of rolling and impact noise,” in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 253, no. 8. Institute of Noise Control Engineering, 2016, pp. 274–280.

[112] J. Maillard and J. Jagla, “Real time auralization of non-stationary traffic noise-quantitative and perceptual validation in an urban street,” in *Proceedings of the AIA-DAGA Conference on Acoustics, Merano, Italy*, 2013, pp. 18–21.

[113] F. Georgiou, “Modeling for auralization of urban environments,” Ph.D. dissertation, Ph. D. Thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2018.

[114] F. Stevens, D. T. Murphy, and S. L. Smith, “Soundscape auralisation and visualisation: A cross-modal approach to soundscape evaluation,” *DAFx 2018*, 2018.

[115] P. Malecki, J. Piechowicz, and J. Wiciak, “Auralization of selected forests in poland,” in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 255, no. 3. Institute of Noise Control Engineering, 2017, pp. 4152–4156.

[116] C. Dora, M. A. Phillips, and M. Phillips, *Transport, environment and health.* WHO Regional Office Europe, 2000, no. 89.

[117] R. Pieren, T. Bütler, and K. Heutschi, “Auralization of accelerating passenger cars using spectral modeling synthesis,” *Applied Sciences*, vol. 6, no. 1, p. 5, 2015.

[118] A. Hoffmann, *AURALIZATION, PERCEPTION AND DETECTION OF TYRE–ROAD NOISE*, ser. Doktorsavhandlingar vid Chalmers tekniska högskola. Ny serie, no: 4150. Department of Civil and Environmental Engineering, Applied Acoustics, Chalmers University of Technology,, 2016.

[119] H. G. Jonasson, “Acoustical source modelling of road vehicles,” *Acta Acustica united with Acustica*, vol. 93, no. 2, pp. 173–184, 2007.

[120] R. Smyth, H. Rice, P. McDonald, and A. Gerdelan, “Simulation of vehicle noise in the virtual city,” in *INTER-NOISE and NOISE-CON Congress*

*and Conference Proceedings*, vol. 2010, no. 6. Institute of Noise Control Engineering, 2010, pp. 4896–4905.

[121] N. F. Viemeister, "Temporal modulation transfer functions based upon modulation thresholds," *The Journal of the Acoustical Society of America*, vol. 66, no. 5, pp. 1364–1380, 1979.

[122] G. Zachos, J. Forssen, W. Kropp, and L. Estevez-Mauriz, "Background traffic noise synthesis," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, 2016, pp. 3502–3508.

[123] J. Kang, "Numerical modeling of the sound fields in urban squares," *The Journal of the Acoustical Society of America*, vol. 117, no. 6, pp. 3695–3706, 2005.

[124] K. Jian, *Urban sound environment.* CRC Press, 2006.

[125] G. Puglisi, J. Kang, Y. Smyrnova, and A. Astolfi, "Effect of vegetation on sound fields in idealised urban open spaces," *Proc. of AIA-DAGA*, pp. 18–21, 2013.

[126] F. Georgiou, M. Hornikx, and A. Kohlrausch, "Auralization of a car pass-by using impulse responses computed with a wave-based method," *Acta Acustica united with Acustica*, vol. 105, no. 2, pp. 381–391, 2019.

[127] M. Hornikx, D. Botteldooren, T. Van Renterghem, and J. Forssén, "Modelling of scattering of sound from trees by the pstd method," in *Forum Acusticum 2011.* European Accoustics Association (EAA), 2011, pp. 839–844.

[128] F. Stevens, D. T. Murphy, L. Savioja, and V. Välimäki, "Modeling sparsely reflecting outdoor acoustic scenes using the waveguide web," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1566–1578, 2017.

[129] W. H. Organization *et al.*, "Iso 9613-2: Acoustics-attenuation of sound propagation outdoors, part 2: General method of calculation," *Geneva: International Organization for Standardization*, 1996.

[130] H. G. Jonasson and S. Storeheier, "Nord 2000. new nordic prediction method for road traffic noise," 2001.

[131] E. Salomons, D. Van Maercke, J. Defrance, and F. de Roo, "The harmonoise sound propagation model," *Acta acustica united with acustica*, vol. 97, no. 1, pp. 62–74, 2011.

[132] S. Mandal, "Brief introduction of virtual reality & its challenges," *International Journal of Scientific & Engineering Research*, vol. 4, no. 4, pp. 304–309, 2013.

[133] J. Carmigniani and B. Furht, "Augmented reality: an overview," *Handbook of augmented reality*, pp. 3–46, 2011.

[134] M. Speicher, B. D. Hall, and M. Nebeling, "What is mixed reality?" in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–15.

[135] T. Lentz, D. Schröder, M. Vorländer, and I. Assenmacher, "Virtual reality system with integrated sound field simulation and reproduction," *EURASIP journal on advances in signal processing*, vol. 2007, no. 1, p. 070540, 2007.

[136] R. Duraiswaini, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of hrtfs [head related transfer functions]," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. IEEE, 2004, pp. iv–iv.

[137] C. S. Reddy and R. M. Hegde, "Horizontal plane hrtf interpolation using linear phase constraint for rendering spatial audio," in *2016 24th European Signal Processing Conference (EUSIPCO).* IEEE, 2016, pp. 1668–1672.

[138] F. Grijalva, L. C. Martini, D. Florencio, and S. Goldenstein, "Interpolation of head-related transfer functions using manifold learning," *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 221–225, 2017.

[139] S. Viollon, C. Lavandier, and C. Drake, "Influence of visual setting on sound ratings in an urban environment," *Applied acoustics*, vol. 63, no. 5, pp. 493–511, 2002.

[140] L. Maffei, M. Masullo, F. Aletta, and M. Di Gabriele, "The influence of visual characteristics of barriers on railway noise perception," *Science of the Total Environment*, vol. 445, pp. 41–47, 2013.

[141] L. Maffei, T. Iachini, M. Masullo, F. Aletta, F. Sorrentino, V. P. Senese, and F. Ruotolo, "The effects of vision-related aspects on noise perception of wind turbines in quiet areas," *International Journal of Environmental Research and Public Health*, vol. 10, no. 5, pp. 1681–1697, 2013. [Online]. Available: https://www.mdpi.com/1660-4601/10/5/1681

[142] T. Cassidy, *Environmental psychology: Behaviour and experience in context.* Psychology Press, 2013.

[143] T. Hempel and N. Chouard, "Evaluation of interior car sound with a new specific semantic differential design," *The Journal of the Acoustical Society of America*, vol. 105, no. 2, pp. 1280–1280, 1999.

[144] M. Schütte, U. Müller, S. Sandrock, B. Griefahn, C. Lavandier, and B. Barbot, "Perceived quality features of aircraft sounds: An analysis of the measurement characteristics of a newly created semantic differential," *Applied Acoustics*, vol. 70, no. 7, pp. 903–914, 2009.

[145] Merriam-Webster Online, "Merriam-Webster Online Dictionary," 2009. [Online]. Available: http://www.merriam-webster.com

[146] A. Farnell, “An introduction to procedural audio and its application in computer games,” in *Audio mostly conference*, 2007, pp. 1–31.

[147] R. A. Moog, “Midi: Musical instrument digital interface,” *Journal of the Audio Engineering Society*, vol. 34, no. 5, pp. 394–404, 1986.

[148] V. Lazzarini, S. Yi, J. Heintz, Ø. Brandtsegg, I. McCurdy *et al.*, *Csound: a sound and music computing system.* Springer, 2016.

[149] N. Bøttcher, “Can interactive procedural audio affect the motorical behaviour of players in computer games with motion controllers?” in *Audio Engineering Society Conference: 49th International Conference: Audio for Games.* Audio Engineering Society, 2013.

[150] H. Hacıhabiboğlu, “Procedural synthesis of gunshot sounds based on physically motivated models,” in *Game Dynamics.* Springer, 2017, pp. 47–69.

[151] J. H. McDermott and E. P. Simoncelli, “Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis,” *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.

[152] R. Guski, “Psychological methods for evaluating sound quality and assessing acoustic information,” *Acta Acustica united with Acustica*, vol. 83, no. 5, pp. 765–774, 1997.

[153] M. Russ, *Sound synthesis and sampling.* Taylor & Francis, 2004.

[154] R. Mignot and V. Välimäki, “Extended subtractive synthesis of harmonic musical tones,” in *Audio Engineering Society Convention 136.* Audio Engineering Society, 2014.

[155] X. Serra and J. Smith, “Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition,” *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[156] A. De Cheveigné and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[157] K. Karplus and A. Strong, “Digital synthesis of plucked-string and drum timbres,” *Computer Music Journal*, vol. 7, no. 2, pp. 43–55, 1983.

[158] D. A. Jaffe and J. O. Smith, “Extensions of the karplus-strong plucked-string algorithm,” *Computer Music Journal*, vol. 7, no. 2, pp. 56–69, 1983.

[159] L. Turchet, S. Serafin, S. Dimitrov, and R. Nordahl, “Physically based sound synthesis and control of footsteps sounds,” in *Proceedings of digital audio effects conference*, vol. 11, 2010.

[160] K. v. d. Doel, “Physically based models for liquid sounds,” *ACM Transactions on Applied Perception (TAP)*, vol. 2, no. 4, pp. 534–546, 2005.

[161] A. Cipriani and M. Giri, *Electronic music and sound design.* Contemponet, 2010, vol. 1.

[162] A. J. Farnell and O. Uk, "Marching onwards: procedural synthetic footsteps for video games and animation," in *Proceedings of the Pure Data Convention.* Citeseer, 2007.

[163] J. Newman, "Driving the sid chip: Assembly language, composition, and sound design for the c64," *G— A— M— E Games as Art, Media, Entertainment*, vol. 1, no. 6, 2017.

[164] K. Collins *et al.*, *Game sound: an introduction to the history, theory, and practice of video game music and sound design.* Mit Press, 2008.

[165] K. Brandenburg, "Mp3 and aac explained," in *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding.* Audio Engineering Society, 1999.

[166] N. Bottcher, H. P. Martinez, and S. Serafin, "Procedural audio in computer games using motion controllers: an evaluation on the effect and perception," *International Journal of Computer Games Technology*, vol. 2013, p. 6, 2013.

[167] K. Squire, *Open-ended video games: A model for developing learning for the interactive age.* MacArthur Foundation Digital Media and Learning Initiative, 2007.

[168] R. Selfridge, D. Moffat, and J. D. Reiss, "Sound synthesis of objects swinging through air using physical models," *Applied Sciences*, vol. 7, no. 11, p. 1177, 2017.

[169] Y. Dobashi, T. Yamamoto, and T. Nishita, "Real-time rendering of aerodynamic sound using sound textures based on computational fluid dynamics," in *ACM SIGGRAPH 2003 Papers*, 2003, pp. 732–740.

[170] P. Bahadoran, A. Benito, T. Vassallo, and J. D. Reiss, "Fxive: A web platform for procedural sound synthesis," in *Audio Engineering Society Convention 144.* Audio Engineering Society, 2018.

[171] C. Verron and G. Drettakis, "Procedural audio modeling for particle-based environmental effects," in *Audio Engineering Society Convention 133.* Audio Engineering Society, 2012.

[172] D. H. M. Kemper and D. Hug, "From foley to function: A pedagogical approach to sound design for novel interactions," *Journal of Sonic Studies*, vol. 6, no. 1, pp. 1–23, 2014.

[173] B. Shen, W. Tan, J. Guo, H. Cai, B. Wang, and S. Zhuo, "A study on design requirement development and satisfaction for future virtual world systems," *Future Internet*, vol. 12, no. 7, p. 112, 2020.

[174] R. M. Schafer, *The soundscape: Our sonic environment and the tuning of the world.* Simon and Schuster, 1993.

[175] 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, 2014.

[176] D. Birchfield, N. Mattar, and H. Sundaram, "Design of a generative model for soundscape creation," in *in Proceedings of the International Computer Music Conference. International Computer Music Association.* Citeseer, 2005.

[177] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).* IEEE, 2017, pp. 344–348.

[178] N. Finney and J. Janer, "Soundscape generation for virtual environments using community-provided audio databases," in *W3C Workshop: Augmented Reality on the Web.* Barcelona, 2010.

[179] V. Akkermans, F. Font Corbera, J. Funollet, B. De Jong, G. Roma Trepat, S. Togias, and X. Serra, "Freesound 2: An improved platform for sharing audio clips," in *Klapuri A, Leider C, editors. ISMIR 2011: Proceedings of the 12th International Society for Music Information Retrieval Conference; 2011 October 24–28; Miami, Florida (USA).* International Society for Music Information Retrieval (ISMIR), 2011.

[180] G. Keizer, *The unwanted sound of everything we want: A book about noise.* PublicAffairs, 2010.

[181] D. Hendy, *Noise: A human history of sound and listening.* Profile books, 2013.

[182] J. J. Alvarsson, S. Wiens, and M. E. Nilsson, "Stress recovery during exposure to nature sound and environmental noise," *International journal of environmental research and public health*, vol. 7, no. 3, pp. 1036–1046, 2010.

[183] K. I. Hume, M. Brink, M. Basner *et al.*, "Effects of environmental noise on sleep," *Noise and health*, vol. 14, no. 61, p. 297, 2012.

[184] T. Münzel, F. P. Schmidt, S. Steven, J. Herzog, A. Daiber, and M. Sørensen, "Environmental noise and the cardiovascular system," *Journal of the American College of Cardiology*, vol. 71, no. 6, pp. 688–697, 2018.

[185] B. Gygi, G. R. Kidd, and C. S. Watson, "Similarity and categorization of environmental sounds," *Perception & psychophysics*, vol. 69, no. 6, pp. 839–855, 2007.

[186] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2017, pp. 776–780.

[187] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[188] 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, 2018.

[189] P. Maijala, Z. Shuyang, T. Heittola, and T. Virtanen, "Environmental noise monitoring using source classification in sensors," *Applied Acoustics*, vol. 129, pp. 258–267, 2018.

[190] M. Śliwińska-Kowalska and K. Zaborowski, "Who environmental noise guidelines for the european region: a systematic review on environmental noise and permanent hearing loss and tinnitus," *International journal of environmental research and public health*, vol. 14, no. 10, p. 1139, 2017.

[191] A. Quaranta, P. Portalatini, and D. Henderson, "Temporary and permanent threshold shift: an overview." *Scandinavian audiology. Supplementum*, vol. 48, pp. 75–86, 1998.

[192] I. ISO, "Acoustics—estimation of noise-induced hearing loss," *International Organization for Standardization, Geneva, Switzerland*, 2013.

[193] B. Berglund, T. Lindvall *et al.*, *Community noise.* Center for Sensory Research, Stockholm University and Karolinska Institute . . . , 1995.

[194] W. D. Ward, A. Glorig, and D. L. Sklar, "Temporary threshold shift from octave-band noise: applications to damage-risk criteria," *The Journal of the Acoustical Society of America*, vol. 31, no. 4, pp. 522–528, 1959.

[195] D. Baguley, D. McFerran, and D. Hall, "Tinnitus," *The Lancet*, vol. 382, no. 9904, pp. 1600–1607, 2013.

[196] J. A. Henry, M. A. Schechter, S. M. Nagler, and S. A. Fausti, "Comparison of tinnitus masking and tinnitus retraining therapy," *Journal of the American Academy of Audiology*, vol. 13, no. 10, pp. 559–581, 2002.

[197] J. A. Henry, M. Schechter, T. Zaugg, S. Griest, P. Jastreboff, J. Vernon, C. Kaelin, M. Meikle, K. Lyons, and B. Stewart, "Clinical trial to compare tinnitus masking and tinnitus retraining therapy," *Acta Oto-Laryngologica*, vol. 126, no. sup556, pp. 64–69, 2006.

[198] J. Selander, M. E. Nilsson, G. Bluhm, M. Rosenlund, M. Lindqvist, G. Nise, and G. Pershagen, "Long-term exposure to road traffic noise and myocardial infarction," *Epidemiology*, pp. 272–279, 2009.

[199] M. Sørensen, Z. J. Andersen, R. B. Nordsborg, T. Becker, A. Tjønneland, K. Overvad, and O. Raaschou-Nielsen, "Long-term exposure to road traffic noise and incident diabetes: a cohort study," *Environmental health perspectives*, vol. 121, no. 2, pp. 217–222, 2013.

[200] M. Sørensen, M. Hvidberg, Z. J. Andersen, R. B. Nordsborg, K. G. Lillelund, J. Jakobsen, A. Tjønneland, K. Overvad, and O. Raaschou-Nielsen, "Road traffic noise and stroke: a prospective cohort study," *European heart journal*, vol. 32, no. 6, pp. 737–744, 2011.

[201] S. A. Stansfeld and M. P. Matheson, "Noise pollution: non-auditory effects on health," *British medical bulletin*, vol. 68, no. 1, pp. 243–257, 2003.

[202] Y. Xiang, W. Jianghua, L. Hui, and C. Yuxiao, "Experimental study of the influence of a silent environment on human sleep quality," *Journal of Tsinghua University (Science and Technology)*, vol. 58, no. 12, pp. 1115–1120, 2018.

[203] P. Knipschild, "Medical effects of aircraft noise: community cardiovascular survey," *International Archives of Occupational and Environmental Health*, vol. 40, no. 3, pp. 185–190, 1977.

[204] G. L. Bluhm, N. Berglind, E. Nordling, and M. Rosenlund, "Road traffic noise and hypertension," *Occupational and environmental medicine*, vol. 64, no. 2, pp. 122–126, 2007.

[205] A. von Eiff, G. Friedrich, W. Langewitz, H. Neus, H. Rüddel, G. Schirmer, and W. Schulte, "Traffic noise and hypertension risk. hypothalamus theory of essential hypertension. second communication," *MMW: Munchener Medizinische Wochenschrift*, vol. 123, no. 11, pp. 420–424, 1981.

[206] E. E. Ryherd, K. P. Waye, and L. Ljungkvist, "Characterizing noise and perceived work environment in a neurological intensive care unit," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 747–756, 2008.

[207] I. 1996, "Acoustics–description, measurement and assessment of environmental noise," 2017.

[208] I. E. Commission, "Iec 61260," *Electroacoustics–Octave-Band and Fractional-Octave-Band Filters*, 1995.

[209] I. E. Commission *et al.*, "Iec 60942: 2003," *Electroacoustics—Sound Calibrators*, 2003.

[210] H. G. Leventhall *et al.*, "Low frequency noise and annoyance," *Noise and Health*, vol. 6, no. 23, p. 59, 2004.

[211] B. British Standard, "4142: 1997:"method for rating industrial noise affecting mixed residential and industrial areas"," *Standards Board*, 1997.

[212] A. Steele, "An environmental impact assessment of the proposal to build a wind farm at langdon common in the north pennines, uk," 1991.

[213] M. MCIEH and J. Hetherington, "A practical evaluation of objective noise criteria used for the assessment of disturbance due to entertainment music," *Journal of Environmental Health Research*, vol. 4, no. 2, 2005.

[214] S. Kephalopoulos, M. Paviotti, and F. Anfosso-Lédée, "Common noise assessment methods in europe (cnossos-eu)," 2012.

[215] S. SAKAMOTO, T. MATSUMOTO, T. TAJIKA, and A. FUKUSHIMA, "Road traffic noise prediction model "asj rtn-model 2013" proposed by the acoustical society of japan–part 1: Outline of the calculation model," in *Proc. ICA*, 2019, pp. 9–13.

[216] G. B. Jónsson and F. Jacobsen, "A comparison of two engineering models for outdoor sound propagation: Harmonoise and nord2000," *Acta Acustica united with Acustica*, vol. 94, no. 2, pp. 282–289, 2008.

[217] E.-l. Europa, "Directive 2002/49/ec of the european parliament and of the council of 25 june 2002 relating to the assessment and management of environmental noise - declaration by the commission in the conciliation committee on the directive relating to the assessment and management of environmental noise," *URL: http://data.europa.eu/eli/dir/2002/49/oj*, 2002.

[218] P. Finne, "Road noise auralisation for planning new roads," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 253, no. 5. Institute of Noise Control Engineering, 2016, pp. 3222–3227.

[219] B. Truax, *Acoustic communication.* Greenwood Publishing Group, 2001.

[220] I. ISO, "Acoustics — soundscape — part 1: Definition and conceptual framework," *International Organization for Standardization, Geneva, Switzerland*, 2014.

[221] M. Niessen, C. Cance, and D. Dubois, "Categories for soundscape: toward a hybrid classification," in *inter-noise and noise-con congress and conference proceedings*, vol. 2010, no. 5. Institute of Noise Control Engineering, 2010, pp. 5816–5829.

[222] S. Payne, W. Davies, and M. Adams, *Research into the practical and policy applications of soundscape concepts and techniques in urban areas (NANR 200).* HMSO, Oct. 2009.

[223] D. A. Hall, A. Irwin, M. Edmondson-Jones, S. Phillips, and J. E. Poxon, "An exploratory evaluation of perceptual, psychoacoustic and acoustical properties of urban soundscapes," *Applied Acoustics*, vol. 74, no. 2, pp. 248–254, 2013.

[224] D. Dubois, C. Guastavino, and M. Raimbault, "A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories," *Acta acustica united with acustica*, vol. 92, no. 6, pp. 865–874, 2006.

[225] M. C. Green and D. Murphy, "Eigenscape: A database of spatial acoustic scene recordings," *Applied Sciences*, vol. 7, no. 11, p. 1204, 2017.

[226] M. Lionello, F. Aletta, and J. Kang, "On the dimension and scaling analysis of soundscape assessment tools: a case study about the" method a" of iso/ts 12913-2: 2018," in *Proceedings of the International Conference on Acoustics ICA*, vol. 2019, 2019.

[227] J. Yong Jeon, J. Young Hong, and P. Jik Lee, "Soundwalk approach to identify urban soundscapes individually," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 803–812, 2013.

[228] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A spatial audio quality inventory (saqi)," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 984–994, 2014.

[229] F. Stevens, D. Murphy, and S. Smith, "Soundscape preference rating using semantic differential pairs and the self-assessment manikin," in *Sound and Music Computing 2016*, Aug. 2016, pp. 455–462, sound and Music Computing 2016 (SMC2016) ; Conference date: 31-08-2016 Through 03-09-2016. [Online]. Available: http://quintetnet.hfmt-hamburg.de/SMC2016/

[230] K. Jian *et al.*, *COST TUD Action TD-0804 Soundscapes of European Cities and Landscapes.* Soundscape-COST, OXFORD, 2013.

[231] A. Fiebig, S. Guidati, and A. Goehrke, "Psychoacoustic evaluation of traffic noise," *NAG, DAGA*, 2009.

[232] C. Lavandier and B. Defréville, "The contribution of sound source characteristics in the assessment of urban soundscapes," *Acta acustica united with Acustica*, vol. 92, no. 6, pp. 912–921, 2006.

[233] P. Ricciardi, P. Delaitre, C. Lavandier, F. Torchia, and P. Aumond, "Sound quality indicators for urban places in paris cross-validated by milan data," *The Journal of the Acoustical Society of America*, vol. 138, no. 4, pp. 2337–2348, 2015.

[234] J. Kang, F. Aletta, T. Oberman, M. Erfanian, M. Kachlicka, M. Lionello, and A. Mitchell, "Towards soundscape indices," in *Proceedings of the 23rd International Congress on Acoustics, Aachen, Germany*, 2019, pp. 9–13.

[235] M. Erfanian, A. J. Mitchell, J. Kang, and F. Aletta, "The psychophysiological implications of soundscape: A systematic review of empirical literature and a research agenda," *International journal of environmental research and public health*, vol. 16, no. 19, p. 3533, 2019.

[236] S. Torresin, R. Albatici, F. Aletta, F. Babich, and J. Kang, "Assessment methods and factors determining positive indoor soundscapes in residential buildings: A systematic review," *Sustainability*, vol. 11, no. 19, p. 5290, 2019.

[237] M. Lionello, F. Aletta, and J. Kang, "A systematic review of prediction models for the experience of urban soundscapes," *Applied Acoustics*, vol. 170, p. 107479, 2020.

[238] F. Aletta, T. Oberman, and J. Kang, "Associations between positive health-related effects and soundscapes perceptual constructs: A systematic review," *International journal of environmental research and public health*, vol. 15, no. 11, p. 2392, 2018.

[239] A. Mitchell, T. Oberman, F. Aletta, M. Erfanian, M. Kachlicka, M. Lionello, and J. Kang, "The soundscape indices (ssid) protocol: A method for urban soundscape surveys—questionnaires with acoustical and contextual information," *Applied Sciences*, vol. 10, no. 7, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/7/2397

[240] C. Xu and J. Kang, "Soundscape evaluation: Binaural or monaural?" *The Journal of the Acoustical Society of America*, vol. 145, no. 5, pp. 3208–3217, 2019. [Online]. Available: https://doi.org/10.1121/1.5102164

[241] S. E. Tan, "Megaphones hiding in trees: civic instruction via mediated soundscapes in places of natural beauty in china," *International Communication of Chinese Culture*, vol. 7, no. 2, pp. 189–214, 2020.

[242] W. Yang, "An aesthetic approach to the soundscape of urban public open spaces." Ph.D. dissertation, University of Sheffield, 2005.

[243] J. E. G. Fernando, G. M. Adrian, G. T. Cesar, and F. L.-R. Diego, "Fountains as sound elements in the design of urban public walks soundscapes," in *PROCEEDINGS of the 22nd International Congress on Acoustics, ICA2016*, vol. 19, no. 1. International Congress on Acoustics, 2016, pp. 1–10.

[244] B. Fontana, "The relocation of ambient sound: urban sound sculpture," *Leonardo*, vol. 41, no. 2, pp. 154–158, 2008.

[245] V. Keylin, "Corporeality of music and sound sculpture," *Organised Sound*, vol. 20, no. 2, pp. 182–190, 2015.

[246] D. Misawa, "Transparent sculpture: An embodied auditory interface for sound sculpture," in *Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction*, 2013, pp. 389–390.

[247] M. R. Iturbide, "The expansion of sound sculpture and sound intallation in art," 2014.

[248] K. Burgemeister and C. Hough, "Auralisation for airport noise impact assessments: Measurements and applications," in *Proceedings of ACOUSTICS 2016, November 2016, Brisbane, Australia*. Acoustical Society of Australia, 2016.

[249] A. consultancy, "Hs2 soundlab demonstrations, simulating the sound of trains along the proposed hs2 route," 2018.

[250] C. John and M. Damian, "Listening to the commons," 2017.

[251] Voice and vote, "Women's place in parliament exhibition," 2018.

[252] J. Forssén, T. Kaczmarek, P. Lundén, M. E. Nilsson, and J. Alvarsson, "Auralization of traffic noise within the listen project: Preliminary results for passenger car pass-by," in *Euronoise 2009*. Institute of Acoustics, 2009.

[253] F. Rietdijk, *Auralisation of airplanes considering sound propagation in a turbulent atmosphere*. Chalmers University of Technology, 2017.

[254] A. Sahai, F. Wefers, S. Pick, E. Stumpf, M. Vorländer, and T. Kuhlen, "Interactive simulation of aircraft noise in aural and visual virtual environments," *Applied acoustics*, vol. 101, pp. 24–38, 2016.

[255] A. Sahai, E. Anton, E. Stumpf, F. Wefers, and M. Vorlaender, "Interdisciplinary auralization of take-off and landing procedures for subjective assessment in virtual reality environments," in *Aiaa/ceas Aeroacoustics Conference*, 2013.

[256] R. Pieren, K. Heutschi, J. M. Wunderli, M. Snellen, and D. G. Simons, "Auralization of railway noise: Emission synthesis of rolling and impact noise," *Applied Acoustics*, vol. 127, pp. 34–45, 2017.

[257] K. Heutschi, "Sonroad: New swiss road traffic noise model," *Acta Acustica united with Acustica*, vol. 90, no. 3, pp. 548–554, 2004.

[258] U. Sandberg and J. Ejsmont, *Tyre/road Noise Reference Book*. INFORMEX, 2002. [Online]. Available: https://books.google.co.uk/books?id=yQW9AAAACAAJ

[259] M. E. Braun, S. J. Walsh, and J. L. Horner, "Sound source contributions for the prediction of vehicle pass-by noise," Ph.D. dissertation, Loughborough University, 2014.

[260] C. Pendharkar, "Auralization of road vehicles using spectral modeling synthesis," Ph.D. dissertation, Chalmers University of Technology, 2012.

[261] A. Acri, E. Nijman, M. Klanner, G. Offner, and R. Corradi, "On the influence of cyclic variability on surface noise contribution analysis of internal combustion engines," *Applied Acoustics*, vol. 132, pp. 97–108, 2018.

[262] J. Jagla, J. Maillard, and N. Martin, "Sample-based engine noise synthesis using an enhanced pitch-synchronous overlap-and-add method," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3098–3108, 2012.

[263] D. A. Heitbrink and S. Cable, "Design of a driving simulation sound engine," in *Driving Simulation Conference, North America 2007 (DSC-NA 2007) Ford Motor CompanyNational Highway Traffic Safety AdministrationUniversity of Iowa, Iowa CityTransportation Research Board*, 2007.

[264] S. Baldan, H. Lachambre, S. Delle Monache, and P. Boussard, "Physically informed car engine sound synthesis for virtual and augmented environments," in *2015 IEEE 2nd VR Workshop on Sonic Interactions for Virtual Environments (SIVE)*. IEEE, 2015, pp. 1–6.

[265] J. A. Yamin and M. H. Dado, “Performance simulation of a four-stroke engine with variable stroke-length and compression ratio,” *Applied energy*, vol. 77, no. 4, pp. 447–463, 2004.

[266] D. Min, B. Park, and J. Park, “Artificial engine sound synthesis method for modification of the acoustic characteristics of electric vehicles,” *Shock and Vibration*, vol. 2018, 2018.

[267] S. Wu, “Engine sound simulation and generation in driving simulator,” Ph.D. dissertation, Masters Thesis, Missouri University of Science and Technology, 2018.

[268] R. Bosch and P. T. Girling, *Automotive handbook*. Society of Automotive Engineers, US, 1996.

[269] G. Fontana and E. Galloni, “Variable valve timing for fuel economy improvement in a small spark-ignition engine,” *Applied Energy*, vol. 86, no. 1, pp. 96–105, 2009.

[270] Z. Lou and G. Zhu, “Review of advancement in variable valve actuation of internal combustion engines,” *Applied Sciences*, vol. 10, no. 4, p. 1216, 2020.

[271] S. Yousufuddin and M. Masood, “Effect of ignition timing and compression ratio on the performance of a hydrogen–ethanol fuelled engine,” *International Journal of Hydrogen Energy*, vol. 34, no. 16, pp. 6945 – 6950, 2009, 4th Dubrovnik Conference. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360319909008362

[272] C. Gong, Z. Li, Y. Chen, J. Liu, F. Liu, and Y. Han, “Influence of ignition timing on combustion and emissions of a spark-ignition methanol engine with added hydrogen under lean-burn conditions,” *Fuel*, vol. 235, pp. 227–238, 2019.

[273] J. O. Smith, “Physical modeling using digital waveguides,” *Computer music journal*, vol. 16, no. 4, pp. 74–91, 1992.

[274] S. Baldan, S. Delle Monache, and D. Rocchesso, “The sound design toolkit,” *SoftwareX*, vol. 6, pp. 255–260, 2017.

[275] D. Potente, “General design principles for an automotive muffler,” in *Proceedings of ACOUSTICS*, 2005, pp. 153–158.

[276] T. Yasuda, C. Wu, N. Nakagawa, and K. Nagamura, “Studies on an automobile muffler with the acoustic characteristic of low-pass filter and helmholtz resonator,” *Applied Acoustics*, vol. 74, no. 1, pp. 49–57, 2013.

[277] G. Sheng, *Vehicle noise, vibration, and sound quality*. SAE, 2012.

[278] J. Forssén, A. Hoffmann, and W. Kropp, “Auralization model for the perceptual evaluation of tyre–road noise,” *Applied Acoustics*, vol. 132, pp. 232–240, 2018.

[279] A. Southern and D. Murphy, "A method for plausible road tyre noise auralization," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3881–3881, 2017.

[280] A. Southern and D. Murphy, "A framework for road traffic noise auralisation," in *Euronoise 2015*. Institute of Acoustics, 2009.

[281] E. N. Directive, "Commission directive (eu) 2015/996 of 19 may 2015 establishing common noise assessment methods according to directive 2002/49/ec of the european parliament and of the council," *Off. J. Eur. Union L*, vol. 168, p. 58, 2015.

[282] C. A. Shaffer, *A practical introduction to data structures and algorithm analysis*. Prentice Hall Upper Saddle River, NJ, 1997.

[283] R. Graf, C.-Y. Kuo, A. Dowling, and W. Graham, "On the horn effect of a tyre/road interface, part i: Experiment and computation," *Journal of Sound and Vibration*, vol. 256, no. 3, pp. 417–431, 2002.

[284] J. A. Ballesteros, E. Sarradj, M. D. Fernández, T. Geyer, and M. J. Ballesteros, "Noise source identification with beamforming in the pass-by of a car," *Applied Acoustics*, vol. 93, pp. 106–119, 2015.

[285] P. Chiariotti, M. Martarelli, and P. Castellini, "Acoustic beamforming for noise source localization–reviews, methodology and applications," *Mechanical Systems and Signal Processing*, vol. 120, pp. 422–448, 2019.

[286] D. Yang, Z. Wang, B. Li, Y. Luo, and X. Lian, "Quantitative measurement of pass-by noise radiated by vehicles running at high speeds," *Journal of Sound and Vibration*, vol. 330, no. 7, pp. 1352–1364, 2011.

[287] K. Collins and R. Dockwray, *The Routledge Companion to Screen Music and Sound*. Routledge, 2017, drive, Speed and Narrative in the Soundscapes of Racing Games. [Online]. Available: http://hdl.handle.net/10034/605882

[288] F. Meng, G. Behler, and M. Vorländer, "A synthesis model for a moving sound source based on beamforming," *Acta acustica united with acustica*, vol. 104, no. 2, pp. 351–362, 2018.

[289] K. Turkowski, "Filters for common resampling tasks," in *Graphics gems*. Academic Press Professional, Inc., 1990, pp. 147–165.

[290] M. Z. Hussain, M. Irshad, M. Sarfraz, and N. Zafar, "Interpolation of discrete time signals using cubic spline function," in *2015 19th International Conference on Information Visualisation*. IEEE, 2015, pp. 454–459.

[291] K. Heutschi, R. Pieren, M. Müller, M. Manyoky, U. W. Hayek, and K. Eggenschwiler, "Auralization of wind turbine noise: Propagation filtering and vegetation noise synthesis," *ACTA Acustica united with Acustica*, vol. 100, no. 1, pp. 13–24, 2014.

[292] T. Van Renterghem, E. Salomons, and D. Botteldooren, "Parameter study of sound propagation between city canyons with a coupled fdtd-pe model," *Applied Acoustics*, vol. 67, no. 6, pp. 487–510, 2006.

[293] J. Blauert, H. Lehnert, J. Sahrhage, and H. Strauss, "An interactive virtual-environment generator for psychoacoustic research. i: Architecture and implementation," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 94–102, 2000.

[294] K. Iu and K. Li, "The propagation of sound in narrow street canyons," *The Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 537–550, 2002.

[295] S. S. Soares, "Hybrid acoustic model for sound propagation in a street canyon," Ph.D. dissertation, Masters Thesis, Eindhoven University of Technology, 2019.

[296] J. Huopaniemi, L. Savioja, and M. Karjalainen, "Modeling of reflections and air absorption in acoustical spaces a digital filter design approach," in *Applications of Signal Processing to Audio and Acoustics, 1997. 1997 IEEE ASSP Workshop on*, 1997.

[297] B. Xie, "On the low frequency characteristics of head-related transfer function," *Chinese Journal of Acoustics*, vol. 28, no. 2, pp. 116–128, 2009.

[298] M. E. Braun, S. J. Walsh, J. L. Horner, and R. Chuter, "Noise source characteristics in the iso 362 vehicle pass-by noise test: Literature review," *Applied Acoustics*, vol. 74, no. 11, pp. 1241–1265, 2013.

[299] K. Vansant, H. Bériot, C. Bertolini, and G. Miccoli, "An update and comparative study of acoustic modeling and solver technologies in view of pass-by noise simulation," *SAE International Journal of Engines*, vol. 7, no. 3, pp. 1593–1609, 2014.

[300] J. Blauert, *The technology of binaural listening.* Springer, 2013.

[301] A. D. May, *Traffic flow fundamentals.* Pearson, First Edition, 1990.

[302] "Unity User Manual," https://docs.unity3d.com/Manual/index.html, accessed: 2020-05-01.

[303] C.-H. Lin and P.-H. Hsu, "Integrating procedural modelling process and immersive vr environment for architectural design education," in *MATEC Web of Conferences*, vol. 104. EDP Sciences, 2017, p. 03007.

[304] T. Alatalo, M. Pouke, T. Koskela, T. Hurskainen, C. Florea, and T. Ojala, "Two real-world case studies on 3d web applications for participatory urban planning," in *Proceedings of the 22nd International Conference on 3D Web Technology*, 2017, pp. 1–9.

[305] D. Mourtzis, V. Zogopoulos, and E. Vlachou, "Augmented reality supported product design towards industry 4.0: a teaching factory paradigm," *Procedia manufacturing*, vol. 23, pp. 207–212, 2018.

[306] M. Puckette *et al.*, "Pure data: another integrated computer music environment," *Proceedings of the second intercollege computer music concerts*, pp. 37–41, 1996.

[307] N. Moody, "Libpd unity integration: a libpd wrapper for unity," Jul. 2018.

[308] "Wwise," https://www.audiokinetic.com/products/wwise/, accessed: 2020-02-12.

[309] "Audio Spatializer SDK," https://docs.unity3d.com/Manual/AudioSpatializerSDK.html, accessed: 2020-01-29.

[310] H. Kim, L. Hernaggi, P. J. Jackson, and A. Hilton, "Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019, pp. 120–126.

[311] M. Julien and J. Jan, "Auralization of urban traffic noise - quantitative and perceptual validation," in *Conference: Congres Francais d'Acoustique, At: Poitiers, France*, vol. 19, no. 1. Congres Francais d'Acoustique, 2014, pp. 1361–1366.

[312] A. Hoffmann and W. Kropp, "Auralization of simulated tyre noise: Psychoacoustic validation of a combined model," *Applied Acoustics*, vol. 145, pp. 220–227, 2019.

[313] A. Southern and D. Murphy, "Comparison of road tyre noise auralisation methods," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 253, no. 7. Institute of Noise Control Engineering, 2016, pp. 1056–1062.

[314] M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman, "Fast and easy crowdsourced perceptual audio evaluation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 619–623.

[315] M. Schoeffler, F.-R. Stöter, H. Bayerlein, B. Edler, and J. Herre, "An experiment about estimating the number of instruments in polyphonic music: A comparison between internet and laboratory results." in *ISMIR*, 2013, pp. 389–394.

[316] T. H. Pedersen, S. Antunes, and B. Rasmussen, "Online listening tests on sound insulation of walls: A feasibility study," in *Proceedings of EURONOISE 2012*. European Acoustics Association-EAA, 2012, pp. 1219–1224.

[317] N. Welch and J. H. Krantz, "The world-wide web as a medium for psychoacoustical demonstrations and experiments: Experience and results," *Behavior Research Methods, Instruments, & Computers*, vol. 28, no. 2, pp. 192–196, 1996.

[318] M. Schoeffler, F.-R. Stöter, B. Edler, and J. Herre, "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the itu-r recommendation bs. 1534 (mushra)," in *1st Web Audio Conference*, 2015, pp. 1–6.

[319] O. Björklund *et al.*, "The design principles of an online listening test for investigating the perception of heavy metal harmony," Ph.D. dissertation, Helsingfors universitet, 2005.

[320] L. Blin, O. Boeffard, and V. Barreaud, "Web-based listening test system for speech synthesis and speech conversion evaluation," Ph.D. dissertation, University of Helsinki, 2008.

[321] M. Schoeffler, J. L. Gernert, M. Neumayer, S. Westphal, and J. Herre, "On the validity of virtual reality-based auditory experiments: a case study about ratings of the overall listening experience," *Virtual Reality*, vol. 19, no. 3, pp. 181–200, 2015.

[322] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen, "The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks." *Psychological review*, vol. 113, no. 4, p. 700, 2006.

[323] K. Vancleef, J. C. Read, W. Herbert, N. Goodship, M. Woodhouse, and I. Serrano-Pedraza, "Two choices good, four choices better: For measuring stereoacuity in children, a four-alternative forced-choice paradigm is more efficient than two," *PLoS One*, vol. 13, no. 7, p. e0201366, 2018.

[324] G. Belojevic, B. Jakovljevic *et al.*, "Factors influencing subjective noise sensitivity in an urban population," *Noise and Health*, vol. 4, no. 13, p. 17, 2001.

[325] P. Lundquist, K. Holmberg, U. Landstrom *et al.*, "Annoyance and effects on work from environmental noise at school," *Noise and Health*, vol. 2, no. 8, p. 39, 2000.

[326] F. Minichilli, F. Gorini, E. Ascari, F. Bianchi, A. Coi, L. Fredianelli, G. Licitra, F. Manzoli, L. Mezzasalma, and L. Cori, "Annoyance judgment and measurements of environmental noise: A focus on italian secondary schools," *International journal of environmental research and public health*, vol. 15, no. 2, p. 208, 2018.

[327] S. Bech, "Selection and training of subjects for listening tests on sound-reproducing equipment," *Journal of the Audio Engineering Society*, vol. 40, no. 7/8, pp. 590–610, 1992.

[328] S. E. Olive, "Differences in performance and preference of trained versus untrained listeners in loudspeaker tests: A case study," *Journal of the Audio Engineering Society*, vol. 51, no. 9, pp. 806–825, 2003.

[329] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[330] C. Mendonça and S. Delikaris-Manias, "Statistical tests with mushra data," in *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[331] T. Sporer, J. Liebetrau, and S. Schneider, "Statistics of mushra revisited," in *Audio Engineering Society Convention 127*. Audio Engineering Society, 2009.

[332] Y. Chan, "Biostatistics 102: quantitative data–parametric & non-parametric tests," *blood Press*, vol. 140, no. 24.08, p. 79, 2003.

[333] A. Field, *Discovering Statistics using IBM SPSS Statistics*. London: SAGE Publications, 2013.

[334] W. A. Teder-Sälejärvi, S. A. Hillyard, B. Röder, and H. J. Neville, "Spatial attention to central and peripheral auditory stimuli as indexed by event-related potentials," *Cognitive Brain Research*, vol. 8, no. 3, pp. 213–227, 1999.

[335] S. A. Amman and M. Das, "An efficient technique for modeling and synthesis of automotive engine sounds," *IEEE Transactions on Industrial Electronics*, vol. 48, no. 1, pp. 225–234, 2001.

[336] L. Maffei, M. Masullo, A. Pascale, G. Ruggiero, and V. P. Romero, "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.

[337] J. Y. Hong and J. Y. Jeon, "The effects of audio–visual factors on perceptions of environmental noise barrier performance," *Landscape and Urban Planning*, vol. 125, pp. 28–37, 2014.

[338] T. Van Renterghem and D. Botteldooren, "View on outdoor vegetation reduces noise annoyance for dwellers near busy roads," *Landscape and urban planning*, vol. 148, pp. 203–215, 2016.

[339] T. Yu, H. Behm, R. Bill, and J. Kang, "Audio-visual perception of new wind parks," *Landscape and Urban planning*, vol. 165, pp. 1–10, 2017.

[340] L. E. Steg, A. E. Van Den Berg, and J. I. De Groot, *Environmental psychology: An introduction*. BPS Blackwell, 2013.

[341] H. Li and S.-K. Lau, "A review of audio-visual interaction on soundscape assessment in urban built environments," *Applied Acoustics*, vol. 166, p. 107372, 2020.

[342] 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.

[343] A. Mitchell, T. Oberman, F. Aletta, M. Erfanian, M. Kachlicka, M. Lionello, and J. Kang, "The soundscape indices (ssid) protocol: A method for urban soundscape surveys—questionnaires with acoustical and contextual information," *Applied Sciences*, vol. 10, no. 7, p. 2397, 2020.

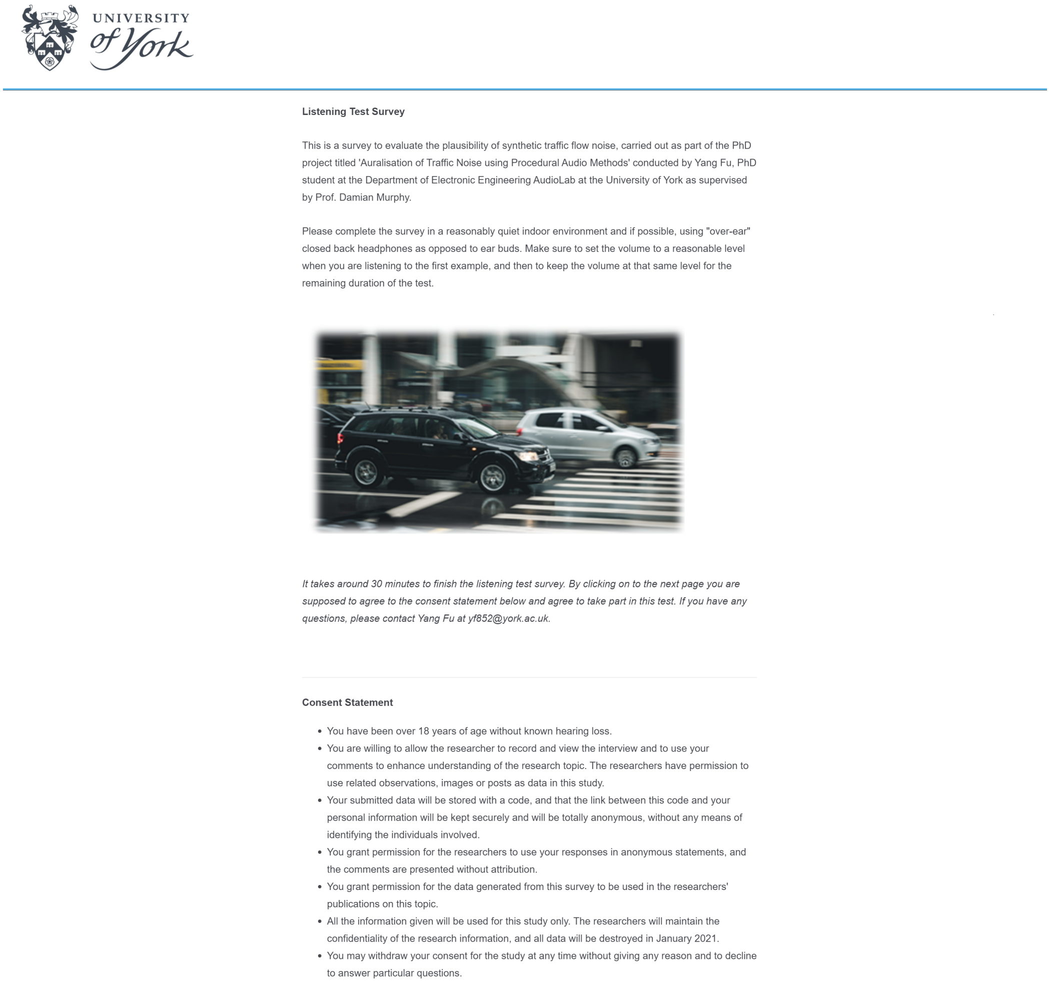
# Appendix A

# Abbreviations

| | |
|---|---|
| AR | Augmented Reality |
| BDC | Bottom Dead Centre |
| BRIR | Binaural Room Impulse Response |
| DFT | Discrete Fourier Transform |
| END | European Noise Directive |
| FEM | Finite Element Method |
| FM | Frequency Modulation |
| GAM | Geometrical Acoustic Methods |
| HCI | Human-Computer Interaction |
| HRTF | Head-Related-Transfer-Function |
| IDFT | Inverse Discrete Fourier Transform |
| IEC | International Electrotechnical Commission |
| ILD | Interaural Level Difference |
| ISO | International Standards Organisation |
| ITD | Interaural Time Difference |
| LTI | Linear Time-Invariant System |
| MIDI | Musical Instrument Digital Interface |
| MR | Mixed Reality |
| MTF | Modulation Transfer Function |
| MUSHRA | Multiple Stimuli with Hidden Reference and Anchor |
| PA | Public Address |
| PD | Pure Data |
| PTS | Permanent Threshold Shift |
| RPM | Revolutions per Minute |
| RT | Reverberation Time |
| SAM | Self-Assessment Manikin |
| SAQI | Spatial Audio Quality Inventory |
| SD | Semantic Differential |
| SIL | Sound Intensity Level |
| SNR | Signal-to-Noise Ratio |
| SPL | Sound Pressure Level |
| TDC | Top Dead Centre |
| TM | Tinnitus Masking |
| TTS | Temporary Hearing Loss |
| VR | Virtual Reality |
| VVT | Variable Valve Timing |
| WFS | Wave Field Synthesis |
| WHO | World Health Organisation |

# Appendix B

# Introduction and Consent Statement for the listening tests



UNIVERSITY of York

**Listening Test Survey**

This is a survey to evaluate the plausibility of synthetic traffic flow noise, carried out as part of the PhD project titled 'Auralisation of Traffic Noise using Procedural Audio Methods' conducted by Yang Fu, PhD student at the Department of Electronic Engineering AudioLab at the University of York as supervised by Prof. Damian Murphy.

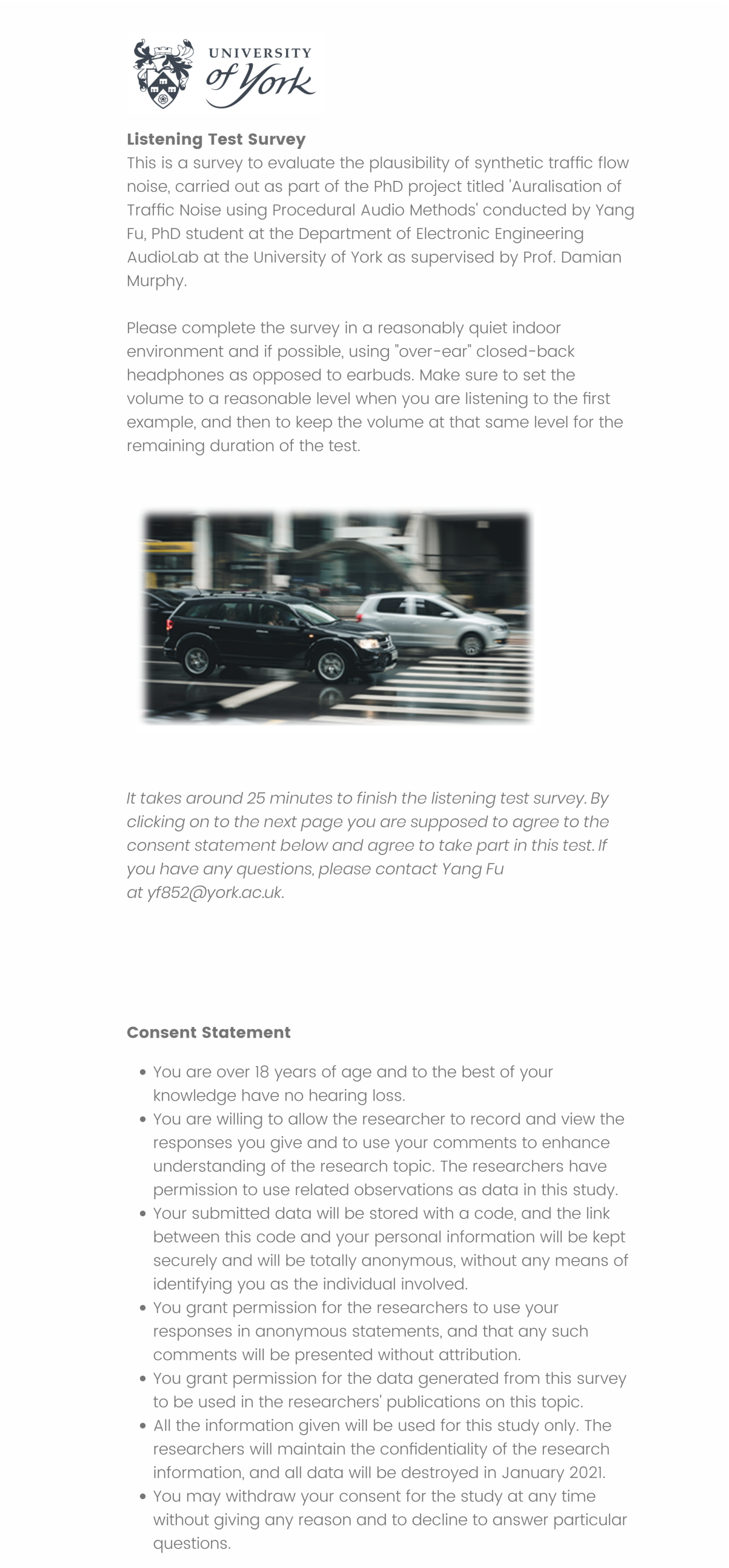
Please complete the survey in a reasonably quiet indoor environment and if possible, using "over-ear" closed back headphones as opposed to ear buds. Make sure to set the volume to a reasonable level when you are listening to the first example, and then to keep the volume at that same level for the remaining duration of the test.

*It takes around 30 minutes to finish the listening test survey. By clicking on to the next page you are supposed to agree to the consent statement below and agree to take part in this test. If you have any questions, please contact Yang Fu at yf852@york.ac.uk.*

**Consent Statement**

- You have been over 18 years of age without known hearing loss.
- You are willing to allow the researcher to record and view the interview and to use your comments to enhance understanding of the research topic. The researchers have permission to use related observations, images or posts as data in this study.
- Your submitted data will be stored with a code, and that the link between this code and your personal information will be kept securely and will be totally anonymous, without any means of identifying the individuals involved.
- You grant permission for the researchers to use your responses in anonymous statements, and the comments are presented without attribution.
- You grant permission for the data generated from this survey to be used in the researchers' publications on this topic.
- All the information given will be used for this study only. The researchers will maintain the confidentiality of the research information, and all data will be destroyed in January 2021.
- You may withdraw your consent for the study at any time without giving any reason and to decline to answer particular questions.

**Figure B.1:** *Introduction and Consent Statement for the ABX listening test.*



**Listening Test Survey**
This is a survey to evaluate the plausibility of synthetic traffic flow noise, carried out as part of the PhD project titled 'Auralisation of Traffic Noise using Procedural Audio Methods' conducted by Yang Fu, PhD student at the Department of Electronic Engineering AudioLab at the University of York as supervised by Prof. Damian Murphy.

Please complete the survey in a reasonably quiet indoor environment and if possible, using "over-ear" closed-back headphones as opposed to earbuds. Make sure to set the volume to a reasonable level when you are listening to the first example, and then to keep the volume at that same level for the remaining duration of the test.

*It takes around 25 minutes to finish the listening test survey. By clicking on to the next page you are supposed to agree to the consent statement below and agree to take part in this test. If you have any questions, please contact Yang Fu at yf852@york.ac.uk.*

**Consent Statement**

- You are over 18 years of age and to the best of your knowledge have no hearing loss.
- You are willing to allow the researcher to record and view the responses you give and to use your comments to enhance understanding of the research topic. The researchers have permission to use related observations as data in this study.
- Your submitted data will be stored with a code, and the link between this code and your personal information will be kept securely and will be totally anonymous, without any means of identifying you as the individual involved.
- You grant permission for the researchers to use your responses in anonymous statements, and that any such comments will be presented without attribution.
- You grant permission for the data generated from this survey to be used in the researchers' publications on this topic.
- All the information given will be used for this study only. The researchers will maintain the confidentiality of the research information, and all data will be destroyed in January 2021.
- You may withdraw your consent for the study at any time without giving any reason and to decline to answer particular questions.

**Figure B.2:** *Introduction and Consent Statement for the MUSHRA listening test.*

# Appendix C

# Listening test Data

## C.1 Listening Test Data – ABX test

Table C.1: The count number for the correct selection of ‘X’ from ‘A’ or ‘B’ for every participant

| **Subject Number** | **Count of correct ‘X’ pick-up (20 trials for each test case)** | | |
|---|---|---|---|
| | **Case 1: Engine Sound** | **Case 2: Engine+Tyre Sound** | **Case 3: Pass-by Sound** |
| Subject 1 | 20 | 14 | 11 |
| Subject 2 | 19 | 13 | 10 |
| Subject 3 | 20 | 12 | 11 |
| Subject 4 | 20 | 16 | 12 |
| Subject 5 | 20 | 18 | 14 |
| Subject 6 | 19 | 13 | 11 |
| Subject 7 | 19 | 15 | 12 |
| Subject 8 | 20 | 16 | 13 |
| Subject 9 | 20 | 12 | 12 |
| Subject 10 | 20 | 13 | 12 |
| Subject 11 | 20 | 13 | 10 |
| Subject 12 | 20 | 15 | 12 |
| Subject 13 | 19 | 15 | 14 |
| Subject 14 | 20 | 14 | 14 |
| Subject 15 | 20 | 13 | 13 |
| Subject 16 | 20 | 15 | 13 |
| Subject 17 | 20 | 16 | 16 |
| Subject 18 | 19 | 18 | 15 |
| Subject 19 | 20 | 14 | 12 |
| Subject 20 | 20 | 15 | 10 |
| Subject 21 | 20 | 19 | 13 |
| Subject 22 | 20 | 18 | 12 |
| Subject 23 | 20 | 17 | 10 |
| Subject 24 | 20 | 13 | 11 |
| Subject 25 | 20 | 16 | 14 |
| Subject 26 | 20 | 17 | 16 |
| Subject 27 | 20 | 16 | 14 |
| Total Count | 535 | 406 | 337 |
| Mean Count | 19.81 | 15.04 | 12.48 |
| Mean Rate | 99.1% | 75.2% | 62.4% |

Table C.2: Probability that a subject is guessing

| **Subject Number** | **Probability that a subject is guessing** | | |
|---|---|---|---|
| | **Case 1: Engine Sound** | **Case 2: Engine+Tyre Sound** | **Case 3: Pass-by Sound** |
| Subject 1 | $9.54 \times 10^{-7}$ | 0.037 | 0.16 |
| Subject 2 | $1.91 \times 10^{-5}$ | 0.074 | 0.18 |
| Subject 3 | $9.54 \times 10^{-7}$ | 0.12 | 0.16 |
| Subject 4 | $9.54 \times 10^{-7}$ | $4.62 \times 10^{-3}$ | 0.12 |
| Subject 5 | $9.54 \times 10^{-7}$ | $1.81 \times 10^{-4}$ | 0.037 |
| Subject 6 | $1.91 \times 10^{-5}$ | 0.074 | 0.16 |
| Subject 7 | $1.91 \times 10^{-5}$ | 0.015 | 0.12 |
| Subject 8 | $9.54 \times 10^{-7}$ | $4.62 \times 10^{-3}$ | 0.074 |
| Subject 9 | $9.54 \times 10^{-7}$ | 0.12 | 0.12 |
| Subject 10 | $9.54 \times 10^{-7}$ | 0.074 | 0.12 |
| Subject 11 | $9.54 \times 10^{-7}$ | 0.074 | 0.18 |
| Subject 12 | $9.54 \times 10^{-7}$ | 0.015 | 0.12 |
| Subject 13 | $1.91 \times 10^{-5}$ | 0.015 | 0.037 |
| Subject 14 | $9.54 \times 10^{-7}$ | 0.037 | 0.037 |
| Subject 15 | $9.54 \times 10^{-7}$ | 0.074 | 0.074 |
| Subject 16 | $9.54 \times 10^{-7}$ | 0.015 | 0.074 |
| Subject 17 | $9.54 \times 10^{-7}$ | $4.62 \times 10^{-3}$ | $4.62 \times 10^{-3}$ |
| Subject 18 | $1.91 \times 10^{-5}$ | $1.81 \times 10^{-4}$ | 0.015 |
| Subject 19 | $9.54 \times 10^{-7}$ | 0.037 | 0.12 |
| Subject 20 | $9.54 \times 10^{-7}$ | 0.015 | 0.18 |
| Subject 21 | $9.54 \times 10^{-7}$ | $1.91 \times 10^{-5}$ | 0.074 |
| Subject 22 | $9.54 \times 10^{-7}$ | $1.81 \times 10^{-4}$ | 0.12 |
| Subject 23 | $9.54 \times 10^{-7}$ | $1.09 \times 10^{-3}$ | 0.18 |
| Subject 24 | $9.54 \times 10^{-7}$ | 0.074 | 0.16 |
| Subject 25 | $9.54 \times 10^{-7}$ | $4.62 \times 10^{-3}$ | 0.037 |
| Subject 26 | $9.54 \times 10^{-7}$ | $1.09 \times 10^{-3}$ | $4.62 \times 10^{-3}$ |
| Subject 27 | $9.54 \times 10^{-7}$ | $4.62 \times 10^{-3}$ | 0.037 |

Table C.3: The total points of 'plausibility to expectation' given by the participants

| **Subject Number** | **The total points of 'plausibility to expectation'** | |
|---|---|---|
| | **Procedural Audio** | **Granular synthesis** |
| Subject 1 | 62 | 58 |
| Subject 2 | 59 | 61 |
| Subject 3 | 56 | 64 |
| Subject 4 | 58 | 62 |
| Subject 5 | 54 | 66 |
| Subject 6 | 57 | 63 |
| Subject 7 | 64 | 56 |
| Subject 8 | 62 | 58 |
| Subject 9 | 62 | 58 |
| Subject 10 | 57 | 63 |
| Subject 11 | 53 | 67 |
| Subject 12 | 63 | 57 |
| Subject 13 | 64 | 56 |
| Subject 14 | 57 | 63 |
| Subject 15 | 58 | 62 |
| Subject 16 | 60 | 60 |
| Subject 17 | 56 | 64 |
| Subject 18 | 58 | 62 |
| Subject 19 | 58 | 62 |
| Subject 20 | 59 | 61 |
| Subject 21 | 57 | 63 |
| Subject 22 | 54 | 66 |
| Subject 23 | 62 | 58 |
| Subject 24 | 63 | 57 |
| Subject 25 | 57 | 63 |
| Subject 26 | 58 | 62 |
| Subject 27 | 63 | 57 |
| Mean Score | 58.93 | 61.07 |
| Standard Deviation | 3.18 | 3.18 |

# C.2 Listening Test Data – MUSHRA test

Table C.4: The rating of stimuli in MUSHRA test Case 1

| **Subject Number** | **Stimuli** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **P6** | **G1P5** | **G2P4** | **G3P3** | **G4P2** | **G5P1** | **G6 (HR)** | **Anchor** |
| Subjective 1 | 62.00 | 61.00 | 68.00 | 63.00 | 57.00 | 73.00 | 86.00 | 33.00 |
| Subjective 2 | 64.00 | 67.00 | 66.00 | 66.00 | 64.00 | 63.00 | 78.00 | 2.00 |
| Subjective 3 | 80.00 | 81.00 | 82.00 | 78.00 | 80.00 | 84.00 | 95.00 | 18.00 |
| Subjective 4 | 50.00 | 58.00 | 50.00 | 58.00 | 57.00 | 62.00 | 80.00 | 10.00 |
| Subjective 5 | 52.00 | 53.00 | 50.00 | 56.00 | 59.00 | 55.00 | 79.00 | 12.00 |
| Subjective 6 | 60.00 | 40.00 | 49.00 | 51.00 | 42.00 | 43.00 | 90.00 | 3.00 |
| Subjective 7 | 60.00 | 60.00 | 55.00 | 62.00 | 63.00 | 62.00 | 89.00 | 2.00 |
| Subjective 8 | 42.00 | 51.00 | 48.00 | 63.00 | 67.00 | 69.00 | 80.00 | 20.00 |
| Subjective 9 | 74.00 | 69.00 | 79.00 | 76.00 | 72.00 | 74.00 | 82.00 | 5.00 |
| Subjective 10 | 40.00 | 39.00 | 32.00 | 35.00 | 36.00 | 34.00 | 79.00 | 18.00 |
| Subjective 11 | 58.00 | 60.00 | 57.00 | 65.00 | 59.00 | 67.00 | 75.00 | 10.00 |
| Subjective 12 | 76.00 | 71.00 | 66.00 | 59.00 | 59.00 | 73.00 | 82.00 | 30.00 |
| Subjective 13 | 69.00 | 62.00 | 74.00 | 67.00 | 57.00 | 55.00 | 84.00 | 9.00 |
| Subjective 14 | 40.00 | 47.00 | 40.00 | 49.00 | 48.00 | 44.00 | 68.00 | 2.00 |
| Subjective 15 | 69.00 | 61.00 | 57.00 | 62.00 | 63.00 | 70.00 | 98.00 | 18.00 |
| Subjective 16 | 60.00 | 70.00 | 65.00 | 65.00 | 59.00 | 67.00 | 80.00 | 10.00 |
| Subjective 17 | 73.00 | 77.00 | 76.00 | 75.00 | 69.00 | 74.00 | 99.00 | 30.00 |
| Subjective 18 | 61.00 | 52.00 | 59.00 | 53.00 | 54.00 | 59.00 | 81.00 | 9.00 |
| Subjective 19 | 35.00 | 53.00 | 40.00 | 39.00 | 44.00 | 39.00 | 99.00 | 2.00 |
| Subjective 20 | 22.00 | 24.00 | 30.00 | 25.00 | 24.00 | 23.00 | 31.0 | 11.00 |

Table C.5: The rating of stimuli in MUSHRA test Case 2

| **Subject Number** | **Stimuli** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **P6** | **G1P5** | **G2P4** | **G3P3 (HR)** | **G4P2** | **G5P1** | **G6** | **Anchor** |
| Subjective 1 | 59.00 | 67.00 | 68.00 | 77.00 | 62.00 | 69.00 | 70.00 | 20.00 |
| Subjective 2 | 58.00 | 65.00 | 66.00 | 80.00 | 63.00 | 63.00 | 64.00 | 3.00 |
| Subjective 3 | 82.00 | 74.00 | 80.00 | 95.00 | 75.00 | 77.00 | 75.00 | 25.00 |
| Subjective 4 | 53.00 | 57.00 | 59.00 | 88.00 | 54.00 | 51.00 | 47.00 | 13.00 |
| Subjective 5 | 54.00 | 56.00 | 51.00 | 90.00 | 57.00 | 56.00 | 55.00 | 15.00 |
| Subjective 6 | 48.00 | 36.00 | 39.00 | 85.00 | 43.00 | 37.00 | 45.00 | 5.00 |
| Subjective 7 | 62.00 | 63.00 | 64.00 | 86.00 | 57.00 | 56.00 | 60.00 | 1.00 |
| Subjective 8 | 48.00 | 42.00 | 43.00 | 79.00 | 39.00 | 45.00 | 50.00 | 12.00 |
| Subjective 9 | 73.00 | 70.00 | 64.00 | 84.00 | 70.00 | 66.00 | 65.00 | 3.00 |
| Subjective 10 | 35.00 | 40.00 | 32.00 | 92.00 | 38.00 | 37.00 | 39.00 | 12.00 |
| Subjective 11 | 59.00 | 56.00 | 55.00 | 86.00 | 58.00 | 67.00 | 65.00 | 11.00 |
| Subjective 12 | 73.00 | 73.00 | 72.00 | 90.00 | 69.00 | 66.00 | 72.00 | 25.00 |
| Subjective 13 | 61.00 | 68.00 | 70.00 | 95.00 | 59.00 | 55.00 | 60.00 | 15.00 |
| Subjective 14 | 52.00 | 54.00 | 55.00 | 89.00 | 47.00 | 49.00 | 50.00 | 3.00 |
| Subjective 15 | 59.00 | 67.00 | 63.00 | 77.00 | 58.00 | 51.00 | 60.00 | 23.00 |
| Subjective 16 | 65.00 | 77.00 | 66.00 | 83.00 | 61.00 | 61.00 | 62.00 | 11.00 |
| Subjective 17 | 75.00 | 72.00 | 69.00 | 98.00 | 70.00 | 73.00 | 74.00 | 29.00 |
| Subjective 18 | 54.00 | 51.00 | 55.00 | 79.00 | 49.00 | 50.00 | 54.00 | 12.00 |
| Subjective 19 | 40.00 | 43.00 | 43.00 | 90.00 | 44.00 | 37.00 | 42.00 | 5.00 |
| Subjective 20 | 26.00 | 29.00 | 25.00 | 34.00 | 22.00 | 29.00 | 23.00 | 9.00 |

Table C.6: The rating of stimuli in MUSHRA test Case 3

| **Subject Number** | **Stimuli** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **P6 (HR)** | **G1P5** | **G2P4** | **G3P3** | **G4P2** | **G5P1** | **G6** | **Anchor** |
| Subjective 1 | 88.00 | 68.00 | 66.00 | 59.00 | 52.00 | 61.00 | 62.00 | 19.00 |
| Subjective 2 | 79.00 | 70.00 | 76.00 | 77.00 | 70.00 | 69.00 | 75.00 | 4.00 |
| Subjective 3 | 90.00 | 72.00 | 79.00 | 69.00 | 65.00 | 74.00 | 76.00 | 20.00 |
| Subjective 4 | 87.00 | 50.00 | 56.00 | 62.00 | 76.00 | 70.00 | 69.00 | 11.00 |
| Subjective 5 | 86.00 | 47.00 | 49.00 | 50.00 | 65.00 | 67.00 | 54.00 | 16.00 |
| Subjective 6 | 81.00 | 28.00 | 31.00 | 29.00 | 48.00 | 35.00 | 36.00 | 7.00 |
| Subjective 7 | 79.00 | 53.00 | 59.00 | 62.00 | 61.00 | 67.00 | 65.00 | 3.00 |
| Subjective 8 | 80.00 | 39.00 | 44.00 | 35.00 | 28.00 | 40.00 | 41.00 | 11.00 |
| Subjective 9 | 89.00 | 68.00 | 66.00 | 59.00 | 60.00 | 61.00 | 69.00 | 5.00 |
| Subjective 10 | 79.00 | 36.00 | 29.00 | 45.00 | 51.00 | 35.00 | 33.00 | 11.00 |
| Subjective 11 | 75.00 | 55.00 | 54.00 | 52.00 | 59.00 | 57.00 | 69.00 | 10.00 |
| Subjective 12 | 91.00 | 76.00 | 70.00 | 69.00 | 75.00 | 73.00 | 71.00 | 22.00 |
| Subjective 13 | 98.00 | 62.00 | 70.00 | 65.00 | 55.00 | 59.00 | 63.00 | 14.00 |
| Subjective 14 | 76.00 | 48.00 | 44.00 | 59.00 | 55.00 | 56.00 | 53.00 | 4.00 |
| Subjective 15 | 80.00 | 69.00 | 72.00 | 70.00 | 57.00 | 59.00 | 61.00 | 20.00 |
| Subjective 16 | 79.00 | 59.00 | 56.00 | 55.00 | 61.00 | 63.00 | 65.00 | 16.00 |
| Subjective 17 | 92.00 | 57.00 | 59.00 | 44.00 | 60.00 | 55.00 | 49.00 | 22.00 |
| Subjective 18 | 91.00 | 57.00 | 59.00 | 44.00 | 60.00 | 55.00 | 49.00 | 22.00 |
| Subjective 19 | 90.00 | 39.00 | 35.00 | 44.00 | 46.00 | 39.00 | 43.00 | 9.00 |
| Subjective 19 | 54.00 | 28.00 | 26.00 | 37.00 | 36.00 | 33.00 | 29.00 | 16.00 |

Table C.7: The rating of stimuli in MUSHRA test Case 4

| **Subject Number** | **Stimuli** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **P6** | **G1P5** | **G2P4** | **G3P3 (HR)** | **G4P2** | **G5P1** | **G6** | **Anchor** |
| Subjective 1 | 47.00 | 57.00 | 58.00 | 81.00 | 67.00 | 72.00 | 71.00 | 25.00 |
| Subjective 2 | 38.00 | 45.00 | 46.00 | 72.00 | 49.00 | 47.00 | 53.00 | 13.00 |
| Subjective 3 | 75.00 | 74.00 | 71.00 | 98.00 | 75.00 | 77.00 | 75.00 | 14.00 |
| Subjective 4 | 55.00 | 59.00 | 58.00 | 86.00 | 62.00 | 60.00 | 69.00 | 13.00 |
| Subjective 5 | 53.00 | 46.00 | 43.00 | 76.00 | 61.00 | 58.00 | 57.00 | 16.00 |
| Subjective 6 | 59.00 | 56.00 | 49.00 | 89.00 | 62.00 | 67.00 | 64.00 | 15.00 |
| Subjective 7 | 71.00 | 68.00 | 69.00 | 65.00 | 67.00 | 66.00 | 65.00 | 13.00 |
| Subjective 8 | 38.00 | 42.00 | 44.00 | 79.00 | 59.00 | 65.00 | 58.00 | 22.00 |
| Subjective 9 | 79.00 | 77.00 | 76.00 | 88.00 | 74.00 | 76.00 | 77.00 | 9.00 |
| Subjective 10 | 59.00 | 60.00 | 55.00 | 76.00 | 74.00 | 77.00 | 76.00 | 16.00 |
| Subjective 11 | 54.00 | 47.00 | 49.00 | 86.00 | 58.00 | 57.00 | 64.00 | 18.00 |
| Subjective 12 | 71.00 | 76.00 | 77.00 | 92.00 | 69.00 | 74.00 | 74.00 | 21.00 |
| Subjective 13 | 52.00 | 54.00 | 58.00 | 99.00 | 69.00 | 75.00 | 70.00 | 19.00 |
| Subjective 14 | 64.00 | 62.00 | 69.00 | 69.00 | 55.00 | 57.00 | 58.00 | 7.00 |
| Subjective 15 | 52.00 | 60.00 | 59.00 | 82.00 | 64.00 | 61.00 | 60.00 | 18.00 |
| Subjective 16 | 71.00 | 69.00 | 66.00 | 89.00 | 73.00 | 71.00 | 70.00 | 13.00 |
| Subjective 17 | 79.00 | 81.00 | 80.00 | 94.00 | 77.00 | 79.00 | 71.00 | 17.00 |
| Subjective 18 | 43.00 | 41.00 | 49.00 | 69.00 | 58.00 | 59.00 | 64.00 | 17.00 |
| Subjective 19 | 53.00 | 55.00 | 52.00 | 92.00 | 64.00 | 67.00 | 67.00 | 16.00 |
| Subjective 20 | 38.00 | 43.00 | 36.00 | 40.00 | 41.00 | 43.00 | 40.00 | 18.00 |

Table C.8: The rating of stimuli in MUSHRA test Case 5

| **Subject Number** | **Stimuli** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **P6** | **G1P5** | **G2P4** | **G3P3 (HR)** | **G4P2** | **G5P1** | **G6** | **Anchor** |
| Subjective 1 | 56.00 | 51.00 | 52.00 | 84.00 | 57.00 | 52.00 | 61.00 | 27.00 |
| Subjective 2 | 49.00 | 55.00 | 56.00 | 89.00 | 49.00 | 51.00 | 52.00 | 23.00 |
| Subjective 3 | 62.00 | 64.00 | 61.00 | 97.00 | 62.00 | 67.00 | 65.00 | 13.00 |
| Subjective 4 | 45.00 | 37.00 | 50.00 | 91.00 | 52.00 | 48.00 | 44.00 | 6.00 |
| Subjective 5 | 59.00 | 48.00 | 51.00 | 84.00 | 61.00 | 70.00 | 70.00 | 12.00 |
| Subjective 6 | 65.00 | 66.00 | 54.00 | 83.00 | 62.00 | 64.00 | 65.00 | 8.00 |
| Subjective 7 | 70.00 | 70.00 | 65.00 | 75.00 | 71.00 | 66.00 | 69.00 | 17.00 |
| Subjective 8 | 58.00 | 49.00 | 57.00 | 86.00 | 59.00 | 53.00 | 55.00 | 14.00 |
| Subjective 9 | 78.00 | 74.00 | 75.00 | 84.00 | 76.00 | 76.00 | 73.00 | 10.00 |
| Subjective 10 | 56.00 | 57.00 | 52.00 | 88.00 | 52.00 | 49.00 | 53.00 | 20.00 |
| Subjective 11 | 67.00 | 62.00 | 60.00 | 79.00 | 68.00 | 63.00 | 65.00 | 17.00 |
| Subjective 12 | 71.00 | 76.00 | 77.00 | 92.00 | 69.00 | 74.00 | 74.00 | 19.00 |
| Subjective 13 | 66.00 | 64.00 | 67.00 | 97.00 | 59.00 | 61.00 | 55.00 | 20.00 |
| Subjective 14 | 64.00 | 69.00 | 68.00 | 67.00 | 65.00 | 62.00 | 59.00 | 8.00 |
| Subjective 15 | 63.00 | 62.00 | 57.00 | 84.00 | 58.00 | 66.00 | 68.00 | 6.00 |
| Subjective 16 | 74.00 | 79.00 | 76.00 | 87.00 | 73.00 | 74.00 | 75.00 | 2.00 |
| Subjective 17 | 59.00 | 70.00 | 80.00 | 96.00 | 79.00 | 76.00 | 72.00 | 18.00 |
| Subjective 18 | 52.00 | 50.00 | 49.00 | 77.00 | 53.00 | 55.00 | 58.00 | 9.00 |
| Subjective 19 | 51.00 | 47.00 | 55.00 | 97.00 | 58.00 | 60.00 | 57.00 | 4.00 |
| Subjective 20 | 40.00 | 51.00 | 56.00 | 55.00 | 53.00 | 54.00 | 41.00 | 8.00 |

Table C.9: The rating of stimuli in MUSHRA test Case 6

| **Subject Number** | **Stimuli** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **P6** | **G1P5** | **G2P4** | **G3P3 (HR)** | **G4P2** | **G5P1** | **G6** | **Anchor** |
| Subjective 1 | 53.00 | 48.00 | 55.00 | 93.00 | 56.00 | 59.00 | 62.00 | 14.00 |
| Subjective 2 | 37.00 | 44.00 | 39.00 | 87.00 | 35.00 | 39.00 | 38.00 | 13.00 |
| Subjective 3 | 72.00 | 68.00 | 65.00 | 96.00 | 63.00 | 75.00 | 74.00 | 19.00 |
| Subjective 4 | 49.00 | 45.00 | 49.00 | 92.00 | 45.00 | 42.00 | 48.00 | 18.00 |
| Subjective 5 | 49.00 | 50.00 | 57.00 | 88.00 | 49.00 | 43.00 | 42.00 | 2.00 |
| Subjective 6 | 69.00 | 65.00 | 61.00 | 76.00 | 61.00 | 58.00 | 60.00 | 7.00 |
| Subjective 7 | 58.00 | 66.00 | 60.00 | 79.00 | 61.00 | 66.00 | 63.00 | 9.00 |
| Subjective 8 | 57.00 | 49.00 | 50.00 | 92.00 | 54.00 | 52.00 | 58.00 | 17.00 |
| Subjective 9 | 75.00 | 68.00 | 67.00 | 91.00 | 70.00 | 66.00 | 61.00 | 6.00 |
| Subjective 10 | 59.00 | 47.00 | 61.00 | 79.00 | 55.00 | 49.00 | 58.00 | 11.00 |
| Subjective 11 | 69.00 | 73.00 | 72.00 | 76.00 | 70.00 | 65.00 | 64.00 | 10.00 |
| Subjective 12 | 68.00 | 66.00 | 75.00 | 80.00 | 70.00 | 71.00 | 69.00 | 8.00 |
| Subjective 13 | 63.00 | 62.00 | 65.00 | 72.00 | 69.00 | 68.00 | 64.00 | 23.00 |
| Subjective 14 | 61.00 | 60.00 | 70.00 | 85.00 | 78.00 | 79.00 | 84.00 | 12.00 |
| Subjective 15 | 63.00 | 72.00 | 67.00 | 94.00 | 68.00 | 66.00 | 59.00 | 17.00 |
| Subjective 16 | 74.00 | 79.00 | 76.00 | 92.00 | 73.00 | 74.00 | 75.00 | 16.00 |
| Subjective 17 | 45.00 | 49.00 | 70.00 | 95.00 | 53.00 | 62.00 | 68.00 | 10.00 |
| Subjective 18 | 52.00 | 50.00 | 49.00 | 91.00 | 53.00 | 55.00 | 58.00 | 9.00 |
| Subjective 19 | 52.00 | 67.00 | 65.00 | 90.00 | 68.00 | 71.00 | 67.00 | 6.00 |
| Subjective 20 | 56.00 | 59.00 | 56.00 | 70.00 | 59.00 | 49.00 | 54.00 | 3.00 |

Table C.10: The rating of stimuli in MUSHRA test Case 7

| **Subject Number** | **Stimuli** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **P6** | **G1P5** | **G2P4** | **G3P3** | **G4P2** | **G5P1** | **G6 (HR)** | **Anchor** |
| Subjective 1 | 68.00 | 74.00 | 75.00 | 88.00 | 77.00 | 70.00 | 62.00 | 7.00 |
| Subjective 2 | 49.00 | 44.00 | 62.00 | 84.00 | 60.00 | 61.00 | 70.00 | 11.00 |
| Subjective 3 | 52.00 | 59.00 | 62.00 | 86.00 | 69.00 | 77.00 | 73.00 | 10.00 |
| Subjective 4 | 63.00 | 39.00 | 48.00 | 93.00 | 46.00 | 50.00 | 68.00 | 6.00 |
| Subjective 5 | 52.00 | 57.00 | 55.00 | 85.00 | 49.00 | 48.00 | 44.00 | 6.00 |
| Subjective 6 | 61.00 | 62.00 | 59.00 | 98.00 | 60.00 | 48.00 | 55.00 | 8.00 |
| Subjective 7 | 50.00 | 56.00 | 49.00 | 76.00 | 53.00 | 57.00 | 59.00 | 11.00 |
| Subjective 8 | 57.00 | 69.00 | 72.00 | 98.00 | 80.00 | 73.00 | 77.00 | 2.00 |
| Subjective 9 | 77.00 | 74.00 | 62.00 | 80.00 | 60.00 | 67.00 | 75.00 | 3.00 |
| Subjective 10 | 69.00 | 70.00 | 63.00 | 74.00 | 61.00 | 68.00 | 58.00 | 14.00 |
| Subjective 11 | 63.00 | 70.00 | 71.00 | 92.00 | 67.00 | 64.00 | 59.00 | 12.00 |
| Subjective 12 | 59.00 | 60.00 | 77.00 | 89.00 | 73.00 | 70.00 | 69.00 | 21.00 |
| Subjective 13 | 60.00 | 66.00 | 69.00 | 76.00 | 68.00 | 70.00 | 79.00 | 12.00 |
| Subjective 14 | 79.00 | 68.00 | 70.00 | 81.00 | 77.00 | 76.00 | 69.00 | 13.00 |
| Subjective 15 | 52.00 | 61.00 | 70.00 | 88.00 | 56.00 | 68.00 | 66.00 | 13.00 |
| Subjective 16 | 77.00 | 74.00 | 53.00 | 87.00 | 71.00 | 76.00 | 54.00 | 11.00 |
| Subjective 17 | 77.00 | 74.00 | 53.00 | 87.00 | 71.00 | 49.00 | 58.00 | 8.00 |
| Subjective 18 | 21.00 | 29.00 | 37.00 | 80.00 | 35.00 | 40.00 | 28.00 | 6.00 |
| Subjective 19 | 66.00 | 69.00 | 59.00 | 86.00 | 48.00 | 65.00 | 62.00 | 6.00 |
| Subjective 20 | 29.00 | 34.00 | 40.00 | 61.00 | 35.00 | 36.00 | 48.00 | 7.00 |

# Appendix D

# Digital Assets

The following items are presented as supplementary materials for this thesis.

### Audio Files

This folder contains all stimuli for the listening tests conducted in this study.

**1) ABX Listening Test Material** This folder contains the stimuli used for the ABX listening test. Three sub-folders are included, corresponding to the stimuli in test Case 1–3.

**2) MUSHRA Listening Test Material** This folder contains the stimuli used for the MUSHRA listening test. Seven sub-folders are included, corresponding to the stimuli in test Case 1–7.

### Computer Code

This folder contains the main codes for the implementation of the auralisation framework for traffic flow sounds proposed in this thesis.

**1) Simulation of engine sound synthesis** This folder contains the engine sound synthesis algorithm developed in this thesis. The simulation of this algorithm is presented as a PD patch, with the engine sound synthesizer written in C programming language and used as PD objects. The code is developed based on [264, 274] following the open-source license (GPLv3).

**2) Simulation of tyre sound synthesis** This folder contains the Matlab scripts for the tyre sound synthesis algorithm developed in this thesis.

**3) Simulation of sound propagation effects** This folder contains the the Matlab scripts for the distance attenuation and the Doppler shift algorithms developed in this thesis.

**4) Simulation of low frequency correction for HRTFs** This folder contains the Matlab scripts for the low frequency correction algorithm for the KEMAR HRTFs dataset developed in this thesis.