UNIVERSITY OF SOUTHAMPTON

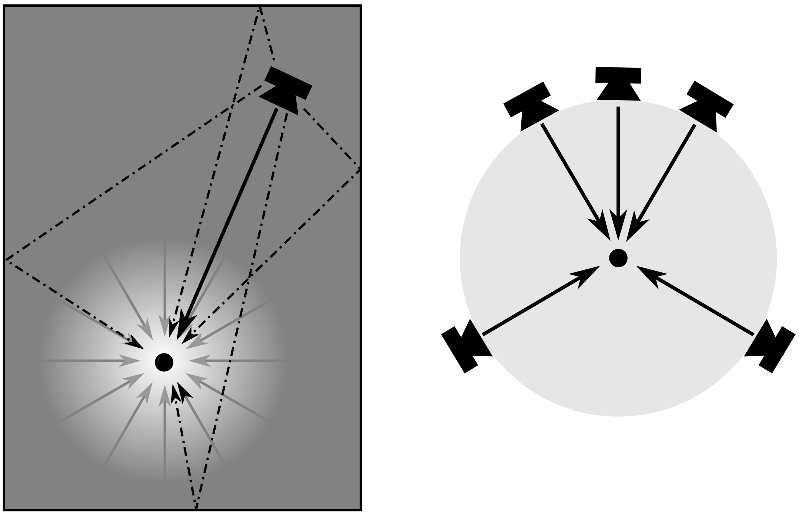
FACULTY OF ENGINEERING AND THE ENVIRONMENT

INSTITUTE OF SOUND AND VIBRATION RESEARCH

The Diffuse Sound Object

by

Michael Patrick Cousins



A thesis submitted in partial fulfilment for the degree of Doctor of Philosophy

April 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND THE ENVIRONMENT

Institute of Sound and Vibration Research

Doctor of Philosophy

THE DIFFUSE SOUND OBJECT

By Michael Patrick Cousins

The theoretical diffuse sound field is one that is generated using an infinite number of uncorrelated plane waves from all directions. Spatial audio, in contrast, is generally delivered using a finite number of loudspeakers, potentially as few as only two. Therefore there is a reduction in the diffuseness of the sound field when reproducing diffuse sound fields using loudspeakers. It is therefore important that sound fields such as reverberation, rain or audience noise–that are approximations of a theoretical diffuse sound field–are reproduced with minimal perceptual degradation to maximise the spatial audio experience. In this research, the perception of diffuseness has been investigated thoroughly using a series of listening tests. Additionally, the relationships between the subjective perceived diffuseness and objective metrics for measuring diffuseness have been investigated leading to a simple metric for perceived diffuseness based on the spatial coherence.

Contents

List of Figures

List of Tables

Academic Thesis: Declaration of Authorship

I, Michael Patrick Cousins, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

The Diffuse Sound Object

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;

2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;

3. Where I have consulted the published work of others, this is always clearly attributed;

4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;

5. I have acknowledged all main sources of help;

6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

7. Parts of this work have been published as:

   M. Cousins, F. M. Fazi, S. Bleeck, and F. Melchior, "Maximising perceived diffuseness in loudspeaker systems with height using optimised relative loudspeaker levels," *Institute of Acoustics: Reproduced Sound*, 2015.

   M. Cousins, F. Fazi, S. Bleeck, and F. Melchior, "Subjective Diffuseness in Layer-based Loudspeaker Systems with Height," *Audio Engineering Society 139th Convention,* New York, 2015.

   M. Cousins, S. Bleeck, F. Melchior, and F. Fazi, "Relation between acoustic measurements and the perceived diffuseness of a synthesised sound field," *22nd International Congress on Acoustics*, Buenos Aires, 2016.

   M. Cousins, S. Bleeck, F. Melchior, and F. Fazi, "The Effect of Inter-channel Correlation Coefficient on Perceived Diffuseness," *4th International Conference on Spatial Audio*, Graz, 2017

Signed:

Date:

Acknowledgements

I would like to thank people not just because of the work they have put into–directly or indirectly–helping me finish my PhD, but also for the roles in which they have played making my life enjoyable over the course of the last 4 years.

First and foremost I must thank my supervisors and VAAE colleagues. All of you are experts in your respective fields and so the technical knowledge that you have imparted over the years is a given. However additionally... Andrew, Jon and Chris at the BBC, thank you for keeping the project moving in the right direction. Frank, I would always feel excited and motivated after every meeting thanks your endless enthusiasm. Stefan, thank you for your encouragement and support. And finally, Filippo, thank you for being the driving force behind my PhD. Thank you for being there week in, week out, answering the technical questions with a smile as well as helping with all the paperwork. You have a great work ethic and have put far more energy and time into my PhD than you were required to or had to spare. This was truly appreciated.

As for my general enjoyment of my time over the last four years; that is entirely thanks to my family and friends. It is impossible to get stressed or anxious when you know your family at home are so supportive. As for my friends and housemates who have also made the last four years so enjoyable, I will simply say thank you and list you in alphabetical order as I could not possibly pick favourites. Cara, Chris, Dermot, Emma, Horacio, Hugo, James, John Sharpe, Juan, Kat, Laura, Lee-Anne, Lizzy, Marine, Mehdy, Michaela, Nathan, Nicola, Raed, Richard, Sarah, Seany, Stacey, Ste and Thalia amongst others.

And finally and most importantly, I need to thank my perfect girlfriend Ana. You are simultaneously the primary reason that I am so glad I decided to do a PhD in Southampton, and the primary reason I was able to complete it. Whilst the people I have met, the things I have learned and the qualifications I have gained during my time in Southampton have made the experience enjoyable, it is meeting you that has made my PhD an unforgettable experience that is the most important decision I have made in my life so far. Somehow, you managed to make me work harder but at the same time have more fun that I thought possible. I have enjoyed every part of my PhD more, because you were a part of it. I love you very much and thank you.

Chapter 1

Introduction

1.1 Overview

This project “The Diffuse Sound Object” is an investigation into the perception of diffuseness in reproduced audio for application in next generation, object-based spatial audio.

This introductory chapter covers the basic motivation for this research and lays out the main contributions of the research. Firstly, the basic concepts of object-based spatial audio are described and why diffuse sound fields deserve particular attention. The following section looks at how this research provides further insight into reproduced diffuse sound fields. And the final section provides a detailed overview of the structure of this thesis.

1.2 Spatial Audio, Object-based Audio and Diffuseness

High quality reproduced audio is dependent not only on the ability of a system to produce a wide dynamic range over a wide frequency range. We as humans can also gain a large amount of spatial information from what we hear, often called spatial impression. We get an indication of where sounds are coming from, their size and the acoustic environment.

A typical signal chain for spatial audio is shown in figure 1.1. A source is captured, processed, reproduced over loudspeakers and the resultant sound field is sampled by the ears of the listener who perceives some spatial information about the source. There are many factors that determine how a sound field is perceived by a listener and there are many factors in the signal chain that will determine the reproduced sound field.

The reproduced sound field is dependent on the source properties and recording environment such as the source’s frequency content, timbre and directivity, the distance to the microphone(s) and reverberation of the recording environment. The microphones used, their arrangement and directivity as well as any processing applied such as panning,

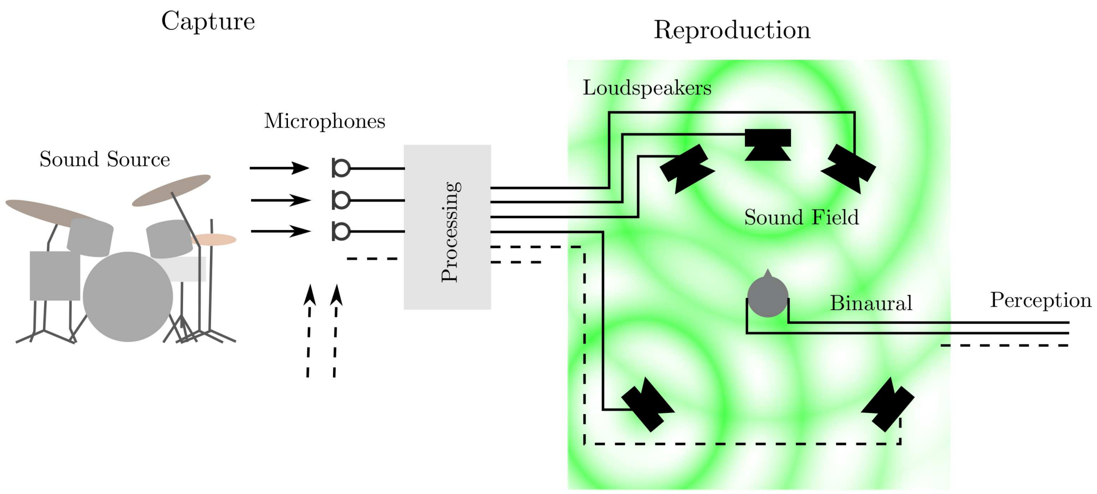


Figure 1.1: Signal chain of spatial audio.

compression, convolution and equalisation. The loudspeakers directivity, gains and arrangement as well as the listening room acoustics will all affect the reproduced sound field. It is this reproduced sound field that is then sampled binaurally by the listener. Aural and interaural cues contained within this binaural signal along with other cues such as visual cues and head rotation cues are combined by the brain to form the final complete spatial impression of the sound field.

Historically, spatial audio has been channel-based with complex scenes composed of many sources rendered to a few loudspeaker channels for loudspeakers in known positions relative to the listener. Microphone techniques, panning algorithms and loudspeaker layouts are implemented so the acoustic signal at the ears of a listener translate to aural and interaural cues that relate to the desired spatial impression. This channel-based approach allows good spatial impression from relatively few channels. However, if the reproduction system does not match the sound engineer's listening system, then the spatial information is distorted. The listener is also not able to alter the way in which the component parts of the signal are combined to account for listener preference or listening environment.

The alternative is object-based audio formats that transmit separate audio objects with metadata describing the location/size of the source. The reproduction system can render the auditory scene to give an accurate spatial representation of the scene based on the system's available resources. This may be a standardised layout of loudspeakers like 2-channel stereo, 5.1 or 22.2; an arbitrary layout of loudspeakers; a binaural headphone system or a wave field synthesis system. Object-based audio additionally allows user interaction. Adjustment of the sound scene based on listener preference and listening environment. For example louder voice for the hard of hearing, efficient different languages, or adjusting source positioning for different screen sizes in audio visual material.

This object-based approach works well for point sources that are easily modelled as a single mono object with a position. However, this is limiting when describing more spatially complex sources and sound fields. Signals such as reverberation, rain and audience applause can be seen as many distributed sound sources that cannot be described solely by a single mono audio object with a position. For example a grand piano has a large soundboard that resonates producing the sound. A single mono stream cannot describe this. A sound engineer would typically use multiple microphones panned to different positions to give a sense of the large source. As the source becomes larger the number of separate mono streams required to represent it increases. As the sound field produced by a source approaches that of a diffuse field (e.g. late reverberation) the sound field can still be approximated using multiple sources positioned on a sphere around the listener however this is becomes inefficient to encode. The theoretical diffuse sound field is one that is composed of an infinite number of uncorrelated plane waves coming from all directions simultaneously. For real sound sources, a point source can be seen as not diffuse, and the sound field in a reverberation room used for diffuse sound field measurements can be seen as highly diffuse. These highly diffuse sound fields are the focus of this research as many of the desirable attributes associated with spatial impression are dependent on the ability of a sound system to represent the size and extent of a source, the size and extent of its environment, in addition to its location.

1.3 This Research

The aim of this research was to better understand the perception of diffuse sound fields and especially the salient components of a diffuse sound field that affect the perception of diffuseness when reproducing diffuse sound fields using loudspeakers. From this understanding, optimal ways of representing diffuse sound fields in object-based audio can be developed along with spatial audio tools for use by recording engineers. This was achieved using a series of subjective listening tests in combination with a range of objective sound field measurement techniques. This led to the development of an algorithm for the prediction of perceived diffuseness that in turn provides guidance on methods to maximise the perceived diffuseness.

The series of listening tests were designed to vary the sound field and elicit the change in perceived diffuseness. These experiments allow a broad understanding of the perception of diffuseness and provided data that could be compared to objective quantities of the sound field. A range of methods for physically measuring sound field diffuseness were tested and the coherence between spatially separated points was found to correlate well with the perception of diffuseness in the majority of cases. From this, a novel metric of perceived diffuseness was developed. A further listening test was performed to optimise the parameters of the metric and finally the metric was critically tested and the limitations of the metric discussed. The metric is simple, runs in real-time and is robust to factors such as slight listener movements and the frequency content of the input material. The nature of the metric reveals possible methods to maximise the perceived diffuseness. An example production tool is suggested that–based on the metric–optimally pans the frequency components of a multichannel signal to loudspeakers to maximise the perceived diffuseness for a central listener looking directly forwards.

1.4 Terminology used in the Thesis

It is worth commenting on the terminology used during this thesis and how it evolved over the course of writing the thesis. Throughout this thesis the terms perceived diffuseness and envelopment are very similar and can in some cases be considered equivalent. However, where possible, the term perceived diffuseness is used to determine if the listener thinks that the sound field has the same perceptual qualities they would expect from a theoretical diffuse sound field. Whereas envelopment could be considered a perceptual sensation, part of which is related to the physical diffuseness.

An contrived example where these might differ might be in a concert hall where lateral energy is often associated with good spatial impression. Lateral reflections are known to reduce the correlation between the ears which is good for apparent source width (ASW) as well as listener envelopment (LEV). However, objectively, and without the visual reference, the listener might recognise the sound is not coming from all directions equally so would not be perceptually diffuse. From the point of view of object based audio, reproducing sound from the sides of the listener is trivial in comparison to producing sound from all around the listener. Therefore the focus is on the perceived diffuseness and not on envelopment.

Despite these subtle distinctions, when explaining the concept of perceived diffuseness to listeners, the description is nearly identical to the standardised description of envelopment from Zacharov et al. (2016) and so in general the two terms can be seen as functionally equivalent.

The final potential source of confusion is that the term “envelopment" is used in experiment 6. In this case the material is not just static noise. Instead there are multiple distinguishable components of the scene and the concept of perceived diffuseness becomes ambiguous. Therefore the decision was made to use the higher level attribute envelopment to hopefully make the task simpler and more repeatable for listeners at the possible expense of the collected data being slightly biased away from that of the earlier experiments.

The result of this is that where noise stimuli are used the words “perceived diffuseness" are used. For musical stimuli the word “envelopment" is used. However, for noise-like stimuli especially, the two terms are equivalent and therefore research on more general spatial impression is also included in the literature review.

Section 2.5.1 explains the terms envelopment and diffuseness in more detail.

1.5 Thesis Overview

The existing literature is presented in the first chapter of this research. Firstly, the current state of object-based audio in relation to diffuse sound fields is summarised (section 2.2). Diffuse sound fields are then described in more detail including the mathematical model in section 2.3 and a number of methods for measuring diffuseness are described from a range of application fields in section 2.4. The focus of this research is to deal with diffuse sound fields in a perceptually motivated manner and so the existing research on the perception of sound fields is summarised in section 2.5. Following this, listening test methodologies are reviewed as the subjective testing forms the backbone of this research (section 2.6). Finally methods of decorrelation are described as these describe the only current method for handling spatially extended sources in object-based audio (section 2.7).

One of the main outcomes of this research is a greater understanding of the perception of diffuse sound fields. Chapters 3 to 6 cover the first four listening tests that systematically tested a selection of different factors of the sound field. Simultaneously, physical measurements of the sound field were taken to investigate the relation between the objective and the subjective diffuseness. In each experiment different parameters that affect the diffuseness of the sound field are tested. Experiment 1 (chapter 3) was designed to allow subjectively optimised diffuseness reproduction for various loudspeaker layouts by using different loudspeaker gains for head-height and non-head-height loudspeakers. This allows the different loudspeaker layouts to be thoroughly examined in Experiment 2 (chapter 4). In this experiment the diffuseness of the sound field was varied by changing the arrangement of the loudspeakers and the gains of the loudspeakers. Standardised loudspeaker layouts were investigated along with different parameters of the arrangement of loudspeakers such as the number of loudspeakers at head-height. The validity and importance of the results from the first experiment were also investigated. Experiment 3 (chapter 5) investigates the perceptual and objective effect on the perceived diffuseness of using non-uncorrelated signals. The Inter-Channel Correlation Coefficient (ICCC) was varied for different loudspeaker layouts. The subjective data revealed the perceived diffuseness to be highly dependent on the ICCC but the objective measures revealed the ICCC to have very different effects on the sound field at both different frequencies and different listener positions. Therefore, experiment 4 (chapter 6) added the variable of frequency to further refine the effect of the ICCC.

Throughout these subjective experiments, objective measures of the sound field were taken. These allowed the subjective data to be compared to objective data. This highlighted where different metrics were successful and where there were errors or biases.

These comparisons between the objective and subjective data led to the development of a new metric for predicting diffuseness based on the coherence between points in the sound field in chapter 7. Whilst the coherence showed a good fit to the data collected in experiments 1 to 4, to ensure a more generalised solution for different stimuli required some refinement of the metric. Experiment 5 was conducted to subjectively tune the window length used for the FFT in the metric.

The metric was tested with some challenging material in chapter 9 to highlight the remaining limitations of the metric and describe the conditions in which it is successful (and more successful than other metrics) and where it is not.

Chapter 10 summarises how the results of this research relate to spatial audio in general and how these results can be used to improve spatial audio with particular focus on object-based spatial audio including descriptions of tools to maximise perceived diffuseness based on the results of the coherence based diffuseness metric as well as briefly covering the questions that remain following this research that may serve as a platform towards further research.

Chapter 2

Diffuseness in Reverberation Rooms, Concert Halls and Reproduced Audio: Implementation, Measurement and Perception

2.1 Overview

This chapter summarises the relevant literature. Section 2.2 covers the current usage of object-based audio, emphasising existing handling of diffuse sound sources. Section 2.3 introduces the concept of the ideal diffuse sound field. Section 2.4 covers different metrics designed to verify whether a sound field is diffuse or evaluate the diffuseness. Literature relating to Perception is covered in section 2.5. The subjective attributes associated with diffuseness perception are summarised in section 2.5.1 along with other research on the perception of diffuse sound fields in section 2.5.2. A variety of listening test methodologies are compared in section 2.6 and finally methods for increasing diffuseness by means of decorrelation are mentioned in section 2.7.

2.2 Existing Object-based Audio

There are some examples of commercially available object-based audio formats (MPEG4 and Atmos) and they handle non-point source objects differently.

In MPEG4 audio channels are transmitted along side BInary Format for Scenes (BIFS) metadata which is encoded as scene information contained in a node structure. BIFS is a compressed version of Virtual Reality Modelling Language (VRML)(Hosseini and Georganas, 2002) in which nodes describe attributes such as position, audio effects, or mixing of sources (Koenen, 2002).

Advanced Audio BIFS (AABIFS) add some additional functionality including the WideSound node (Schmidt and Schroeder, 2004). The WideSound node allows the shape and size of audio sources to be specified (figure 2.1) as well as parameters (density, decorrelation strength and diffuse select) that define the amount of decorrelation applied by the renderer and ensure different decorrelators are used when attached to the same source. This turns single point sources into many point sources. All the point sources are then spatialised using techniques such as Vector Based Amplitude Panning (VBAP) for loudspeaker reproduction (Plogsties et al., 2003) or HRTF based rendering for headphone listening (Jang et al., 2005).

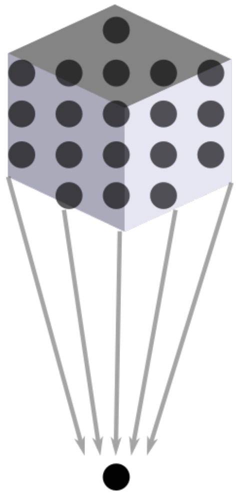


Figure 2.1: Cube built of uncorrelated sources specified using the WideSound node (adapted from Schmidt and Schroeder (2004)).

Atmos is Dolby's object-based format for cinema (Loftis, 2014). This format is designed to improve localisation and allow for irregular loudspeaker positioning. Atmos uses object-based techniques for rendering up to 118 localisable sources. These are layered over a 9.1 "bed" that encompasses the remaining point sources, and often diffuse and ambient sound fields in a conventional channel-based format. This is a simple solution to representing diffuse fields, however, considering the resources of the system with many discrete channels available it seems there is no real quality gain over existing channel-based formats–for the diffuse components at least. Also, in cinema, the high channel count is less of an issue but for consumer audio, the bandwidth required for 128 discrete channels may not be appropriate.

The main focus of these object-based systems is the improvements available in localisation, system independence or interactivity (Jang et al., 2005; Plogsties et al., 2003).

Existing systems either describe diffuse sound fields in terms of a set of independent discrete sources or apply decorrelation to a single source. Decorrelation algorithms have their own drawbacks (section 2.7) and so the current tools are unlikely to give the desired quality. However, the use of many independent sources is likely inefficient requiring a lot of data with little consideration of the perception of diffuse sound fields.

The target of this thesis is to discover the limitations of these existing object-based formats when reproducing diffuse sound fields to allow high quality, perceptually motivated reproduction.

2.3 The Ideal Diffuse Sound Field

This section will introduce the physical parameters that define an ideal, theoretical, diffuse sound field. This helps highlight when the sound field is “truly” diffuse and when it is an approximation of this ideal diffuse sound field. The concept is of sound arriving equally and simultaneously from all directions. A diffuse sound field can therefore be thought of as an infinite number of uncorrelated sources far from the measurement position and from all directions simultaneously. These uncorrelated sources do not add coherently so there is no interference pattern. Therefore the sound field becomes homogeneous (same root mean square (r.m.s.) pressure at all positions) and isotropic (arriving equally from all directions). The temporal correlation function and coherence function between two points should also depend only on their separation (Cook et al., 1955; Jacobsen and Roisin, 2000).

An ideal diffuse sound field has many practical applications including measurement of absorption, noise sound pressure, transmission loss in partitions and the standard diffuse field response/directivity index of a microphone (Cook et al., 1955; Veit and Sander, 1985; Jacobsen and Roisin, 2000).

The sound intensity is one method of looking at the diffuseness. The intensity of a point in the sound field is a vector that describes the power per unit area or the flow of energy through an area (figure 2.2).

The instantaneous sound intensity (I_r) is the amount of energy passing through an area in direction r over a period of time.

$$I_r = \frac{dE_r}{dt.dA} \tag{2.1}$$

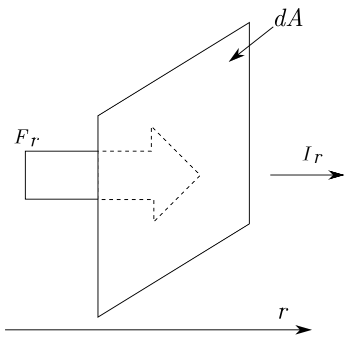


Figure 2.2: The energy transfer through dA due to the force F in the direction r. Adapted from Gade (1982)

The energy E_r is equal to the work done on the area dA by force F_r in the r direction.

$$dE_r = F_r.dr = p_t.dA.dr \tag{2.2}$$

where p_t is the sum of atmospheric p_a and sound pressure p.

Hence the instantaneous intensity is,

$$I_r = p_t.\frac{dr}{dt} = p_a u_r + p u_r \tag{2.3}$$

where u_r is the particle velocity in direction r given by dr/dt and p_t is the sum of atmospheric pressure p_a and sound pressure p. p_a is a d.c component that will, when time averaged, average to zero. The time averaged intensity vector in direction r is therefore the time averaged product of the pressure and particle velocity in direction r.

$$\langle I_r \rangle = \langle p.u_r \rangle \tag{2.4}$$

where $\langle\rangle$ indicates time averaging. This assumes the particle velocity has no d.c. offset (i.e. no flow).

In three dimensions this becomes,

$$\begin{aligned} \overrightarrow{\langle I \rangle} &= \langle p.u_x + p.u_y + p.u_z \rangle & (2.5) \\ &= \langle p.\vec{u} \rangle & (2.6) \end{aligned}$$

where u_x, u_y and u_z are orthogonal components of the particle velocity and $\overrightarrow{I}$ is the intensity vector (Gade, 1982).

The sound intensity includes both active (real) and reactive (imaginary) parts. Figure 2.3 shows the relationship between the pressure and particle velocity for both the active and reactive parts of the intensity .

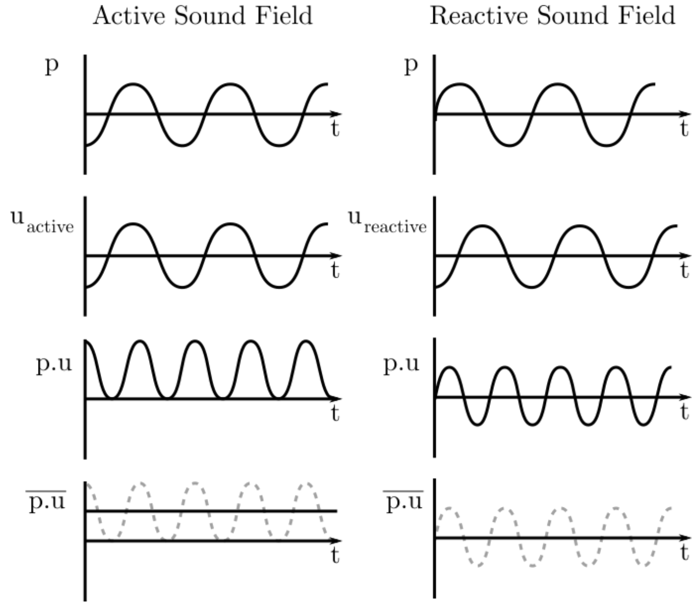


Figure 2.3: Pressure, particle velocity and intensities of active and reactive sound fields. Adapted from Gade (1982).

If the pressure and particle velocity are correlated and in phase the intensity is purely active, for example, a plane wave. All the energy is propagated and there is a phase gradient but no amplitude gradient. This would be homogeneous but not isotropic.

If the pressure and particle velocity are correlated but 90° out of phase, then the intensity is purely reactive. A standing wave is purely reactive because all the energy fluctuates between the sources (real or imaginary) and the medium. No energy is propagated and there is an amplitude gradient but no phase gradient. In this case there will be a instantaneous sound intensity but the time averaged intensity will average to zero. This is not homogeneous or isotropic.

In a diffuse sound field the pressure and particle velocity are uncorrelated at any point in space. There is no amplitude or phase gradient with the r.m.s. sound pressure the same at all points in space (homogeneous). The phase gradient is random and so the time averaged intensity is also zero (isotropic) (Gade, 1982).

The exception to these rules is at pressure maxima and minima of standing waves where there is either no particle velocity or no acoustic pressure. In these cases the the phase gradient or amplitude gradient have no meaning.

This theoretical isotropic and homogeneous sound field generated by infinitely many uncorrelated plane waves from all directions is the upper limit of the type of sound field we are interested in for this research. This mathematical model is the basis for some of the metrics described in the next section.

2.4 Measuring Diffuseness

The metrics described in this section cover the ways in which diffuseness can be measured. Measurements made in reverberation rooms require a diffuse sound field to ensure accurate measurements. In concert hall acoustics, diffuseness has been shown to be a perceptually desirable quality. The metrics described in this section are commonly used to validate reverberation chambers or compare the quality of concert halls by comparing the properties of a measured sound field to those of a theoretical diffuse sound field or an anechoic environment. In this research these metrics are used/adapted to objectively evaluate the diffuseness of a given sound field.

The techniques for measuring diffuseness are summarised well by Loutridis (Loutridis, 2009). They can be roughly separated into directional microphone techniques; angular dependence of the reverberant field (sound intensity measures); frequency irregularity; cross-correlation methods; uniformity and linearity of decay rate; and finally subjective methods. Several of these measures are focused on the diffuseness of reverberation although a diffuse sound field need not be reverberation. The measures that are transferable to non-reverberant diffuse sound fields–as is the case in spatial audio–are covered in this section.

It worth emphasising that the diffuse sound field is only defined quantitatively in its extreme states, not diffuse or perfectly diffuse (Spring and Randall, 1969). Therefore, there is no single “diffuseness” scale. A perfectly diffuse sound field must be both homogeneous and isotropic but a partially diffuse sound field is not defined. The measures covered in this section are designed to test the homogeneity and isotropy of a sound field. However, many of the methods test only one of these and are therefore inherently limited in their validity.

2.4.1 Directional Microphone Measures

Directional microphones can be used to test the isotropy of the sound field. By rotating the microphone, the relative levels from each direction can be compared to show diffuseness. Meyer uses a microphone in a parabolic reflector to show the energy in 10° intervals (Meyer, 1954). These are plotted at different points in the room onto “hedgehog” diagrams. These give a visual representation of the directivity of the sound field (figure 2.4)(Meyer, 1954).

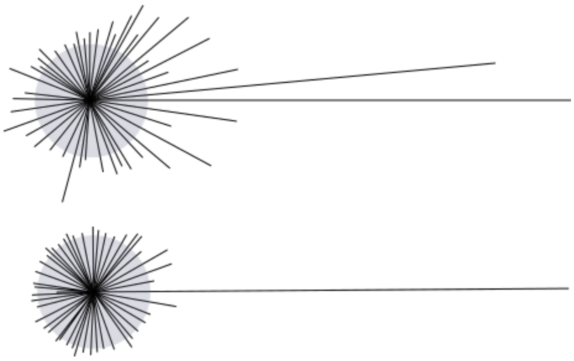


Figure 2.4: Hedgehog plots of directional diffusivity adapted from Meyer (1954). The top plot has good diffuseness, bottom has low diffuseness.

Thiele then took these measurements and was able to calculate the "directional diffusivity" which varies from 1 to 0 as the sound field changes from diffuse field to free field. If $E(\theta,\phi)$ is the energy measured by the directional microphone pointing in a particular direction, the average energy is given by M,

$$M = \frac{\int\limits_{\theta=\theta_0}^{\frac{\pi}{2}} \int\limits_{\phi=0}^{2\pi} E(\theta,\phi)\cos(\theta)\,\mathrm{d}\theta\mathrm{d}\phi}{2\pi \int\limits_{\theta=\theta_0}^{\frac{\pi}{2}} \cos(\theta)\,\mathrm{d}\theta} \tag{2.7}$$

where θ_0 is the lowest used elevation angle. The variation in the energy is therefore given by ΔM,

$$\Delta M = \frac{\int\limits_{\theta=\theta_0}^{\frac{\pi}{2}} \int\limits_{\phi=0}^{2\pi} |E(\theta,\phi) - M| \times \cos(\theta)\,\mathrm{d}\theta\mathrm{d}\phi}{2\pi \int\limits_{\theta=\theta_0}^{\frac{\pi}{2}} \cos(\theta)\,\mathrm{d}\theta} \tag{2.8}$$

the ratio of these values m, is given by,

$$m = \frac{\Delta M}{M} \tag{2.9}$$

And so, the directional diffusivity (d) is the m value normalised by the m value from an anechoic environment (m_o) (and subtracted from 1 so the metric increases with diffuseness)(Meyer, 1954).

$$d = 1 - \frac{m}{m_0} \tag{2.10}$$

An alternate, similar metric is the diffusion index. This is a comparison between the polar response of a directional microphone rotated in the test sound field and the polar response of the same microphone rotated in an anechoic environment. A polar response is generated by rotating a highly directional microphone in a stationary noise field ($R(\theta)$). The polar response measurement is then repeated in an anechoic room $D(\theta)$ (figure 2.5).

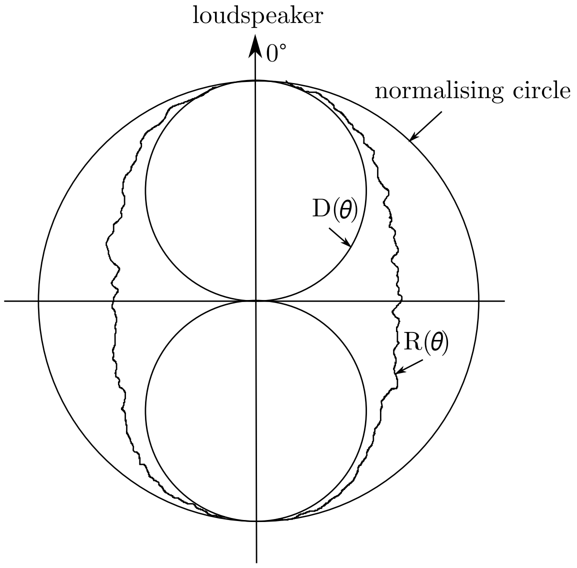


Figure 2.5: Directivity of the sound field and the microphone (adapted from Spring and Randall (1969)).

The diffusion index d is given by,

$$d = (A_0 - A)/A_0 \tag{2.11}$$

where A is the area of the polar response of the measured room and A_0 is the area of the polar response in the anechoic room (Spring and Randall, 1969).

Gover et al. use the same metric but use a a spherical microphone and beam-forming techniques to steer a directional microphone to all directions (Gover et al., 2004).

The use of a microphone with a high directivity index is optimal as the directional microphone measurements are limited by the directivity of the directional microphone (Gover et al., 2002).

All these metrics have been employed to test the diffuseness of rooms but directional index measures are unusable in small room as the contribution of the source always overpowers the reverberant measurement. This causes the result to be based on the amount of absorption and the level of the reverberation more than the diffuseness (Spring and Randall, 1969).

These directional microphone techniques only test the diffuseness at a single point of the sound field and not the sound field as a whole.

Another issue is that these metrics attempt to plot the distribution of the arriving energy with direction. This is of use in reverberation where the number of reflections is high and the directions form a continuous distribution. However, in reproduced audio, this may be less useful as the distribution of sound energy is already known, is discrete, and is dependent only the positions of the loudspeakers.

2.4.2 Sound Intensity Measures

Intensity measurements can be seen as calculating the proportion of the magnitude of the net flow of energy to the total energy. This will be 1 for free field (with all energy from one direction) and 0 for a diffuse field (flow of energy is from all directions and so averages zero).

The intensity is given by the product of the pressure and particle velocity (section 2.3). The particle velocity can be measured directly using an intensity metre (usually a flown wire) or using the pressure gradient, which can be estimated from two spatially separated pressure measurements (assuming the wavelength is much longer than the space between the microphones) (Gade, 1982).

Using Newton's second law $F = ma$, the pressure gradient can be linked to the particle acceleration $d\overrightarrow{u}/dt$ and the density of air ρ_0.

$$\rho_0 \frac{d\overrightarrow{u}}{dt} = -\nabla p \tag{2.12}$$

in direction r this becomes,

$$\rho_0 \frac{du_r}{dt} = -\frac{dp}{dr} \tag{2.13}$$

Therefore the particle velocity u in direction r is given by,

$$u_r = \frac{1}{\rho_0} \int \frac{dp}{dr} dt \tag{2.14}$$

This can be approximated using two closely spaced pressure signals separated by Δr.

$$\widetilde{u_r} = -\frac{1}{\rho_0 \Delta r} \int (P_B - P_A) dt \quad \Delta r \ll \lambda \tag{2.15}$$

where $\widetilde{u_r}$ is the estimate of the component of the particle velocity in direction r (Gade, 1982).

When applied orthogonally, a 3D intensity vector can be found and this is the basis of Directional Audio Coding (DirAC) diffusion estimation (Merimaa and Pulkki, 2005). Dirac processes a B-format signal to separate directional and diffuse parameters in separate frequency bands and time windows. This is then transmitted as metadata along side the mono pressure signal to allow the spatial scene to be recreated at the receiver. The diffuseness estimation compares an intensity vector to the total energy.

The components of the instantaneous intensity vector are I_X, I_Y and I_Z and can be calculated using,

$$I_X(t,f) = \frac{1}{\sqrt{2}Z_0}\mathbb{R}\{W^*(t,f)\cdot X(t,f)\}, \tag{2.16}$$

where $W(t,f)$, and $X(t,f)$ are the Short Time Fourier Transform (STFT) spectra of the zero-order pressure component and first-order component in the x direction respectively of the B-format signal and * denotes the complex conjugate. The expected energy density given by,

$$E(t,f) = \frac{1}{2}\rho_0 Z_0^{-2}\left(|W(t,f)|^2 + \frac{|X(t,f)|^2+|Y(t,f)|^2+|Z(t,f)|^2}{2}\right), \tag{2.17}$$

where ρ_0 is the mean density of air and Z_0 is the acoustic impedance of air. The diffuseness estimation is then given by,

$$\psi(t,f) = 1 - \frac{||E\{[I_X(t,f),\ I_Y(t,f),\ I_Z(t,f)]^T\}||}{cE\{E(t,f)\}}, \tag{2.18}$$

where $E\{.\}$ is the expectation operator and c is the speed of sound. The diffuseness coefficient ranges from 0 to 1 for each frequency bin and time window.

The problems with purely sound intensity measures are that they are sensitive to room modes (which also have an time averaged sound intensity of zero) and a single measurement only tests the isotropy at a single position in the room.

2.4.3 Spatial Correlation/Coherence Measures

The isotropic and homogeneous properties of a diffuse sound field mean the cross-correlation function and the coherence function between two points in a diffuse sound field can be found analytically and are dependent only on the separation between the points (Jacobsen and Roisin, 2000).

In reverberation rooms, the diffuse reverberation is not generated by an infinite array of independent, uncorrelated sources. Instead the room is driven by a finite number of sources (usually 1), but the irregular shape of the room and long reverberation time means that the reflections have random phase and therefore sum to become a diffuse sound field. This leads to a model for a diffuse sound field composed of plane waves with random phase and from all directions. In this case, the correlation function (Cook et al., 1955; Nélisse and Nicolas, 1997) and coherence function (Jacobsen and Roisin, 2000) between two points can be found analytically.

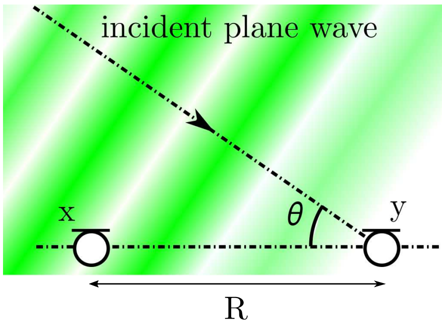


Figure 2.6: Plane wave.

Considering a sinusoidal plane wave as shown in figure 2.6, the pressures p_x and p_y are,

$$\begin{aligned} p_x(t) &= A\cos(\omega t) &= \mathbb{R}\{Ae^{j\omega t}\} \\ p_y(t) &= A\cos(\omega(t+\tau)) &= \mathbb{R}\{Ae^{j\omega(t+\tau)}\} \end{aligned} \tag{2.19}$$

where ω is the angular frequency and τ is the delay between the microphone measurement positions given by,

$$\tau = R\cos(\theta)/c \tag{2.20}$$

where c is the speed of sound, θ is the angle of the incoming plane wave and R is the separation between the measurement points.

The cross-correlation coefficient is, by definition, given by,

$$\rho_{xy} = \frac{\langle P_x(t)P_y(t)\rangle}{\sqrt{\langle P_x(t)^2\rangle\langle P_y(t)^2\rangle}} \tag{2.21}$$

Inputting the signals from equation 2.19 gives a cross-correlation coefficient given by,

$$\begin{aligned} \rho_{xy} &= \mathbb{R}\{e^{-j\omega R\cos\theta/c}\} \qquad (2.22) \\ \rho_{xy}(kR) &= \cos(kR\cos\theta) \end{aligned}$$

Where k is the wave number. The cross-correlation coefficient as a function of kR shows the cross-correlation coefficient to be dependent on frequency, the incident angle of the plane wave and the separation between measurement points.

To find the cross-correlation coefficient between points in a diffuse sound field composed of plane waves of random phase, the plane wave case is integrated for all possible directions (θ and ϕ) leading to,

$$\rho_{xy}(kR) = \cos(kR) \qquad (2.23)$$

for a 1 dimensional case,

$$\rho_{xy}(kR) = \frac{1}{2\pi}\int_0^{2\pi}\cos(kR\cos\theta)\mathrm{d}\theta = J_0(kR) \qquad (2.24)$$

for the 2 dimensional case and,

$$\rho_{xy}(kR) = \frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}\cos(kR\cos\theta)\sin\theta\mathrm{d}\theta\mathrm{d}\phi = \frac{\sin(kR)}{kR} \qquad (2.25)$$

for the 3 dimensional case (Nélisse and Nicolas, 1997).

Jacobsen and Roisin (2000) use coherence in place of correlation. The coherence γ_{xy} between the pressure at 2 points is given by,

$$\gamma^2_{xy}(\omega) = \frac{|S_{xy}(\omega)|^2}{S_{xx}(\omega)S_{yy}(\omega)} \qquad (2.26)$$

The coherence requires the cross-spectrum and autospectra of the two signals. The relationship between the cross-spectral density function S_{xy} and the cross-correlation function R_{xy} is the Fourier transform.

$$S_{xy}(f) = \int_{-\infty}^{\infty} R_{xy}(\tau)e^{j\omega\tau}\mathrm{d}\tau \qquad (2.27)$$

Therefore single sided cross-spectrum G_{xy} is given by,

$$G_{xy}(f) = 2\int_{-\infty}^{\infty} R_{xy}(\tau)e^{j\omega\tau}\mathrm{d}\tau = C_{xy}(f) - jQ_{xy}(f) \qquad (2.28)$$

where C_{xy} and Q_{xy} are the cospectrum and quadspectrum respectively, these are the real and imaginary parts of the cross-spectrum.

The cross-correlation function is therefore linked to the real and imaginary parts of the cross spectrum by,

$$R_{xy}(\tau) = \int_0^\infty [C_{xy}(f)\cos(\omega\tau) + Q_{xy}(f)\sin(\omega\tau)]\mathrm{d}f \tag{2.29}$$

In this equation the cross-spectrum is integrated over all frequencies. This can be replaced by a limited integral over a limited frequency range and the cross-correlation function for a given frequency band.

$$R_{xy}(\tau, \omega_0, \Delta\omega) = \int_{\omega_a}^{\omega_b} [C_{xy}(f)\cos(\omega\tau) + Q_{xy}(f)\sin(\omega\tau)]df \tag{2.30}$$

and the zeroth time lag this gives,

$$R_{xy}(0, \omega_0, \Delta\omega) = \int_{\omega_a}^{\omega_b} C_{xy}(f)\mathrm{d}f \tag{2.31}$$

For the coherence (equation 2.26) the cross-spectrum S_{xy} includes both the real and imaginary parts,

$$|S_{xy}(\omega)|^2 = C_{xy}^2(\omega) + Q_{xy}^2(\omega) \tag{2.32}$$

The imaginary part of the cross-spectrum can be related to the Hilbert transform of the cross-correlation function by,

$$\hat{R}_{xy}(\tau, \omega_0, \Delta\omega) = \int_{\omega_a}^{\omega_b} [C_{xy}(f)\sin(\omega\tau) + Q_{xy}(f)\cos(\omega\tau)]\mathrm{d}f \tag{2.33}$$

where ˆ denotes the Hilbert transform. This gives the zeroth lag Hilbert transform of the cross-correlation in terms of the quadspectrum,

$$\hat{R}_{xy}(0, \omega_0, \Delta\omega) = \int_{\omega_a}^{\omega_b} Q_{xy}(f)\mathrm{d}f \tag{2.34}$$

For the autocorrelation the autospectrum is purely real and even and is therefore given by,

$$R_{xx}(\tau, \omega_0, \Delta\omega) = \int_{\omega_a}^{\omega_b} C_{xx}(f)\cos(\omega\tau)\mathrm{d}f \tag{2.35}$$

The equation for the cross-correlation coefficient function is then used to find the coherence function. The cross-correlation coefficient function is by definition given by,

$$\rho_{xy}(\tau) = \frac{R_{xy}(\tau)}{\sqrt{R_{xx}(0)R_{yy}(0)}} \tag{2.36}$$

This can be found in narrow frequency bands using,

$$\rho_{xy}(\tau, \omega_0, \Delta\omega) = \frac{R_{xy}(\tau, \omega_0, \Delta\omega)}{\sqrt{R_{xx}(0, \omega_0, \Delta\omega)R_{yy}(0, \omega_0, \Delta\omega)}} \tag{2.37}$$

The Hilbert transform of the cross-correlation coefficient function will also be required as this uses the Hilbert transform of the cross-correlation function which includes the imaginary part of the cross spectrum (the quadspectrum).

$$\hat{\rho}_{xy}(\tau, \omega_0, \Delta\omega) = \frac{\hat{R}_{xy}(\tau, \omega_0, \Delta\omega)}{\sqrt{R_{xx}(0, \omega_0, \Delta\omega)R_{yy}(0, \omega_0, \Delta\omega)}} \tag{2.38}$$

Again looking at the zeroth time lag the equations 2.31, 2.34 and 2.38 can be combined in to give the cross-correlation coefficient as a function of frequency using the narrow band cospectrum and autospectra.

$$\rho_{xy}(0, \omega_0, \Delta\omega) = \frac{\int_{\omega_a}^{\omega_b} C_{xy}(\omega)\mathrm{d}\omega}{\sqrt{\int_{\omega_a}^{\omega_b} S_{xx}(\omega)\mathrm{d}\omega \int_{\omega_a}^{\omega_b} S_{yy}(\omega)\mathrm{d}\omega}} \tag{2.39}$$

The quadspectrum in combination with the autospectra gives,

$$\hat{\rho}_{xy}(0, \omega_0, \Delta\omega) = \frac{\int_{\omega_a}^{\omega_b} Q_{xy}(\omega)\mathrm{d}\omega}{\sqrt{\int_{\omega_a}^{\omega_b} S_{xx}(\omega)\mathrm{d}\omega \int_{\omega_a}^{\omega_b} S_{yy}(\omega)\mathrm{d}\omega}} \tag{2.40}$$

Finally, combining equations 2.39, 2.40 with equations 2.26 and 2.32 shows the squared coherence given by the narrow band correlation coefficient and its Hilbert transform.

$$\gamma_{xy}^2(\omega_0, \Delta\omega) = \rho_{xy}^2(0, \omega_0, \Delta\omega) + \hat{\rho}_{xy}^2(0, \omega_0, \Delta\omega) \tag{2.41}$$

This is equivalent to the envelope of the cross-correlation coefficient.

In a diffuse sound field where the cross-correlation coefficient is given by equation 2.25, this leads to the coherence as given by,

$$\gamma_{xy}^2(\omega, r) = \left(\frac{\sin(\omega r/c)}{(\omega r/c)}\right)^2 \tag{2.42}$$

assuming that the spectral resolution is very high ($\Delta\omega \ll \omega$). In this way the coherence can be seen as calculating many correlation measurements at once (Jacobsen and Roisin, 2000). They point out that the use of the coherence is only useful if the sound field is composed of uncorrelated sources. However, because the autocorrelation of random noise is mostly 0 except for at 0 lag, for rooms with sufficiently long reverberation time, the reflections appear uncorrelated. This means that in a reverberation room driven with a single loudspeaker, using a frequency resolution that is too high ($\Delta\omega < 1/T_{60}$) or using pseudorandom noise synchronised to the length of the FFT, will lead to a coherence of 1 for all frequencies. This is expected as the loudspeaker, reverberation, microphone process is a linear time invariant system. The paper by Jacobsen and Roisin demonstrates the validity of equation 2.42 but also demonstrates that if spatial averaging is used, even non-diffuse rooms, can lead to the coherence function given by equation 2.42 (Jacobsen and Roisin, 2000). The measurement of the coherence is not incorporated into a metric for measuring diffuseness (possibly due to these biases) but in this research, where the sound field can be composed of truly uncorrelated discrete sources, the coherence is once again a valid parameter.

2.4.4 Spatial Uniformity Measures

Spatial uniformity techniques compare the pressure at multiple positions in the sound field to test the homogeneity. In the case of reverberation rooms, ISO standard 3741 (ISO, 1999) gives the permissible standard deviation of the Sound Pressure Level (SPL) between points for different frequency ranges for a given upper limit of reproducibility accuracy when making reverberation room measurements (table 2.1).

Octave band Mid-frequencies/Hz	1/3 octave band mid frequencies/Hz	Standard Deviation of SPL/dB
125	100 to 160	1.5
250 and 500	200 to 630	1.0
1000 and 2000	800 to 2,500	0.5
4000 and 8000	3,150 to 10,000	1.0

Table 2.1: Maximum permissible standard deviation of sound pressure level as specified in ISO standard 3741. (ISO, 1999)

The standard deviation is calculated over at least 6 measurement positions, each more than half the wavelength of the lowest frequency of interest apart; no closer to the surfaces of the room than 1m; and at least d_{min} for the source, where d_{min} is given by,

$$d_{min} = 0.4 \times 10^{(L_{Wr} - L_{pr})/20} \tag{2.43}$$

where L_{Wr} and L_{pr} are the sound power level of the source (in dB) and the sound pressure level of the source in the room (in dB) respectively.

An alternative spatial uniformity measure is to quantify a spatially limited diffuse sound field by specifying the area over which the sound pressure level is constant (Veit and Sander, 1985). Viet and Sander simulated a diffuse sound field using 8 loudspeakers. The SPL was then calculated as a microphone moved from the centre of the sound field towards the nearest loudspeaker. This showed the sound field to be constant up to a point and then the SPL becomes non-uniform. This is shown in curve *b* from figure 2.7, with curve *a* showing the equivalent measure with a single loudspeaker in an anechoic environment.

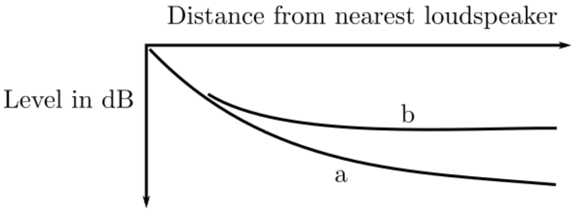


Figure 2.7: SPL of reproduced diffuse sound field as the measurement point moves towards one of the loudspeakers for 1/3 octave band centred at 315 Hz (adapted from Veit and Sander (1985)),

These spatial uniformity measures are applicable to the entire listening area, however, they only measure the pressure so ignore the direction of arrival. For example a plane wave has the same pressure at all positions but is not diffuse.

2.4.5 Angular (Wavenumber) Spectrum Measure

One method of investigating isotropy is using spherical harmonics. The first order of the spherical harmonics is synonymous with the isotropy metrics in section 2.4.2. Whilst the first order has limited spatial resolution (only estimating the average direction of incoming sound), higher order spherical harmonics have higher spatial resolution allowing for better separation of incoming plane waves. Nolan et al. (2016) decompose spherical microphone array simulations into a set of a plane waves. Either the phase or magnitude of each plane wave is then translated to the spherical harmonic domain. The energy of the coefficients in each order are averaged and normalised to show how the magnitude or the phase of the spherical harmonics is distributed across the orders of harmonics.

Firstly the sound field is described as a complex angular spectrum. In the paper by Nolan et al. this complex angular spectrum is theoretical and therefore can be trivially discretised and simulated for 1000 points. The magnitude of the angular spectrum $|P(k,\Omega)|$ is represented as a sum of spherical harmonics,

$$|P(k,\Omega)| = \sum_{n=0}^{\infty} \sum_{m=-n}^{m=n} A_{mn}(k) Y_n^m(\Omega). \tag{2.44}$$

The energy of the spherical harmonic coefficients A_{mn} can then be normalised using,

$$\frac{1}{N} \sum_{m=-n}^{n} |A_{mn}|^2, \tag{2.45}$$

Where N is the number of coefficients contributing to each order.

The phase of the angular spectrum $(\phi(P(k,\Omega)))$ can also be calculated as a sum of spherical harmonics,

$$\phi(P(k,\Omega)) = \sum_{n=0}^{\infty} \sum_{m=-n}^{m=n} B_{mn}(k) Y_n^m(\Omega). \tag{2.46}$$

The energy of the spherical harmonic coefficients that relate to the phase (B_{mn}) can then be normalised using,

$$\frac{1}{N} \sum_{m=-n}^{n} |B_{mn}|^2, \tag{2.47}$$

where N is again the number of coefficients contributing to each order.

Figure 2.8 shows how the normalised energy varies between the harmonic orders in terms of magnitude and phase for three different plane wave situations. In an isotropic sound field (middle and bottom) the sound comes equally from all directions. In the spherical harmonic domain this corresponds to only the zeroth order, omnidirectional component of the magnitude of the plane waves. Sound fields that are less isotropic (Single plane wave) require higher orders to describe them. Additionally, if the incoming plane waves are correlated, then the sound field will still not be diffuse. The phase of the incoming plane waves is therefore also converted into the spherical harmonic domain allowing the correlated and uncorrelated cases to be differentiated.

This method of visualising the sound field is most suited to the suggested use case, reverberation. In this case the plane wave assumption is most appropriate. Reflections are harmonically related to the source and the randomisation is the phase of the reflections. This phase relation between a reflection and the direct sound is always constant. In reproduced audio where the sources are truly uncorrelated, the phase shift between the “plane” waves from different directions varies with time. The randomisation of the relative

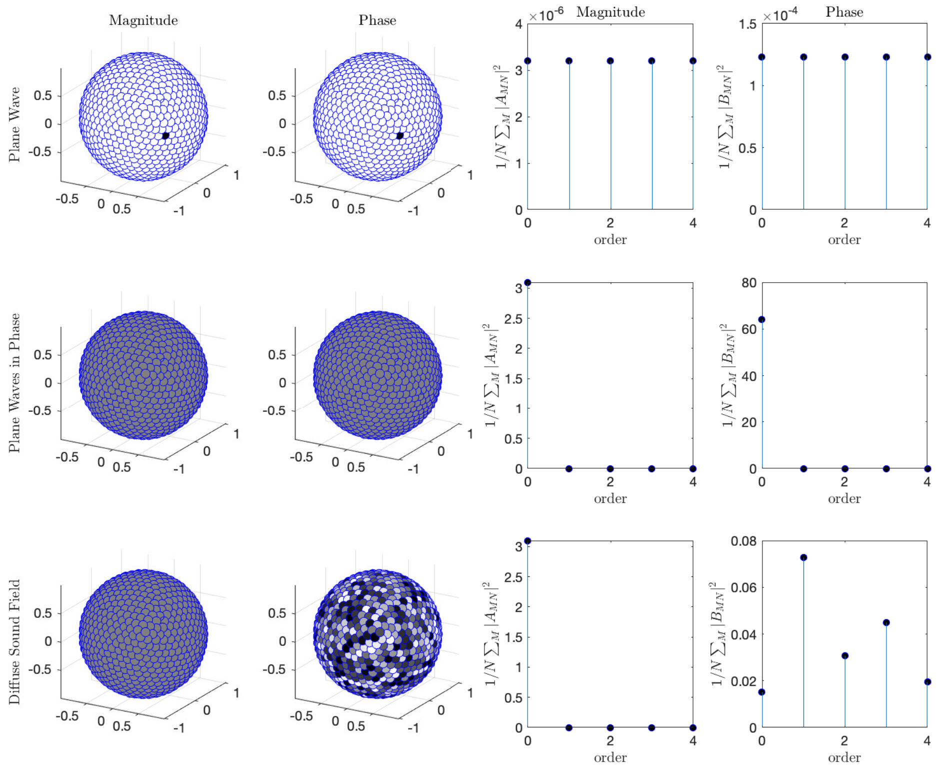


Figure 2.8: Phase and magnitude of the normalised energy for different orders of spherical harmonics for different sound fields. *Top* a single plane wave. *Middle* Many plane waves all in phase. *Bottom* Many plane waves of random phase. (adapted from Nolan et al. (2016))

phase with time means that the phase will likely average to zero over time. Therefore it would be necessary to apply a time window to avoid the zero average phase difference being reflected in the reduction of the higher orders of harmonics. It is not immediately obvious what this time window should be and how to combine separate time windows to give a single value.

Another remaining issue with this method is there is no obvious way to combine the phase part with the amplitude part. As is often a difficulty with measuring diffuseness (a combination of both homogeneity and isotropy) this measurement technique investigates both but does not give any suggestion as to an optimal method for combining the two parts.

The author also highlights that more research needed for low frequency signals where the modal response of the room will encourage the lower order harmonics to dominate. Whilst this effect at low frequencies might actually be highly relevant as highly modal sound fields may be less diffuse, it is not clear exactly how this might affect the measurement and further research would be required.

2.4.6 Perceptually Motivated Measures for Concert Hall Acoustics

The metrics so far compare the properties of a measured sound field to the properties of a theoretical diffuse sound field. In concert hall acoustics the focus is slightly different. The source is generally small and reflections from the walls either broaden the source or are perceived separately as part of the environment. These desirable spatial attributes are discussed in section 2.5.1 where the relevant factor is how these reflections increase the spatial impression over anechoic conditions. This difference is subtle but there are cases where maximum diffuseness according to these metrics may not equate to the value of these metrics in a theoretical diffuse sound field.

Lateral Fraction and Lateral Hall Gain The lateral fraction of Barron and Marshall is a simple measure that compares the impulse response of an omnidirectional microphone to the impulse response of a figure-of-8 microphone with the "dead axis" pointed towards the source (i.e. the lateral energy)(Barron and Marshall, 1981). A source in an anechoic room would not be recorded on the figure-of-8-microphone leading to a low lateral fraction. Conversely, in a highly diffuse environment, the figure-of-8-microphone will pick up a lot of the reflections from the side walls giving a high lateral fraction.

$$LateralFraction = \frac{\int\limits_{0}^{80ms} x_{fig8}(t)^2 \mathrm{d}t}{\int\limits_{0}^{80ms} x_{omni}(t)^2 \mathrm{d}t} \tag{2.48}$$

The lateral fraction is measured over the first 80ms of reverberation for early reflections. The microphones are calibrated to give the same output for a source on-axis for both microphones in anechoic conditions.

The Lateral Hall Gain of Bradley and Soulodre (1995) is used for late reverberation and is given by,

$$LateralHallGain = \frac{\int\limits_{80ms}^{\infty} x_{fig8}(t)^2 \mathrm{d}t}{\int\limits_{0}^{\infty} x_{omni\_anc}(t)^2 \mathrm{d}t} \tag{2.49}$$

In this case the late reverberation is measured at a standard distance of 10 m. The omnidirectional measurement is calculated in anechoic conditions this time.

These metrics were designed based on the perception of early and late reflections being relevant factors of spatial impression (Barron and Marshall, 1981). They have found uses in concert hall acoustics although the lateral fraction varies only a small amount between different halls (Hidaka et al., 1995; Barron, 2001) and the lateral hall gain is found to be strongly dependent on the amount of absorption which may or may not relate to the particular spatial impression (Barron, 2001).

This metric is also not well suited to reproduced audio as the maximum value can be obtained by increasing the level of the loudspeakers at the sides. This will not increase the physical diffuseness and is unlikely to increase the perceived diffuseness but will increase the measured lateral energy.

InterAural Cross-correlation Coefficient (IACC) The InterAural Cross-correlation Coefficient (IACC) is very commonly used. In general these methods are the same and the IACC is given by the maximum absolute value of the normalised interaural cross-correlation function (IACCF) over the range of possible interaural time differences (ITDs) (ISO, 2009).

$$IACCF_{t1,t2}(\tau) = \frac{\int\limits_{t1}^{t2} p_L(t) \cdot p_R(t+\tau)\mathrm{d}t}{\sqrt{\int\limits_{t1}^{t2} p_L^2(t)\mathrm{d}t \int\limits_{t1}^{t2} p_R^2(t)\mathrm{d}t}} \tag{2.50}$$

where $p_L(t)$ and $p_R(t)$ are the left and right ear impulse responses respectively and t_1 and t_2 determine a section of the impulse response.

The IACC is given by the maximum absolute value over the range of possible ITDs.

$$IACC_{t1,t2} = max\left|IACCF_{t1,t2}\right|, \quad -1\mathrm{ms} \leq \tau \leq 1\mathrm{ms} \tag{2.51}$$

The t_1 and t_2 values are commonly 0 ms and 80 ms respectively for early reflections and 80 ms and the length of the impulse response respectively for late reflections. The distinction is made as early reflections are fused to the source whereas late reflections are perceived as part of the environment. The IACC is also commonly calculated in octave bands and averaged. This is partly due to the logarithmic frequency response of the ear and partly due to some frequency bands matching poorly to subjective data. Hidaka et al. use the octave bands at 500 Hz, 1 kHz and 2 kHz based on the assumption of the likely source material and the accuracy with which these bands differentiate different concert halls (Hidaka et al., 1995). At very high and low frequencies there is commonly less energy especially with the types of music performed acoustically in concert halls. The IACC is also always high at low frequencies where the long wavelength relative to the interaural distance means a higher degree of correlation. Therefore, there is a smaller range of possible values making the measurements less accurate and poor differentiators between different concert halls.

There are several notable points regarding the IACC;

- The IACC is high at low frequencies due to longer wavelengths. Griesinger points out the fact that the increase in IACC does not correlate well with subjective data which shows spatial impression to be important at low frequencies (Griesinger, 1999). However it is worth mentioning that this observation of Griesinger is slightly ambiguous. It is understandable that the IACC is a poor measure of envelopment at LF as it tends towards 1 where the envelopment may not tend towards “not enveloping”. However this is not the same as determining that a lower value of IACC would still not be more enveloping than a higher IACC, albeit there is a smaller range of possible IACC values at low frequency than at high frequency as was found by Gribben and Lee (2017).

- Measuring the IACC over a small range of octave bands may lead to ignoring frequencies that may still be relevant to the perceived diffuseness even if those frequencies might be less important.

- A low IACC can be measured from sound fields that are not physically diffuse. Ando and Kurihara ran experiments involving a single reflection and the subjective diffuseness. They found, for a direct sound in front of the listener, a large azimuth angle ($\pm 90°$) gave the highest subjective diffuseness whereas frequencies above 700 Hz require a narrower angle of the reflection. The angle of the single reflection that gave the highest subjective diffuseness also gave the lowest IACC (Griesinger, 1999; Ando and Kurihara, 1986). However, in terms of a theoretical diffuse sound field, there was no change, there was still a single source and single reflection. Of course this distinction may not be important as the subjective perception is of more importance to this research than the physical diffuseness of the sound field.

Despite these issues and considerations, the IACC has been found to correlate well with subjective attributes, Apparent Source Width (section 2.5.1) in (Hidaka et al., 1995) and with perceived diffuseness in (Power et al., 2014).

InterAural Difference (IAD) The InterAural Difference (IAD) is worth inclusion as the metric is very simple and incorporates equalisation to account for the high correlation at low frequencies due to longer wavelengths. The IAD expands on the lateral fraction and compares the level of the equalised difference between the binaural impulse responses of the two ears to the total level of the impulse responses at the two ear signals (Griesinger, 1999).

$$IAD(t) = 10 \log_{10} \left(\frac{(eq(L(t) - R(t)))^2}{(L(t)^2 + R(t))^2} \right) \tag{2.52}$$

At low frequencies the difference signal is low due to the higher correlation between the two ears. Therefore, the difference signal is equalised to give an IAD of 0 for all frequencies in a large reverberant room and this can be estimated using a 6 dB/8ve boost below 300 Hz. This gives the IAD as a function of time and can be calculated in octave bands with 40 ms smoothing.

2.4.7 Summary

These metrics are designed to quantify the differences between non diffuse sound fields such as plane waves, diffuse sound fields and sound fields that lie somewhere in-between. A useful metric in the context of spatial audio is one that correlates well with the perception of the sound field. The actual diffuseness of the sound field is of less importance than it being perceived as diffuse.

2.5 Perceiving Diffuseness

In this section commonly used terminology is explained and existing subjective experiments that relate to diffuse sound fields are covered.

2.5.1 Terminology and Subjective Attributes of Spatial Audio

In any subjective experiment the choice of language can be ambiguous or even biasing. This section covers the difficulties with the terminology when describing diffuse sound fields. Firstly the ambiguity of "diffuseness" is explained and then the research investigating semantic descriptors of sound fields is covered. This allows clarification of the terminology used throughout this research.

The term diffuseness can be ambiguous for two reasons. The first reason for ambiguity is that technically, a sound field can only be either diffuse, or not diffuse, so there cannot be a diffuseness scale. It is logical that a single plane wave might be the lowest possible diffuseness and an ideal diffuse sound field might have the highest diffuseness but there is no single scale in the middle to define moderate diffuseness. The second reason for ambiguity comes from the subjective aspect of this research. Physical diffuseness is important in reverberation rooms where the sound field must be highly diffuse in order to make accurate measurements. It must be completely isotropic and homogeneous so it is equivalent to the theory of an infinite number of uncorrelated sources coming from all directions simultaneously. This allows the measurements to be repeatable. Alternatively, a sound field can be perceptually diffuse, i.e. it sounds as if it is generated by an infinite number of uncorrelated sources far from the listener. This leads to the requirement to specify diffuseness as either physical diffuseness or perceptual diffuseness.

The ambiguity is complicated further as a physically diffuse sound field may not be perceived as an infinite number of uncorrelated sources from all directions. This is found in concert hall acoustics where the early reflections are fused to the source and make it sound wider and the diffuseness of these early reflections is proportional to its size. In this case a better option may be to use the descriptors used by listeners for the effects they hear when diffuseness is increased. However, when listeners are asked to give verbal feedback on what they hear, there tends to be a wide variety of ways of describing it. Fortunately, a lot of work has been made into identifying attributes that are synonymous or antonymous as to select semantic attributes that are independent, easy to understand and hopefully consistent.

In concert hall acoustics, the most common terms used when referring to diffuse reverberation are Apparent/Auditory Source Width (ASW) and Listener EnVelopment (LEV). These were found by (Bradley and Soulodre, 1995) to be independent components of spatial impression. The ASW is associated with early reflections that are indistinguishable from the source making it sound larger. LEV is dependent on the late reverberation and is responsible for the feeling of being surrounded by the room.

For reproduced sound, Berg et al. derived "broad attribute classes" from verbal data using a Repertory Grid Technique (Berg and Rumsey, 2003, 1999). These were,

- Authenticity/naturalness
- Lateral positioning/source size
- Envelopment
- Depth

Semantic descriptors from previous research were investigated including attributes from concert hall acoustics and from reproduced sound. Berg and Rumsey then used the method shown in figure 2.9 to select the consistent and independent attributes shown in table 2.2.

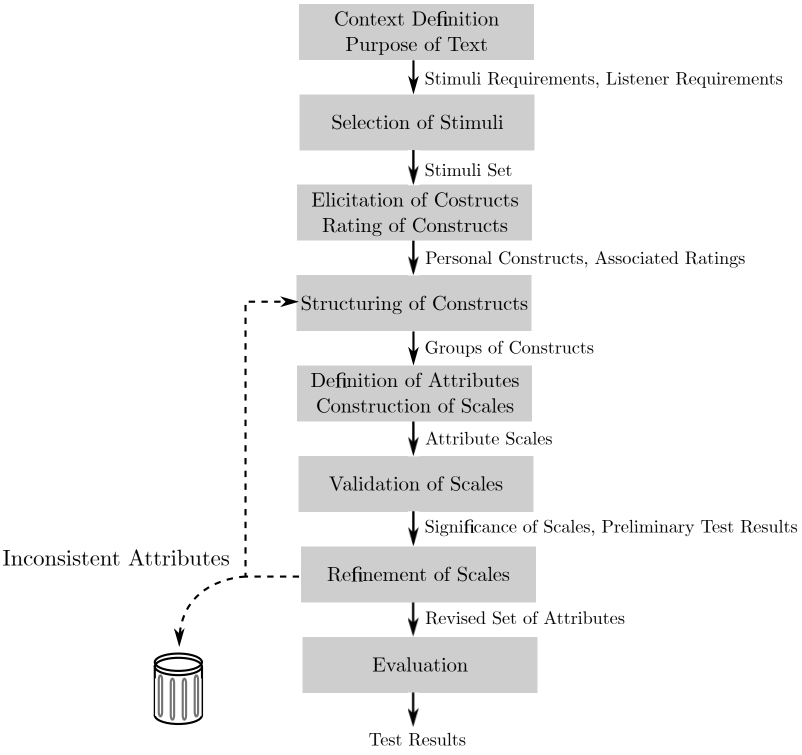


Figure 2.9: Method for evaluating semantic attributes. Adapted from Berg and Rumsey (2003)

We can see from the attributes shown in table 2.2, that there is some extension on the broader classes and on the attributes from concert hall acoustics. ASW is split between individual source width and ensemble width and envelopment is both source envelopment and room envelopment. Even with these well defined classes there are cases in the past of different uses of these terms and therefore there remains ambiguity in the literature. Surroundedness is tentatively proposed by Berg as a new semantic descriptor for the attributes describing sound that surrounds the listener emphasising when the definitions of envelopment from the past may have contradicted each other (Berg, 2009).

Even further distinction can be made when investigating surround sound systems with height where engulfment is proposed as an attribute that is independent of envelopment when referring to loudspeaker systems with height (Paine et al., 2007).

The most recent attempt to standardise the definition of diffuseness is given by Zacharov et al. (2016). They recommend the use of the following description.

> “Degree of being surrounded by a source, scene or ensemble. Typically, envelopment is associated with a scene. Scale: Not enveloping to Completely enveloping Being surrounded by reverberation would be considered highly enveloping. Being surrounded by a large number of dry sources may also be highly enveloping. This may be heard when standing and listening to the rain

Attribute	Description
Naturalness	How similar to a natural (i.e. not reproduced through e.g. loudspeakers) listening experience the sound as a whole sounds.
Presence	The experience of being in the same acoustical environment as the sound source, e.g. to be in the same room.
Preference	If the sound as a whole pleases you. If you think the sound as a whole sounds good. Try to disregard the content of the programme, i.e. do not assess genre of music or content of speech.
Low Frequency Content	The level of low frequencies (the bass register).
Ensemble Width	The perceived width/broadness of the ensemble, from its left flank to its right flank. The angle occupied by the ensemble. The meaning of "the ensemble" is all of the individual sound sources considered together. Does not necessarily indicate the known size of the source, e.g. one knows the size of a string quartet in reality, but the task to assess is how wide the sound from the string quartet is perceived. Disregard sounds coming from the sound source's environment, e.g. reverberation - only assess the width of the sound source.
Individual Source Width	The perceived width of an individual sound source (an instrument or a voice). The angle occupied by this source. Does not necessarily indicate the known size of such a source, e.g. one knows the size of a piano in reality, but the task is to assess how wide the sound from the piano is perceived. Disregard sounds coming from the sound source's environment, e.g. reverberation - only assess the width of the sound source.
Localisation	How easy it is to perceive a distinct location of the source - how easy it is to pinpoint the direction of the sound source. Its opposite is when the source's position is hard to determine - a blurred position.
Source Distance	The perceived distance from the listener to the sound source.
Source Envelopment	The extent to which the sound source envelops/surrounds/exists around you. The feeling of being surrounded by the sound source. If several sound sources occur in the sound excerpt: assess the sound source perceived to be the most enveloping. Disregard sounds coming from the sound source's environment, e.g. reverberation - only assess the sound source.
Room Width	The width/angle occupied by the sounds coming from the sound source's reflections in the room (the reverberation). Disregard the direct sound from the sound source.
Room Size	In cases where you perceive a room/hall, this denotes the relative size of that room.
Room Sound Level	The level of sounds generated in the room as a result of the sound source's action, e g reverberation - i.e. not extraneous disturbing sounds. Disregard the direct sound from the sound source.
Room Envelopment	The extent to which the sound coming from the sound source's reflections in the room (the reverberation) envelops/surrounds/exists around you - i.e. not the sound source itself. The feeling of being surrounded by the reflected sound.

Table 2.2: Attributes of spatial audio from Berg and Rumsey (2003).

> hitting the pavement. Envelopment may occur with reverberation or other aspects of the scene such as applause in a concert hall, atmosphere or air conditioning (room tone). Holes (an absence of sound from a certain directions) in the reproduction would normally reduce envelopment. Envelopment may be subdivided in horizontal and vertical envelopment"

This definition was published in 2016. All experiments carried out after this date used this description of envelopment as the basis for the description given to listeners.

Figure 2.10 is used to visualise different cases that demonstrate how the choice of vocabulary can affect the results.

The functions plotted around the circle represent the perceived directions of the sound. Cases a, b and c represent sound fields composed of discrete components–four in a and b and three in case c–with case d showing a theoretical diffuse sound field. In case a the

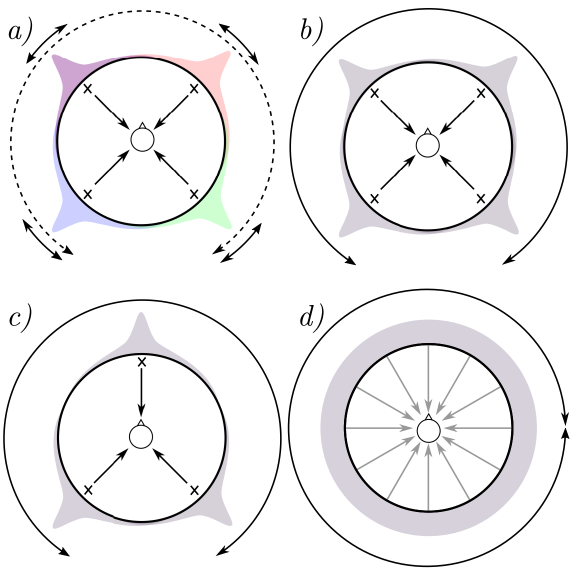


Figure 2.10: Visualisation to help show the differences between apparent source width, envelopment and perceived diffuseness.

discrete components are different scene components–for example voice, guitar, bass and drums– that are relatively easy to focus on and separate whereas in cases *b*, *c* and *d* the sound field components are all noise like and harder to differentiate. In all three of the first cases there are directions where, less or no sound is coming from. From these example cases it is possible to visualise how a listener might give very different answers for the same sound field depending on the specific phrasing of the question.

For example if asked about the extent or span from right to left of a source: in case *a* the separate sound field components are easily separated and the extent could either be of a single source or of the whole scene. In cases *b* and *c*, the noise signals are not separable and the extent is only that of the scene which is identical in both cases despite *b* being perceptually more diffuse. The extent is further complicated in 3D where there may be sources may be horizontally or vertically broad.

Alternatively, listeners could be asked to rate the how surrounding the sound field is. All of these cases are likely to be surrounding because there is sound coming from in front, behind, left and right in all cases. Envelopment is a similar high level attribute that may or may not relate to the diffuseness of the sound field. The feeling of envelopment is an overall feeling of being surrounded by the sound which may not be equivalent to localisation of the sound field to be equally from all directions. For example, the presence of gaps in the perceived direction of arrival does not necessarily limit the envelopment of the sound field.

Finally, another possible term is the "perceived diffuseness". In this context, perceived diffuseness, would relate to how the distribution of perceived directions of arrival relate to that of case *d*. This has the advantage of including any variation between the size of gaps in the directions of possible arrivals, for example the difference between cases *b* and *c*. However, for complex sound scenes such as *a*, where the components of the sound field are separable, the perceived diffuseness is once again ambiguous as it could be for an individual scene component, the whole scene or the average across all scene components.

Therefore, in this thesis, perceived diffuseness is used in general to compare between noise like stimuli as it is a feature of the number of sources as well as their distribution. Envelopment is used for complex sound scenes where the overall perception of being surrounded by the sound field is the most relevant salient feature for that type of sound field. However, when using either the perceived diffuseness or the envelopment as the feature to be rated, in both cases the description is similar and relates to the description of envelopment from Zacharov et al. (2016). The exact wording used in each experiment is reported in the listener methodology sections of the relevant chapters.

2.5.2 Perceptual Evaluations of Diffuseness

Although the majority of the work on perception of envelopment comes from the world of concert hall acoustics, there has also been a lot of work on spatial impression in loudspeaker systems.

Hiyama et al.

The work of Hiyama et al. (2002) provides a valuable basis for the work in this thesis. In their experiments, a horizontal array of 24 loudspeakers was used. The envelopment from this array was compared to layouts with fewer loudspeakers and in a range of configurations. The layouts were also investigated for a range of frequencies. The first results are shown in figure 2.11.

They found a high correlation between the number of loudspeakers and the perceivable differences in envelopment. The layout 12a is indistinguishable from the reference of 24 loudspeakers. As the number of loudspeaker reduces, the difference in envelopment from the reference also increases.

This experiment was followed by another experiment looking at different frequency ranges and also included some additional layouts with loudspeakers at non-even spacings. These results are shown in figure 2.12.

There are significant differences between different layouts which might have the same number of loudspeakers (for example the layouts with four loudspeakers). Equally there is

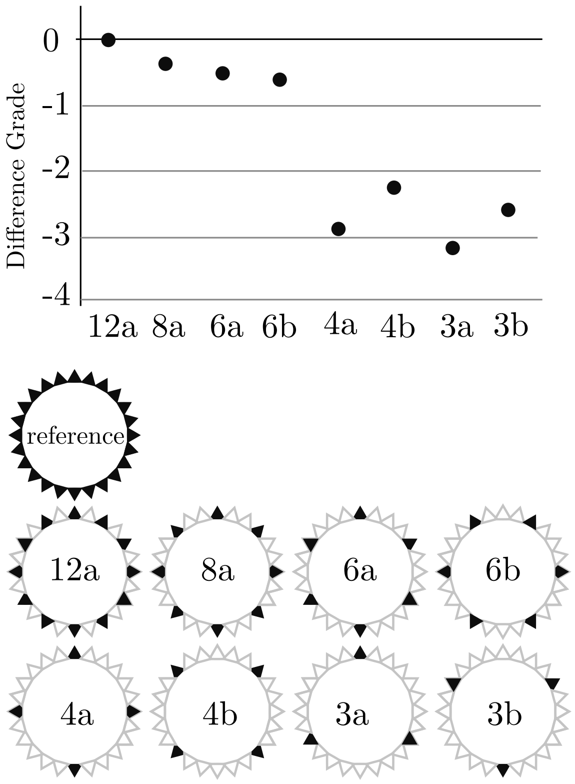


Figure 2.11: Mean difference ratings between the reference of 24 loudspeakers and reduced layouts. Adapted from Hiyama et al. (2002)

also a statistically significant factor of the frequency range used. They recommend that, to maximise envelopment, using a layout featuring a pair of loudspeakers at ±30° and a second pair around ±90-120°. This is in line with the 5.1 standardised layout. This was concluded based on a combination of the subjective results, measurements of the IACC below 1.8 kHz and the transfer function of the external ear above 1.8 kHz. They conclude that the layout of 4 loudspeakers should produce the spatial impression of a diffuse sound field up to 1.8 kHz.

This work by Hiyama is highly relevant to this thesis but lacks the third dimension of height which is increasingly common in new object- based audio formats. Paine et al. (2007) demonstrate how the additional 3rd dimension can increase immersion and relates to engulfment. This is not covered in the work of Hiyama et al.–understandably due to the added complexity arising from the additional dimension. The work of Hiyama et al. is also only ever relative to the layout of 24 loudspeakers. Although this can be assumed to be highly diffuse both physically and perceptually, it is not always clear if 24 loudspeakers is the maximum diffuseness and by using it as an explicit reference there is potential for bias.

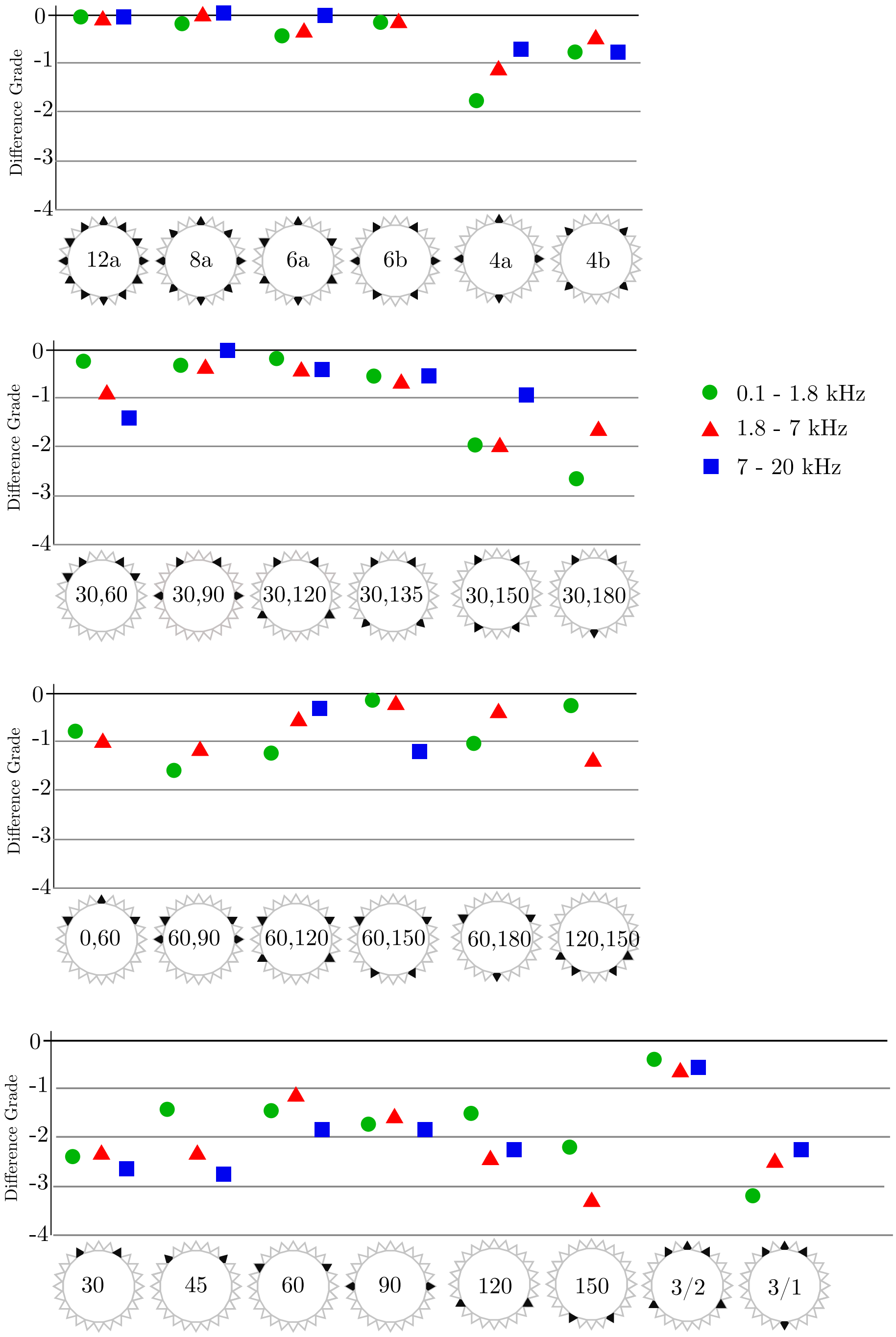


Figure 2.12: Mean difference ratings between the reference of 24 loudspeakers and reduced layouts for 3 different frequency bands. Adapted from Hiyama et al. (2002)

Santala and Pulkki

Santala and Pulkki (2011) took the 2D layout a step further limiting the loudspeakers to a frontal arc ±105° in steps of 15° and asking listeners which loudspeakers they could localise in various combinations. Figure 2.13 shows the results of their first experiment. The dark squares on the x axis indicate the positions of loudspeakers that were turned on for that stimulus. The histograms show which loudspeakers the listeners believed to be switched on.

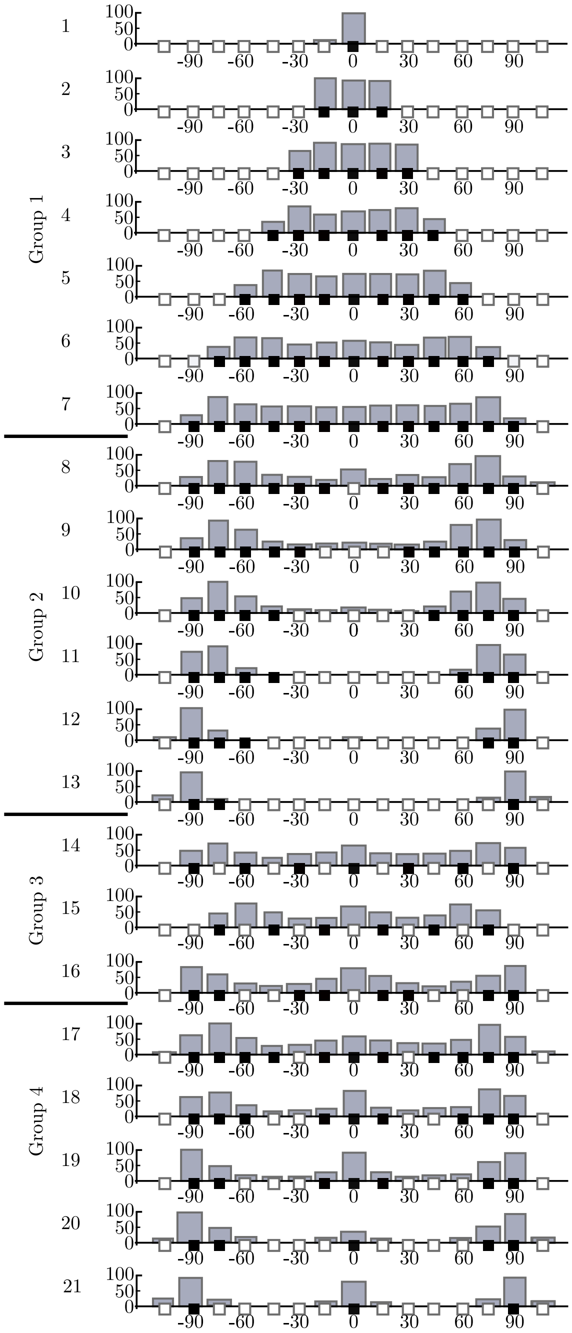


Figure 2.13: Histograms showing how often the particular loudspeaker was indicated as emitting sound, 100% being the maximum. Black boxes denote the loudspeakers that were emitting sound. Adapted from Santala and Pulkki (2011)

These results show that for wider groups of sources, the width was hard to determine accurately (group 1) than for the narrower groups of loudspeakers. Group 2 shows the existence of gaps was reliably perceived although the size of the gap was in general overestimated. In the complex scenes of group 3, the arrangement of loudspeakers was incorrectly determined. With three groups of active loudspeakers (group 4) –as was the case in group 2–the size of gaps overestimated slightly.

The second part of their experiments looked also at the frequency content of the stimuli including the centre frequency and the bandwidth and the results are plotted in figure 2.14. In this experiment difference grades were given from the reference of 13 loudspeakers to each of the reduced layouts.

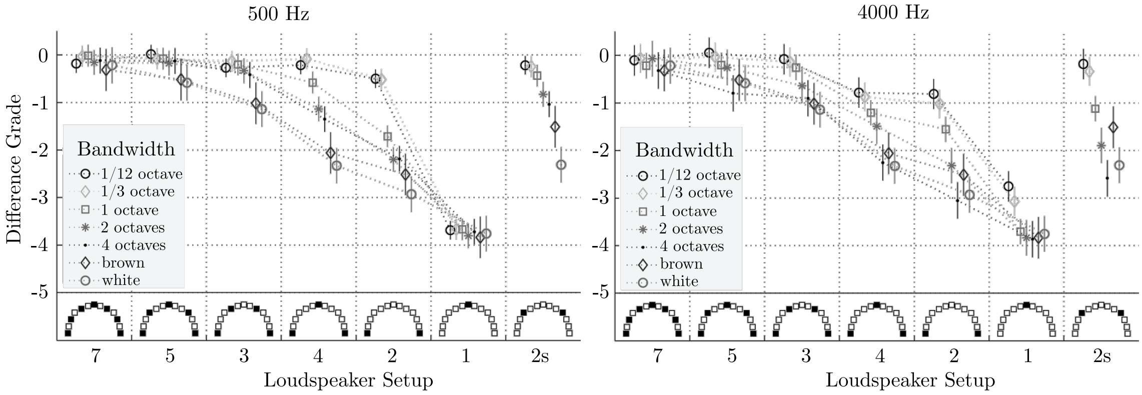


Figure 2.14: Difference in spatial impression between the reference of 13 loudspeakers and the tests subset. Adapted from Santala and Pulkki (2011)

In addition to the frequency dependence found by Hiyama et al., there is also a dependence on the bandwidth with narrow bandwidths appearing more similar to the reference than when a wide bandwidth is presented to the listener.

Once again this is a valuable starting point for this thesis but does not look at the third dimension.

Romblom et al.

Romblom et al. (2016) looked at the perception of diffuseness from the point of view of auralisation of reverberation. Their experiments were to find the threshold for lack of isotropy in the sound field arising from acoustics with different absorption coefficients within the room in three dimensions. They used loudspeakers to vary the loudness in either the front-back, left-right or up-down directions from an even distribution of sound energy to determine perceptual thresholds of level differences. Loudspeakers were placed in two rings at 0°, ±24°, ±54°, ±76°, ±104°, ±124°, ±154°and 180 azimuth angles with the two layers at 20 cm below head-height and 1.4 m above head-height for the upper layer. Both rings having a radius of 2.2 m making the upper ring 2.5 m from the listener's head. The subsets used in each condition are shown in figure 2.15.

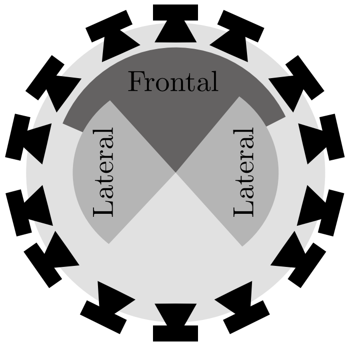


Figure 2.15: Loudspeaker layout used in experiments of Romblom et al. (2016). Head-height and above head-height layers were identical. The Lateral and Frontal conditions involved varying intensity between the loudspeakers contained within the marked arcs respectively and the other loudspeakers. The Height condition varied the intensity between the two layers (in the same layout). Adapted from Romblom et al. (2016).

The in the lateral condition, the loudness of the loudspeakers denoted by inward facing triangles on either the right or left would be varied relative to the rest of the loudspeakers. In the frontal condition the loudspeakers denoted by solid shapes were varied relative to the rest of the loudspeakers, and finally in the height condition, the upper ring was varied relative to the lower ring. An ABX test was used to determine perceptual thresholds and a cut off of 75% correct answers was used as a criteria for the threshold. The results are shown in figure 2.16.

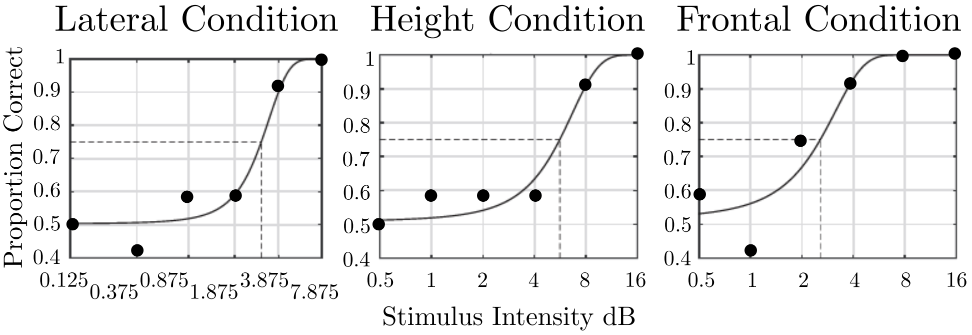


Figure 2.16: Proportion of correctly identifying the reference and the test for different relative gains for each of the conditions. Adapted from Romblom et al. (2016)

This shows the lateral condition to be most sensitive (-2.5 dB) followed by the frontal condition (-3.2 dB) and finally the height condition (-6.8 dB). They propose that the lateral condition has the largest effect on the monaural and interaural cues followed by the frontal and height conditions as is reflected in the results.

This work confirms a threshold of perceivable difference as the sound field is reduced from a highly diffuse to less diffuse. The first experiment of this thesis can be seen as an extension of the height condition in this experiment. For the lateral and frontal conditions it can be assumed the most diffuse loudness distribution is that of all loudspeakers the same level. However, in the height condition, this is not so clear depending on the definition of diffuseness. For example, the time averaged intensity will be lower for the head-height only layout rather than the 3D layout. However, in this experiment there is only a threshold calculated relative to the 3D layout (which may not be the most diffuse).

In all of these experiment the target or reference is the maximum number of loudspeakers. One of the first parts of this thesis is to determine what is subjectively the most diffuse and not to assume it is the maximum number of loudspeakers.

2.6 Listening Test Methodologies

The first part of this research is focused on the subjective evaluation of diffuseness for different stimulus material. This elicitation of subjective data needs to be appropriate to the stimuli and give meaningful results with minimal bias. The common double-blind methods for conducting listening tests are described and compared in this section.

2.6.1 Pairwise

Pairwise comparisons are a simple and very fast method to determine which of two stimuli is better or whether there is no difference. Listeners have the option to select a preference for stimulus A or B. In some cases there is also an option for no preference. If many stimuli are compared in all combinations then it is possible to get a ranking of the stimuli using frequency counts. A limitation is that there is no indication of the absolute differences between stimuli, large differences and small but perceivable differences are both rated equally.

Whilst this test is very fast for a single trial, to compare many stimuli in all combinations can take a lot of time (De Man and Reiss, 2013).

2.6.2 ABX

ABX tests are used to determine whether or not there is a discernible difference between two stimuli (Clark, 1981). Stimuli A and B are compared to the reference stimuli X where one of A or B is the reference stimuli X. If there is no difference then the selection of the correct stimulus A or B will be chance and so the expected value for each stimulus will be 50%. If the differences are large, the listener will identify the reference stimuli 100% of the time. The reference shows the listener exactly what they are looking for so gives them the best chance of discerning any slight differences.

ABX tests are most suited to near negligible degradations in the audio. The test is fast so many repeats can be made quickly. However, if the differences are not close to negligible, then listeners will nearly always correctly identify the correct stimulus and, as with the pairwise comparison method, there is no indication on the amount of degradation.

2.6.3 Triple Stimulus with Hidden Reference - BS.1116-1

The BS.1116-1 is a standardised test methodology commonly used to quantify small degradations in sound quality. It is very similar to the ABX test but also includes some indication of the amount of degradation (ITU-R, 1997). Once again the listener can switch between 3 stimuli, the reference and 2 test stimuli with one of these stimuli the same as the reference. The listener then rates both stimuli on a continuous scale labelled with 5 intervals. Therefore the stimuli they believe to be the reference should be rated at the maximum and the other rated according to the perceived amount of degradation. The 5 intervals are usually labelled as shown in table 2.3.

Grade	Impairment
5.0	Imperceptible
4.0	Perceptible, but not annoying
3.0	Slightly, annoying
2.0	Annoying
1.0	Very annoying

Table 2.3: 5 point impairment scale from BS.1116-1. (ITU-R, 1997)

This test is more time consuming than an ABX test but has the advantage of being able to compare different amounts of degradation. This test can still be used to determine if a listener can tell any difference between two stimuli. Whilst well suited to small degradations, it is less suited to medium or large impairments for reasons covered in the next section.

2.6.4 MUltiple Stimulus with Hidden Reference and Anchors (MUSHRA) - BS.1534-2

MUltiple Stimulus with Hidden Reference and Anchor tests (MUSHRA) is standardised in BS.1534-2 and is used for intermediate quality audio systems (ITU-R, 2014a). Systems are compared to a reference as with ABX and BS.1116-1 tests but in the case of MUSHRA, more stimuli are included in a single trial and so can be compared to each other in addition to the reference (figure 2.17). This allows small differences between stimuli that are not close to the reference to be compared more accurately. There is an explicit reference and selection of test stimuli. One of the test stimuli will be the hidden reference, another will be the hidden anchor. The purpose of the hidden reference is twofold, firstly it is a test of the listeners ability to identify the non-degraded signal, secondly, it anchors the top of the scale so the listeners know which stimuli to rate the highest and therefore encourage the use of the top of the scale. The anchors are used to compare the results from different tests to a known degradation. For example, the systems under test may have a range of different degradations e.g. distortion, frequency response or signal to noise ratio. A single test might look at only one of these but it might be of interest to see how distortion degradation compares to frequency response degradation. In this case a common anchor allows the results of the two experiments to be compared. However, the anchor should not be used for rescaling the data. The scales are usually labelled as in figure 2.17 or as in table 2.3.

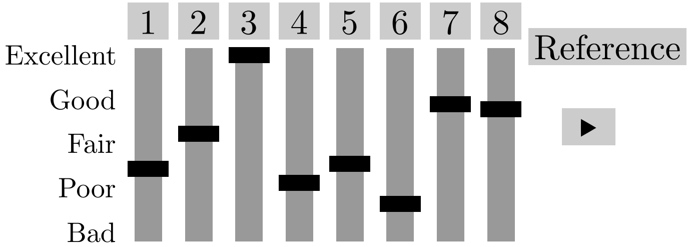


Figure 2.17: Example MUSHRA user interface. Adapted from ITU-R (2014a)

This type of test is very accurate but is not immune to potential bias. Zieliński et al. report on some potential biases of the MUSHRA test methodology depending on the choice of stimuli. They term these biases “range equalisation bias” and “stimulus spacing bias” (figure 2.18).

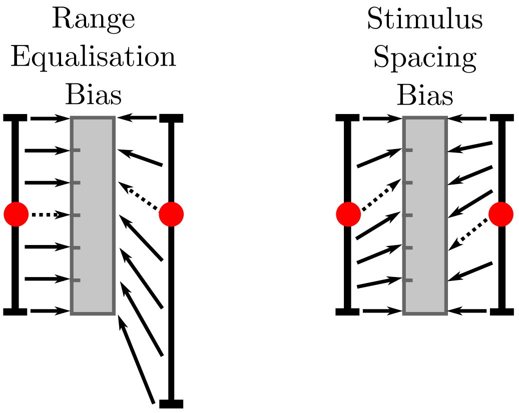


Figure 2.18: Potential Biases of MUSHRA. Left is equalisation bias, right is stimulus spacing bias. The black bars represent a range of stimuli that are mapped by listeners to a scale (the grey box). The red dot represents a stimulus that is the same for all trials but gets mapped to different parts of the output scale based on the other stimuli within the trial. Adapted from Zielinski et al. (2007).

The left figure shows “range equalisation bias". The same stimulus (red dot) is rated with a different absolute value if the total range of the stimuli is different. Listeners automatically place the most and least diffuse stimuli at the extremes of the scale thereby adjusting the limits of the scale to the given stimuli and possibly ignoring the absolute scale labels.

The figure on the right shows “stimulus spacing bias" and how the distribution of the data can effect the absolute rated values. The red dot represents the same stimulus in both trials but mapped to different parts of the scale depending on whether there are more stimuli that should be rated high, or more stimuli that should be rated low. Listeners find the ranking part of the task easier than judging the absolute difference between stimuli. For both of these biases the rank order of the stimuli remains the same, only the absolute ratings are biased. Therefore even with these biases it is possible to draw accurate conclusions on relative performance even if the absolute performance may be biased. Zielinski et al. (2007) found this deviation could be as high as 22% of the scales range .

Depending on the normality of the distribution of the data, ANalysis Of VAriance (ANOVA) can be used to identify statistically significant factors as well as the interactions between factors.

2.7 Methods of Signal Decorrelation

The theoretical diffuse sound field requires that all the signals are uncorrelated. In spatial audio this requires many discrete channels. One method to have multiple uncorrelated signals is to generate uncorrelated signals from a single source by means of decorrelation. There are several ways to perform the decorrelation each with various benefits and drawbacks. In this section a few of the possible methods of decorrelation are summarised.

Pulkki and Merimaa (2006) compare a range of decorrelation techniques. The first is amplitude panning. In this case the direction of arrival is estimated from a B-Format recording and the pressure component of the B-Format signal is amplitude panned to the estimated location. As the angle of arrival will be random for a diffuse sound field, the position will move rapidly and ideally appear diffuse. This is found to be effective at low frequencies. Unfortunately, the fast panning leads to amplitude modulation manifesting as distortion. This becomes a trade-off between the amount of distortion and the possibility of perceiving the movement of frequency components.

Convolutional diffusion involves convolving the audio with short noise bursts. These noise bursts are uncorrelated and flat in frequency response. Noise bursts longer than 50 ms affect the perceived length of the reverberation. However, limiting the length of the noise burst to 50 ms means low frequencies (below 200 Hz) cannot be decorrelated as much as higher frequencies (Pulkki and Merimaa, 2006).

Alternatively, all-pass filters with random phase responses can be used. The all-pass filter has a flat magnitude response to avoid colouration. This can be easily implemented using the Inverse Fast Fourier Transform (IFFT) of a response with flat magnitude and random phase. Any number of random phase responses can be generated and, when convolved with a single mono source, the output will be uncorrelated (Kendall, 1995). (Pulkki and Merimaa, 2006) perform the decorrelation in the STFT domain. This phase randomisation is limited by the length of the FFT window. Long FFT windows ensure minimum distortion but introduces temporal smearing. Transients have wide bandwidth but short duration. When the transient is decorrelated, the energy is rearranged to fill the whole FFT window and the transient becomes noise. This reduces the quality although attempts have been made to remove the transients from the audio prior to decorrelation to avoid smearing (Laitinen et al., 2011).

Decorrelation can be seen as a compromise between the amount of decorrelation and the artefacts generated.

2.8 Summary

This second chapter covered the relevant literature that influenced the research in this thesis. The current tools available in object-based audio were summarised and the lack of efficient ways to represent distributed sources was emphasised. The ideal diffuse sound field was quantified showing an idealised version of the type of source that is currently difficult to represent. This idealistic diffuse sound field can never be truly reproduced and so metrics were introduced that attempt to quantify diffuseness. The appropriate terminology was clarified with respect to the terminology used in other research and existing research into the the perception of diffuse sound fields in spatial audio was summarised. The methodologies of different listening tests were compared and finally, techniques for decorrelation were summarised. This literature was than used to design the experiments described in the next chapters.

Chapter 3

Experiment 1: Relative Loudness Between Head-height and Non-head-height Subsets of Loudspeakers

3.1 Overview

Advocates of 3D loudspeaker layouts claim 3D layouts to be more enveloping and surrounding than 2D layouts. A contrived example to investigate this claim would be 12 head-height loudspeakers with an additional single loudspeaker above the listener with 12 head-height loudspeakers (reported to be maximally diffuse in two dimensions by Hiyama et al. (2002)). In an informal listening test it was noted this single loudspeaker above the listener made little or no difference to the perceived diffuseness. The single loudspeaker had very little effect as the loudspeaker could not be heard over the 12 head-height loudspeakers. However, if the single loudspeaker was 12 times louder than each loudspeaker at head-height then it would be too loud, easily localised, and the sound field not diffuse. It therefore follows that there is some optimal level of this additional loudspeaker relative to the level of the head-height loudspeakers that might take the 2D layout with 12 loudspeakers to a 3D layout that is more diffuse.

In this first experiment an adjustment task allowed listeners to vary the distribution of sound energy between the loudspeakers at head-height and those not at head-height to the point they determined most diffuse using a slider. The total level of the stimulus was maintained across the length of the slider and the layouts of loudspeakers chosen investigated the effect of,

- the number of loudspeakers at head-height,
- the number of loudspeakers in layers not at head-height,
- the position of the non-head-height layer(s) above and/or below the head-height layer.

Section 3.2 covers the choice of stimulus material and the layouts of loudspeakers tested. Section 3.3 the listener response method and UI design. Section 3.4 covers the experimental set-up, the loudspeaker arrangement and system calibration with section 3.5 covering the listeners who sat the test. Section 3.6 examines the results and conclusions of the experiment. These optimised relative levels are evaluated further in the second experiment in chapter 4.

3.2 Stimuli

The goal of this experiment was to find, for a given loudspeaker layout, what loudspeaker gains maximise the perceived diffuseness. This leads to an enormous range of continuous variables. Both the quantity of loudspeakers and their positions in azimuth and elevation as well as the individual gains of each loudspeaker. To minimise the number of variables the focus was placed on the difference between 2D and 3D. Therefore the loudspeakers were evenly distributed in azimuth and the only level adjustment was the relative level between the loudspeakers at head-height and those not at head-height. The details behind the exact choice of stimuli are covered in this section.

3.2.1 Stimulus Material

It was important that all the loudspeaker signals had minimal inter-channel correlation. This is difficult with recorded audio signals and decorrelation techniques add artefacts that may bias the results (section 2.7). Uncorrelated pink noise signals were chosen over white noise as the logarithmic nature of pink noise is more relevant to humans' logarithmic hearing of frequency. Static noise stimuli are not subject to material preference bias, do not vary with time and therefore, should give repeatable results.

Uncorrelated pink noise was generated using the dsp.ColoredNoise object in Matlab. These signals are generated using independent random sequence generators to ensure a correlation coefficient of 0 (for a long sequence).

3.2.2 Loudspeaker Layouts

Each stimulus was a 3D layout of loudspeakers divided into the subset of loudspeakers at head-height–the 2D component–and the subset of loudspeakers not at head-height–the additional 3D component. The Audio Lab at the University of Southampton features 37 loudspeakers that could be split into 10 subsets of loudspeakers, 3 head-height subsets (table 3.1) and 7 non-head-height subsets (table 3.2). Each 3D stimulus comprised of a head-height subset and a non-head-height subset and allowed the listener to cross-fade between the two.

Azimuth	Elevation	$n = 12$	$n = 6$	$n = 4$
0	0	X		
± 30°	0	X	X	X
± 60°	0	X		
± 90°	0	X	X	
± 120°	0	X		
± 150°	0	X	X	X
180°	0	X		

Table 3.1: Azimuth and elevations of loudspeakers in the head-height subsets.

Azimuth	Elevation	12/n/13	8/n/8	8/n/0	0/n/8	0/n/4w	0/n/4	0/n/1
± 45°	-56°	X						
± 135°	-56°	X						
0°	-20°	X	X	X				
± 45°	-17°	X	X	X				
± 90°	-24°	X	X	X				
± 135°	-17°	X	X	X				
180°	-20°	X	X	X				
0°	27°	X	X		X			
± 45°	24°	X	X		X	X		
± 90°	32°	X	X		X			
± 135°	24°	X	X		X	X		
180°	27°	X	X		X			
± 45°	52°	X					X	
± 135°	52°	X					X	
0°	90°	X						X

Table 3.2: Azimuth and elevation angles of the loudspeakers in the non-head-height subsets.

The subsets were chosen to investigate; the number of loudspeakers at head-height; the number of loudspeakers in the layer(s) not at head-height; and the position of the non-head-height layer(s) either above and/or below the head-height layer. Standardised layouts were not used in the first test as they are not evenly distributed in azimuth around the listener and therefore would not have a constant energy from the front and the back as the relative level between head-height and non-head-height subsets was varied.

The non-head-height subsets are labelled in the format $m_B/n/m_A$ where m_B is the number of loudspeakers below head-height, m_A is the number of loudspeakers above the head-height and n is the number of loudspeakers at head-height.

The head-height subsets, $n = 12$, $n = 6$ and $n = 4$, compare different numbers of head-height loudspeakers (table 3.1). The $n = 12$ loudspeaker subset is designed to have maximum horizontal diffuseness (Hiyama et al., 2002). The subset $n = 6$ is still evenly distributed around the listener but perceptually less diffuse than the $n = 12$ subset (Hiyama et al., 2002). The $n = 4$ subset has no loudspeakers at ±90° and is not evenly distributed around the listener however the energy from in front of the listener and from behind the listener is the equal. The $0/n/8$, $0/n/4$ and $0/n/1$ subsets were chosen to compare variations with number of non-head-height loudspeakers. Subsets of 4 elevated loudspeakers in wide and narrow arrangements ($0/n/4w$ and $0/n/4$) were chosen to compare different elevations with the same number of loudspeakers. The $8/n/0$, $8/n/8$ and $0/n/8$ subsets were chosen to compare how layers above and/or below the head-height layer would be adjusted. The $8/n/8$ and $12/n/13$ subsets were chosen to compare the effect of number of channels when there are many non-head-height loudspeakers. All combinations of head-height and non-head-height subsets were investigated, leading to 21 unique subset pairs. Each stimulus was assessed and adjusted twice by each listener.

3.3 Listener Response Method

Listeners were asked to perform an adjustment task that involved moving a slider to find the most diffuse setting which corresponds to a specific relative level between the two loudspeaker subsets. To avoid the expectation of the listener biasing the results, the listeners were not informed of the effect of the slider. Listeners were asked to;

> “Move the slider from left to right, in order to vary the spatial attributes of a noise stimulus. Your task is to find the point at which you perceive it most diffuse.”

The definition of diffuseness for this experiment was,

> "Diffuseness is defined in this experiment as the sound coming from all directions with equal intensity. Therefore, the sound should ideally be impossible to localise and without any gaps (areas you perceive there is no sound coming from) all in three dimensions."

A user interface was designed in Max 6.1 (figure 3.1) that featured the slider, a button to play or pause the stimulus, a button to progress to the next stimulus, and a slider that allowed ±2 dB of gain (however all listeners left the gain at 0 dB) .

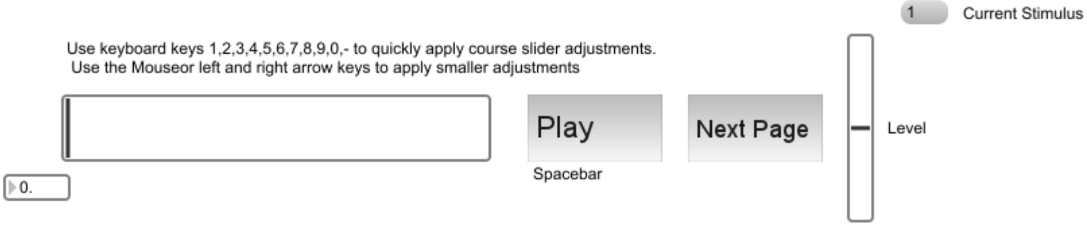

Use keyboard keys 1,2,3,4,5,6,7,8,9,0,- to quickly apply course slider adjustments.
Use the Mouseor left and right arrow keys to apply smaller adjustments

Figure 3.1: Screenshot of user interface.

All subsets were aligned to the same SPL at the central listening position. The total level of the stimuli was maintained across the length of the slider using a constant power (-3 dB) cross-fade between the head-height and non-head-height subsets. The slider value (s_s) ranged from 0 to 1. The levels of the head-height (X) and non-head-height (Y) subsets are given in dB by,

$$X = 10 \log_{10}(s) \tag{3.1}$$

and

$$Y = 10 \log_{10}(1 - s) \tag{3.2}$$

respectively where $s = s_s$, if the non-head-height subset was on the left or $s = 1 - s_s$, if the head-height subset was on the left.

This leads to an Inter-Subset Level Difference (ISLD) in dB of,

$$ISLD = 10 \log_{10}(s/(1 - s)) \tag{3.3}$$

Positive ISLD values indicate the head-height subset is that number of dB louder than the non-head-height subset. Negative ISLD values indicate the non-head-height subset is louder.

Keyboard shortcuts were provided to allow the slider to be changed quickly without looking at the computer screen. The stimulus presentation order, which of the two subsets was presented on the left, and the slider start position (left or right) were all randomised. This encouraged the listener to listen to the stimulus rather than make judgements based on previous ratings. The 42 stimuli were rated in two sessions of 20 minutes with a 5 minute interval.

3.4 Reproduction System

3.4.1 Room and Loudspeakers

The Audio Lab at the University of Southampton measures 4.80 m×3.97 m×2.56 m and had a reverberation time of 0.12 s ±0.02 s in 1/3 octave bands between 125 Hz and 8 kHz.

The loudspeakers were 37 Kef HS3001SE driven by a PC with 40 channel RME audio interface were arranged into 6 layers all evenly distributed in azimuth around the listener (figure 3.2).

Loudspeakers were mounted at uniform height from the floor and so the elevation angle varies with distance to the listener.

3.4.2 Calibration

All individual loudspeakers were aligned in level, time and frequency. Firstly, digital gain was applied to each loudspeaker so the Sound Pressure Level (SPL) at the listening position would be the same for a given input signal. To avoid any potential precedence effects, all loudspeakers were also time aligned to within 100 μs (ITU-R, 2014b) using digital delay.

The frequency response also varied between loudspeakers due to strong ceiling and floor reflections and the necessary positioning of loudspeakers close to the walls of the room. Frequency response matching was implemented using 1/6 octave band equalisation for each loudspeaker individually. Five seconds of pink noise was replayed over each loudspeaker; recorded using a calibrated B&K free-field microphone type 4190 with B&K preamplifier type 2669 pointed towards the loudspeaker; and analysed using a bank of 1/6 octave filters implemented in Matlab. The target response was the 1/6 octave band filtered input signal weighted below 95 Hz to mimic the low frequency roll-off of the loudspeaker. The error,

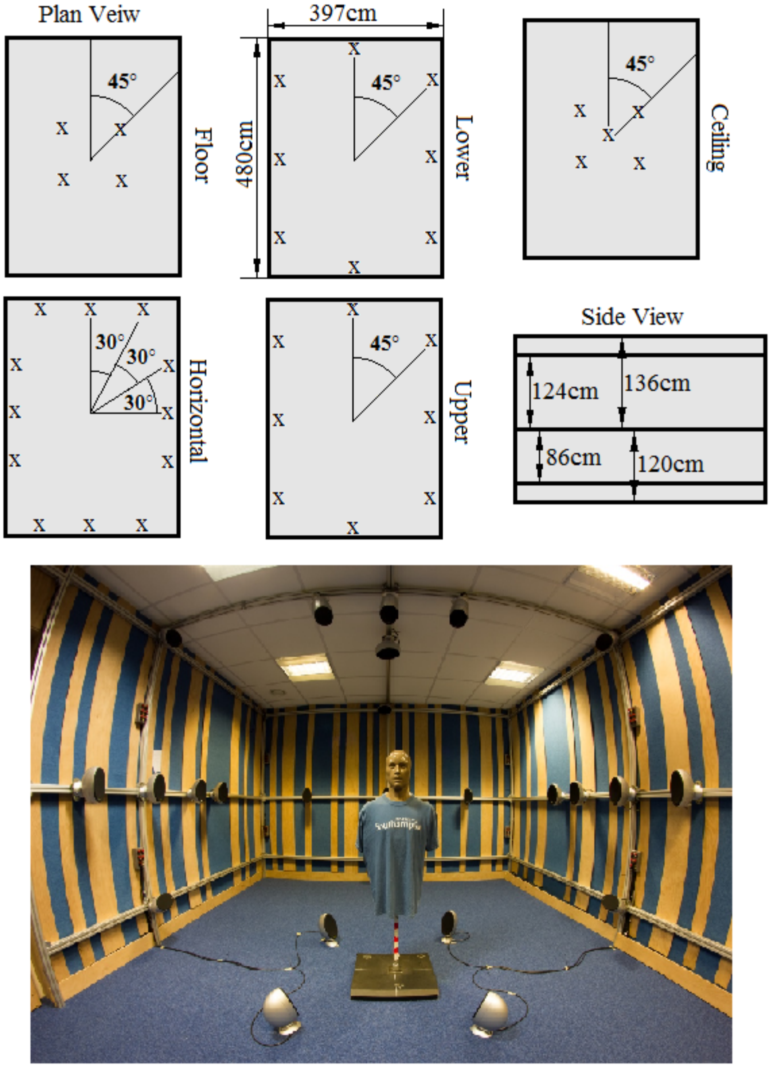


Figure 3.2: The Audio Lab at the University of Southampton.

in dB, between the target response and the measured response was added to the gain coefficients for each sub-band and the process was repeated until the error was within ±0.5 dB for all sub-bands above 95 Hz. The iterative process allowed the frequency selectivity of the filters and any extraneous noise during the "record" stage to not affect the choice of sub-band coefficients. An example 1/6 octave band analysis of the signal, at various stages in the equalisation process, is shown with the resultant sub-band coefficients and residual error in figure 3.3. The coefficients generated were then applied to the filtered stimulus material (in this case also pink noise).

Some of the frequency irregularity is due to reflections in the room. Over equalisation could lead to difficulties at other frequencies as the listener moves slightly and so 1/6 octave band equalisation was used in place of techniques that invert the frequency response. Unfortunately there may still be slight colouration differences between stimuli

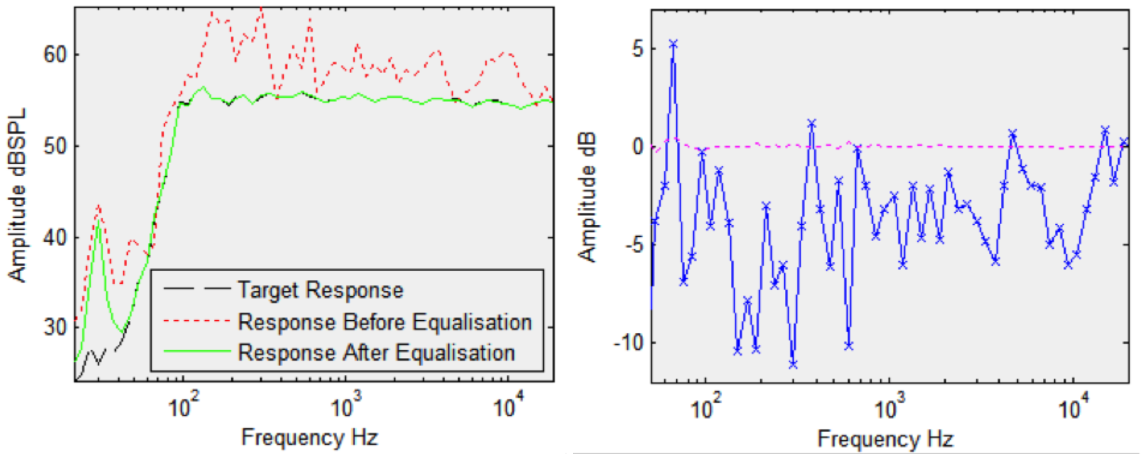


Figure 3.3: 1/6 octave band equalisation of the loudspeaker at head-height, directly in front of the listener. The blue line shows the sub band gains and the pink line shows the remaining error between the target response and the response after equalisation.

as the ears of the listener are not exactly at the alignment position although there was a clear improvement in tonal colour consistency following the equalisation.

3.5 Subjects

The listening test was approved by the ethics and research governance committee (ID: 11554). The listening test was sat by 16 PhD/Masters students at the University of Southampton with self-reported normal hearing.

3.6 Results

The results are first screened for listener quality and consistency. The remaining results are then analysed, compared and conclusions are drawn.

3.6.1 Post-screening

The listeners were compared on their ability to produce repeatable results. The difference between the two repeats of the same stimulus for each listener was used as a metric for consistency. The figure 3.4 shows the mean absolute difference between repeats for each listener.

Most listeners are, on average, less than 0.2 between the repeats of the same stimulus. Listeners who had, on average, greater than 0.2 difference between their adjustment for the same stimulus were excluded. This removed listeners 4, 5, 11, 14 and 15. Visual inspection of the histograms for each stimuli showed the remaining listeners to be fairly consistent

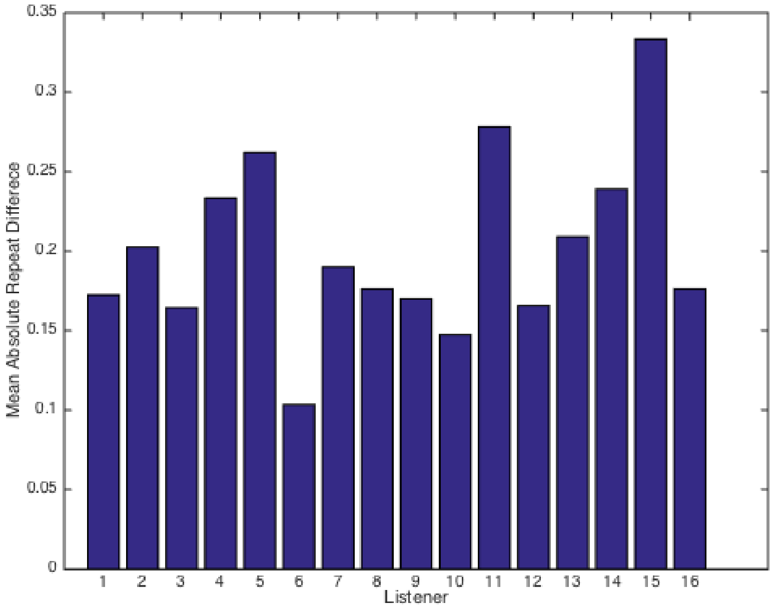


Figure 3.4: Listener consistency based on the difference between the adjustments given to the same stimulus across the two repeats.

with the data close to normal distribution. The cut-off of 0.2 is arbitrary and chosen to remove some of the least consistent listeners whilst retaining a sufficient number of data points.

With inconsistent listeners removed, the two repeats for each listener and each stimulus were averaged and are plotted in figure 3.5.

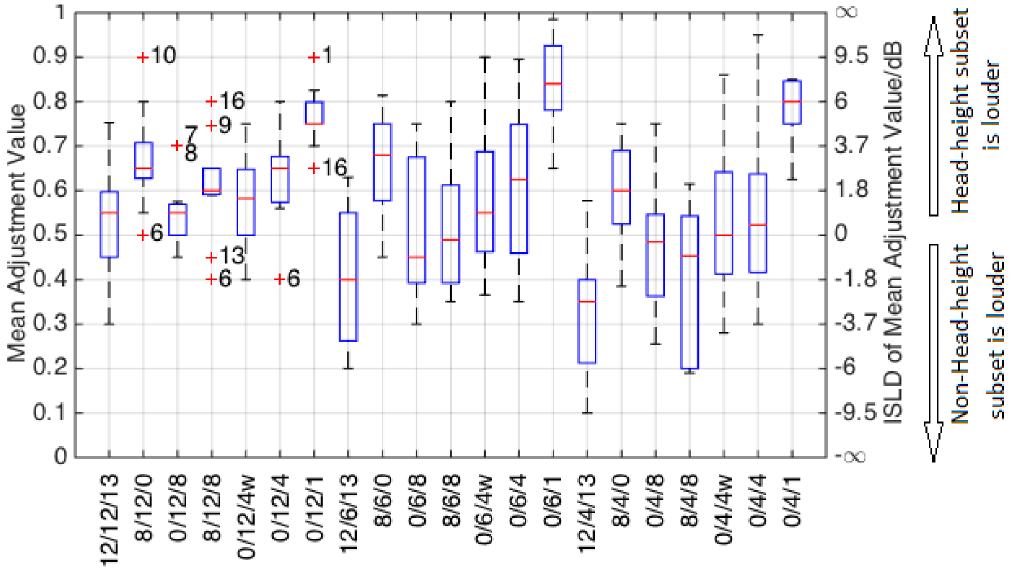


Figure 3.5: Box and whisker plot of data from all consistent listeners. Point labels are listener IDs.

There are several outliers for the $n = 12$ stimuli however, as many listeners have already been removed, this may be slightly exaggerated by the small number of data points. None of the outliers are very far from the median so no further listeners were excluded. The means slider values of the remaining data (following removal of the inconsistent listeners) are used from this point. The mean is used in place of the median due to the low number of data points and the removal of inconsistent listeners means the advantages of the median being unbiased by outliers is less relevant. The slider values are averaged in the following plots but these values are converted to the ISLD and labelled as the $ISLD_{Mean}$ as the ISLD is a more tangible concept than the slider value. The results show the choice of level to vary depending on both the number of loudspeakers in the head-height and non-head-height layers, as well as the position of the non-head-height layer.

3.6.2 Analysis and Discussion

The mean slider values are plotted in figure 3.6 and labelled with the associated ISLD.

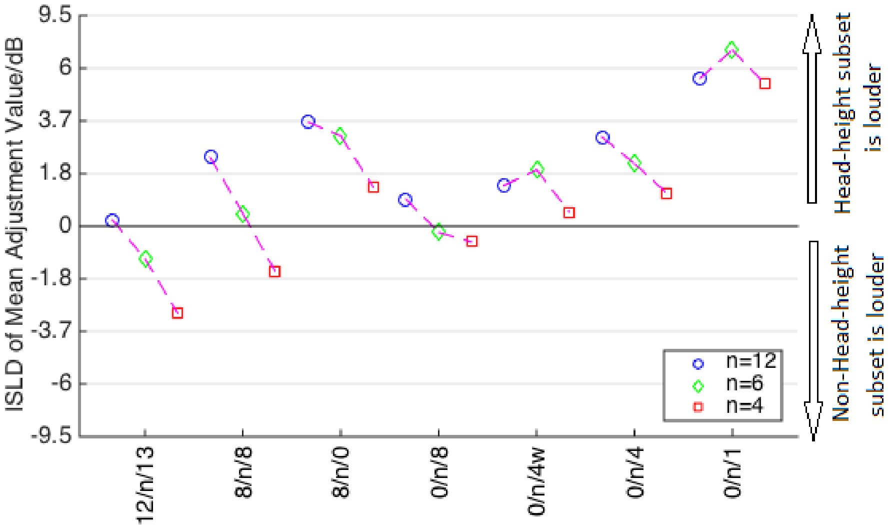


Figure 3.6: ISLD of mean adjustment from consistent listeners. Pink lines show the effect of changing of the number of head-height loudspeakers.

Most of the $ISLD_{Mean}$ are above 0 dB meaning, in general, the head-height subset was preferred to be louder than the non-head-height subset. The number of loudspeakers in each subset as well as the position of the subset appears to make a difference to the preferred ISLD.

Repeated measures ANalysis Of VAriance (ANOVA) shows both the head-height and non-head-height subset to be significant factors ($F(2, 9) = 6.674$, $p = 0.017$ and $F(6, 5) = 16.815$, $p = 0.004$ respectively) with p-values less than 0.05 indicating statistical significance at the 95% confidence level and high F-statistic indicating good "explained variance to unexplained variance" ratio.

The significant differences when looking at pairwise comparisons tend to be between the more extreme layout differences. This is to be expected considering the difficulty of the task. Romblom et al. (2016) determined the minimum perceptual level difference between a head-height layer and an elevated layer to be -6.8 dB . In the format of this experiment, this would be equivalent to an ISLD of -6.8 dB for a layout of 0/14/14 being the just noticeable difference relative to an ISLD of 0. Therefore the fairly wide range seen here is to be expected and because there are relatively few listeners, the statistical significance suffers.

Number of Head-Height Loudspeakers In general, as the number of head-height loudspeakers increases from $n = 4$ to $n = 12$, so does the ISLD (level of the head-height subset relative to the non-head-height subset). This is regardless of the number or distribution of the non-head-height subset.

Pairwise comparison in table 3.3 show insignificant differences between similar subsets. This is to be expected considering the difficulty of the task and therefore the wide standard deviation. There is a significant difference in the preferred ISLD between $n = 12$ and $n = 4$.

	$n = 6$	$n = 12$
$n = 4$	0.128	**0.016**
$n = 6$		0.200

Table 3.3: Significance values from pairwise comparisons between head-height subsets. Significance has Bonferroni adjustment for multiple comparisons.

Number of Non-Head-Height Loudspeakers The $0/n/8$, $0/n/4$ and $0/n/1$ non-head-height subsets were included to investigate the effect of the number of elevated loudspeakers. Stimuli with more elevated loudspeakers are adjusted to a lower ISLD. The clearest result is that of the $0/n/1$ subset that was consistently adjusted to be low in level relative to the head-height subset. This was the only ISLD that was statistically significant in the pairwise comparisons between the non-head-height subsets (table 3.4). Informal listener feedback following the test indicated they found the subsets with fewer non-head-height loudspeakers (especially $0/n/1$) very localisable and therefore adjusted them to a low level to hide this highly localisable subset.

	$8/n/8$	$8/n/0$	$0/n/8$	$0/n/4w$	$0/n/4$	$0/n/1$
$12/n/13$	0.337	**0.008**	0.271	0.322	0.032	**0.000**
$8/n/8$		0.073	1.000	1.000	0.925	**0.000**
$8/n/0$			0.242	1.000	1.000	**0.006**
$0/n/8$				1.000	0.860	**0.000**
$0/n/4w$					1.000	**0.008**
$0/n/4$						**0.002**

Table 3.4: Significance values from pairwise comparisons between non-head-height subsets. Significance has Bonferroni adjustment for multiple comparisons.

Distribution of Loudspeakers Although the number of loudspeakers in the upper hemisphere appears to be directly related to the $ISLD_{Mean}$, the number is not the only contributing factor. Of the $8/n/0$, and $0/n/8$ stimuli, the $8/n/0$ subset was kept at the lowest level relative to the head-height layer. This may be because surrounding sounds from the lower level are unexpected or less desirable. The $0/n/4w$ is rated at a similar ISLD to the $0/n/4$ stimuli with the same number of loudspeakers indicating the same number of channels in the same hemisphere makes little difference to the choice of level distribution. However, if you consider the limit, if the elevation increased to 90° this arrangement of four loudspeakers would be equivalent to $0/n/1$ and so the ISLD would likely change change in that scenario. However, from the data here, the small change in elevation from approximately 24° to 52° appears to have no effect on the most perceptually diffuse ISLD.

ISLD vs. Gains of Individual Loudspeakers The general trend is that the optimum ISLD increases as either, the number of head-height channels increases, or the number of non-head-height channels decreases. Intuitively this equivalent to turning down the subset with the fewer loudspeakers. For this reason the optimum ISLD values were plotted against the ratio of the number of loudspeakers between the head-height and non-head-height layers in figure 3.7. Also plotted are two comparison lines, the ISLD if both subsets were at the same level (ISLD=0) and the ISLD if the Inter-Subset Channel Level Difference (ISCLD) equals 0. The ISCLD=0 can be seen as the ISLD adjustment that would have been given had the listener decided the most diffuse ISLD is when each individual channel in the head-height and non-head-height layers are at the same level.

Neither of these curves line up perfectly with the $ISLD_{Mean}$. Ideally a model could be used to predict the optimum ISLD for a given layout. However, the $ISLD_{Mean}$ varies not only based on the ratio of the number of channels, but also their position (in the case of $0/n/8$ vs. $8/n/0$). This implies some psychoacoustic behaviour that cannot be captured by

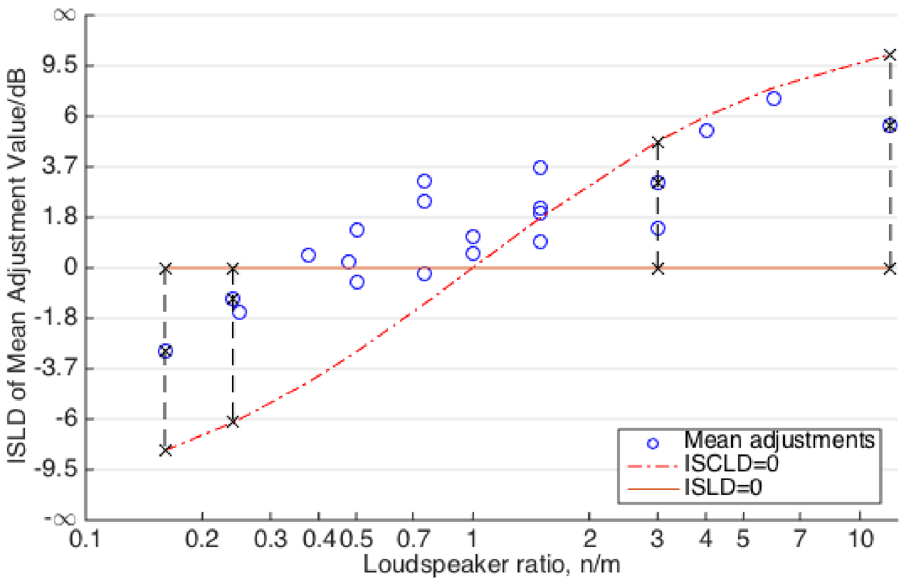


Figure 3.7: Comparison of the ISLD of the mean adjustment value to the ISLD of equal level from each subset or equal level from each channel. Black lines identify four stimuli where ISLD varies most between $ISLD_{Mean}$, ISLD=0 and ISCLD=0 and so were also tested in experiment 2 to validate and quantify these results.

the ratio of the number of loudspeakers. Therefore any metric based only on the number of loudspeakers in both subsets can never be predict all the data. It is also not clear from the data in this experiment the magnitude of the difference in diffuseness when choosing the wrong ISLD.

ISLD and the Ratio of Loudspeakers in Existing Loudspeaker Systems The ratio of the number of head-height and number of non-head-height loudspeakers can be calculated for several standardised loudspeaker layouts and compared to the results in figure 3.7. The layouts used in later parts of this thesis are 9.1(5+4) and 22.2. For these layouts the respective "loudspeaker ratios" are 1.25 and 0.833 respectively. For both of these layouts, using the ISCLD=0–the red dash-dot line in figure 3.7–seems a reasonable choice and confirms the current practice of aligning all loudspeakers to the same level.

3.7 Summary

In this first experiment the relative gains between the loudspeakers at head-height and those not at head-height were optimised to maximise the perceived diffuseness. This was found to correlate with the ratio of head-height to non-head-height loudspeakers but not sufficiently strongly to create an accurate model. For the loudspeaker systems tested, the $ISLD_{Mean}$ can be seen as an optimised level distribution to maximise the perceived diffuseness. For loudspeaker systems not included in this experiment, the ratio of the number of loudspeakers is a good way to approximate an optimal ISLD and for standardised layouts without extreme ratios in the number of head-height loudspeakers to non-head-height loudspeakers, an ISCLD=0 with all individual loudspeakers at equal level is a suitable choice. This optimisation experiment should allow different loudspeaker layouts to be compared fairly and in an unbiased manner especially for unusual loudspeaker layouts with large differences in the number of head-height and non-head-height loudspeakers. The next experiment includes some stimuli chosen to validate this experiment and to quantify the improvements in the perceived diffuseness when optimising the ISLD.

Chapter 4

Experiment 2: Subjective Diffuseness of Different Arrangements of Loudspeakers

4.1 Overview

The second experiment was designed to critically assess the diffuse field reproduction performance of existing loudspeaker systems and investigate the relative effect of different parameters of layer based loudspeaker systems. The results from experiment 1 (chapter 3) were used to maximise the perceived diffuseness for each arrangement of loudspeakers.

Section 4.2 covers the choice of stimuli and how they relate to those in the first experiment. Section 4.3 covers the experimental design. The reproduction system and the subjects are covered in sections 4.4 and 4.5 respectively. The results are covered in section 4.6 and conclusions are drawn from the results from both of the first two experiments. Section 4.7 attempts to link the subjective results from this second listening test to objective measures of the sound field.

4.2 Stimuli

Stimuli were chosen to investigate a range of different factors covered in this section. These factors led to the choice of stimuli shown in table 4.1. The layouts are labelled in the format $m_B/n/m_A$ where m_B is the number of loudspeakers below head-height, m_A is the number of loudspeakers above the head-height and n is the number of loudspeakers at head-height.

These layouts were chosen to investigate the factors laid out in the following subsections.

	Azimuth	Elevation	Stereo*	5.0	9.0	9.0b	22.0	0/4/0	0/6/0	0/8/0	0/12/0	0/6/1	0/6/4	0/6/8	8/0/0	0/0/8	0/12/1	0/12/4	12/4/13	12/6/13	12/12/13*
4 × Floor	±45°	-56°																	✓	✓	✓
	±135°	-56°																	✓	✓	✓
8 × Lower	0°	-20°					✓								✓				✓	✓	✓
	±45°	-17°					✓								✓				✓	✓	✓
	±90°	-24°													✓				✓	✓	✓
	±135°	-17°													✓				✓	✓	✓
	180°	-20°													✓				✓	✓	✓
12 × Head-height	0°	0°		✓	✓	✓	✓			✓	✓						✓	✓			✓
	±30°	0°	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			✓	✓	✓	✓	✓
	±60°	0°					✓			✓	✓						✓	✓			✓
	±90°	0°					✓		✓	✓	✓	✓	✓	✓			✓	✓		✓	✓
	±120°	0 °		✓	✓	✓	✓			✓	✓						✓	✓			✓
	±150°	0°						✓	✓		✓	✓	✓	✓			✓	✓	✓	✓	✓
	180°	0°					✓			✓	✓						✓	✓			✓
10 × Upper	0°	27					✓							✓		✓			✓	✓	✓
	±30°	26°			✓																
	±45	24°					✓						✓	✓		✓			✓	✓	✓
	±90°	32°					✓							✓		✓			✓	✓	✓
	±135°	24°			✓		✓						✓	✓		✓			✓	✓	✓
	180°	27°					✓							✓		✓			✓	✓	✓
5 × Ceiling	±45°	52°				✓												✓	✓	✓	✓
	±135°	52°				✓												✓	✓	✓	✓
	0°	90°					✓					✓					✓		✓	✓	✓
Head-height and non-head-height subsets generate same SPL. ISLD=0. (subscript L)																	✓	✓	✓	✓	
SPL difference between head-height and non-head-height subsets is optimised. ISLD=$ISLD_{Mean}$ (subscript M)												✓†	✓†	✓†			✓	✓	✓	✓†	✓†
All loudspeakers reproduce the same SPL at the the listening position. ISCLD=0 (subscript C)			✓†	✓†	✓†	✓†	✓†	✓†	✓†	✓†	✓†				✓†	✓†	✓	✓	✓	✓†	

Table 4.1: Active loudspeakers for all stimuli and choice(s) of relative level between the head-height and non-head-height loudspeakers. *High and low hidden anchors. † Stimuli also tested off-centre.

4.2.1 Comparing Standard Layouts

The standard layouts, stereo, 5.1(ITU-R, 2007), 9.1(Daele and Baelen, 2012; ITU-R, 2014b) and 22.2 (ITU-R, 2014b, 2011) are popular multichannel formats and so are of commercial interest. In this experiment the Low Frequency Effects (LFE) channel(s) were not used and neither were subwoofers, so stimuli are labelled as 5.0, 9.0, 9.0b and 22.0 to avoid confusion. An alternate arrangement of 9.0 (termed 9.0b) using a narrower arrangement of elevated loudspeakers was also included as there is a commercial interest in systems that uses beam-forming or directional loudspeakers to direct sound that reflects off the ceiling to give virtual elevated loudspeakers. This narrow arrangement is more akin to these systems than the standard 9.0 layout specified in (ITU-R, 2014b).

4.2.2 Investigating Individual Parameters of Layer Based Layouts

There are infinite ways in which loudspeakers can be arranges in terms of quantity, azimuth and elevation. Instead of testing all possible loudspeaker arrangements, various parameters that define the arrangement of loudspeakers were investigated. To test these parameters, layouts distributed evenly in azimuth and separated into horizontal layers were used to avoid too many variables.

Number of Head-Height Loudspeakers Although previously tested by Hiyama et al. (2002), the relative effect of the number of loudspeakers in a horizontal plane at head-height is of interest and allows comparison with the earlier research. This was investigated using the stimuli 0/12/0, 0/8/0, 0/6/0 and 0/4/0. The 0/12/0 and 0/6/0 stimuli are evenly distributed in azimuth but 0/4/0 and 0/8/0 are not be due to constraints on the number of channels available for all stimuli and the chosen layouts. In these cases, symmetry was maintained front to back and left to right with the same amount of energy from in front of and behind the listener.

Number of Non-Head-Height Loudspeakers Arrangements of 8, 4 and 1 loudspeaker(s) above head-height were chosen. For each case there were also 6 head-height loudspeakers in order to focus on the improvement over the 2-dimensional layout and therefore allowing comparison between 3D layouts with different numbers of non-head-height loudspeakers and the 2D layout with the same head-height arrangement but no non-head-height loudspeakers. $0/6/m_A$ was chosen as they are evenly distributed in azimuth and moderately diffuse but not maximum horizontal diffuseness. As we are interested in the maximum improvement available from adding non-head-height loudspeakers, the optimal ISLD should be used. This was the mean of the adjusted value from experiment 1 labelled $ISLD_{Mean}$ from section 3.6. Stimuli using $ISLD_{Mean}$ are indicated with subscript M. The wider arrangement of 4 non-head-height loudspeakers from experiment 1 allows better comparison with the 8 non-head-height loudspeakers case as the loudspeakers are also at the same elevation.

Positioning of a Single Layer of Loudspeakers 8 loudspeakers were place below, at, and above head-height (8/0/0, 0/8/0 and 0/0/8 respectively). The loudspeakers in the 0/8/0 stimuli could not be evenly distributed around the listener so, as before, front-back and left-right symmetry was maintained.

4.2.3 Quantifying Improvement in Perceived Diffuseness Relative to 2D Layouts

It had been found previously by Hiyama et al. (2002) that adding more than 12 loudspeakers in a head-height layer did not increase perceived diffuseness. It was therefore of interest whether adding more loudspeakers not at head-height would have any affect. This was investigated by comparing maximum head-height diffuseness (0/12/0) with the same stimulus with many non-head-height loudspeakers. The 12/12/13 layout from experiment 1 was chosen with $ISLD_{Mean}$ for the ISLD having many additional loudspeakers. The layouts 0/12/1 and 0/12/4 also test possible increases in perceived diffuseness over 2D but using far fewer loudspeakers.

4.2.4 Validating Experiment 1

The $ISLD_{Mean}$ from experiment 1 for each stimulus should be the most perceptually diffuse ISLD possible. However, the results of experiment 1 do not show how much more diffuse the $ISLD_{Mean}$ is over using a simpler method of choosing the ISLD such as ISLD=0 or ISCLD=0. Stimuli were chosen to ensure the results from experiment 1 were valid and to quantify the effect of the ISLD as a parameter of layer based loudspeaker layouts. To best show the effect of the ISLD, the stimuli that gave the greatest range of ISLD between the ISLD=0, ISLD=$ISLD_{Mean}$ and ISCLD=0 were chosen. These were selected based on figure 3.7 to be stimuli, 12/6/13, 12/4/13, 0/12/4 and 0/12/1 in all combinations with the 3 ISLD options. The different ISLD are indicated by subscript L, M and C for ISLD=0, ISLD=$ISLD_{Mean}$ and ISCLD=0 respectively.

4.2.5 ISLD for Layouts Not Included in Experiment 1

For the systems tested, the optimal ISLD can be assumed to be the $ISLD_{Mean}$. For layouts not included in experiment 1, the ISLD=0 or ISCLD=0 (keeping the layers at equal level or individual loudspeakers at equal level respectively) are the two simplest options. The error between using a simple loudspeaker gain such as ISCLD=0 or ISLD=0 and using the “correct" value ($ISLD_{Mean}$) can be seen in figure 4.1. This shows the mean square error between the mean s values from the listeners and the s values that relate to either maintaining equal loudspeaker loudness (where ISCLD=0) or equal head-height to non-head-height loudness (ISLD=0). In the worst three cases (0/6/1, 0/12/1, and 0/4/1), the ISLD=0 has a higher squared error. For this reason ISCLD=0 is recommended the best choice for ISLD when the $ISLD_{Mean}$ is not known and therefore used for all the standardised layouts.

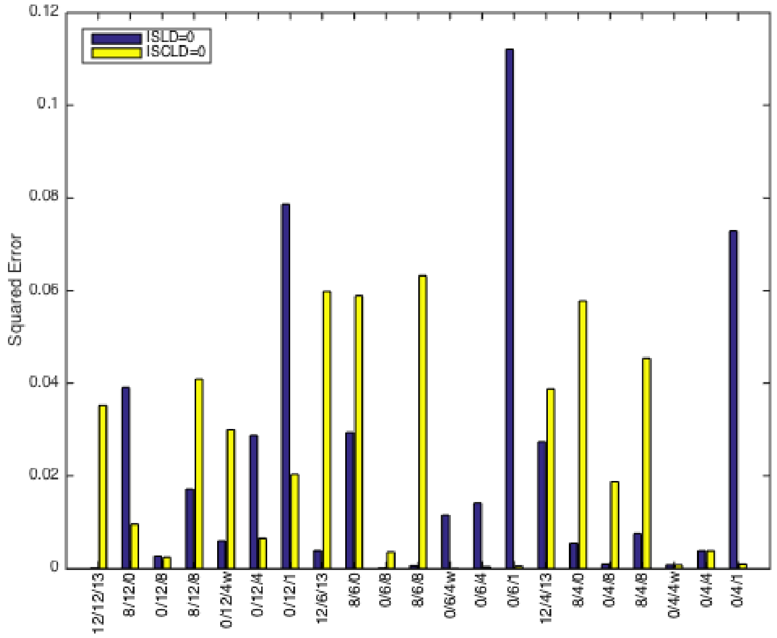


Figure 4.1: Squared error between the mean adjustment value s and the s value that would maintain ISLD=0 or ISCLD=0. said differently, the mean square error between the slider position chosen by the listener and the slider position that would be chosen based on either ISCLD=0 or ISLD=0

4.2.6 Assessing Robustness to Listener Movement

In real world scenarios, the listener is often not in the sweet-spot and therefore good diffuseness reproduction performance should be robust to listener movement. Investigating all of the research questions above in two listening positions gives too many stimuli to realistically run in a single sitting of a listening test. The stimuli chosen to validate the first experiment were therefore not included in the off-centre listening position. However, when designing the UI, there was space for 2 more stimuli off-centre and so the 12/6/13 stimuli with ISCLD=0 and ISLD=$ISLD_{Mean}$ were included to give an impression of whether $ISLD_{Mean}$ is robust to listener movement.

The off-centre position was chosen to be 80 cm to the right of the on-centre listening position. This was half the distance to the nearest head-height loudspeaker and a reasonable approximation to a domestic “large sofa” scenario. Moving right/left affects the interaural differences (ITD and ILD) more than moving forwards/backwards so is likely a more severe test of robustness.

The loudspeakers were not recalibrated for the off-centre position. The listener is off axis to many loudspeakers so equalising the new position might lead to very extreme equalisation especially at high frequency (HF). The equalisation for the centre position was maintained instead.

4.3 Listener Response Methodology

The listener response method chosen for experiment 2 was a MUSHRA-style listening test modified to make it more appropriate for the given task and stimuli. The MUSHRA test lends itself to comparing medium and large degradations and the multiple stimuli presentation allows listeners to compare several stimuli at once improving consistency across stimuli (Section 2.6.4). However, the MUSHRA test is targeted at degradations in sound quality relative to a reference. As it could not be known before the test what the most perceptually diffuse stimuli would be, an upper reference could not be used. However, to maintain a good range of diffuseness across all pages, high and low hidden anchors were used to avoid equalisation bias (Zielinski et al., 2007). The high and low anchors were chosen to be stereo and $12/12/13_M$ as these were predicted to be the least and most perceptually diffuse respectively. These stimuli were found to give a good range thus the range of the slider could be used effectively to compare the range of stimuli although the listeners were unaware of the hidden anchors and there was no task to rate one of the stimuli at 100%.

The second difference from standardised MUSHRA comes in the lack of an existing empirical scale. The standard, “Bad, Poor, Fair, Good, Excellent” scale labels are not ideal as they impose a range on the scale that may not fit well to the actual stimuli. For this test, the relative performance of the systems is of most importance and so the labels were replaced by simple, “low” or “high”. This gave a perceived diffuseness rating between 0 and 100.

A screenshot of the user interface implemented in Max 6.1 is shown in figure 4.2. Keyboard shortcuts were provided to quickly swap between stimuli without looking at the laptop screen but the sliders were adjusted using the mouse. ±2 dB of level adjustment was available but not used by any of the listeners.

It was impractical for the listener to compare the on and off-centre listening positions directly on a single trial of MUSHRA. Instead the on and off-centre positions were run on different pages. This way absolute differences in diffuseness across positions is likely not as reliable as the absolute differences between stimuli in the same position.

On-centre there were 25 stimuli plus the 2 hidden anchors. These were randomised and split between 5 pages. Off-centre there were 15 stimuli plus the 2 hidden anchors. These

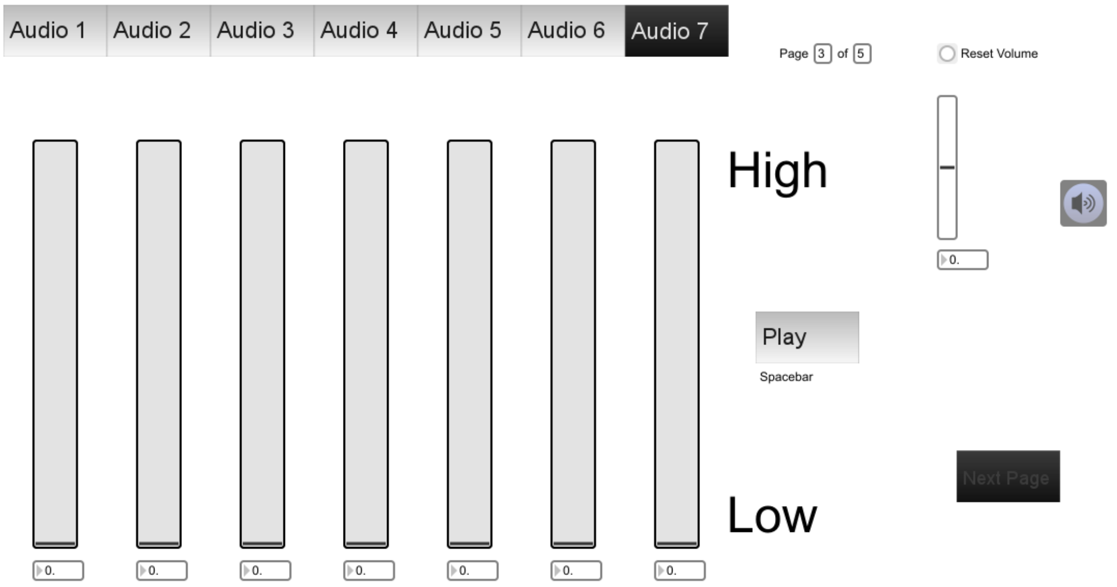


Figure 4.2: MUSHRA-style user interface used in experiment 2 implemented in Max 6.1.

were randomly assigned between 3 pages. This total of 8 pages, each with 7 ratings, were repeated 3 times by each listener taking approximately 45 minutes plus breaks as required. The randomisation was different for each listener ensuring that differences between pages were error rather than bias and therefore allowing comparison between pages.

4.4 Reproduction System

The listening room, reproduction system and calibration is the same as was used in experiment 1 (section 3.4). SPL level alignment using gain, followed by equalisation in 1/6 octave bands (±0.5 dB 95 Hz-20 kHz), and finally time alignment (within 100 μs). In experiment 2 two additional elevated loudspeakers at azimuth ±30° and elevation 26° were also required to allow the arrangement 9.0 to be within tolerances of the standard (ITU-R, 2014b).

4.5 Subjects

The listening test was approved by the ethics and research governance committee (ID: 18709). The experiment was sat by 16 PhD/Masters students at the University of Southampton with no known hearing impairment.

4.6 Results

4.6.1 Post-screening

The listeners were first screened on their rating repeatability. The standard deviation of the three repeats was averaged across all stimuli to give an indication of repeat consistency for each listener. This is plotted in figure 4.3.

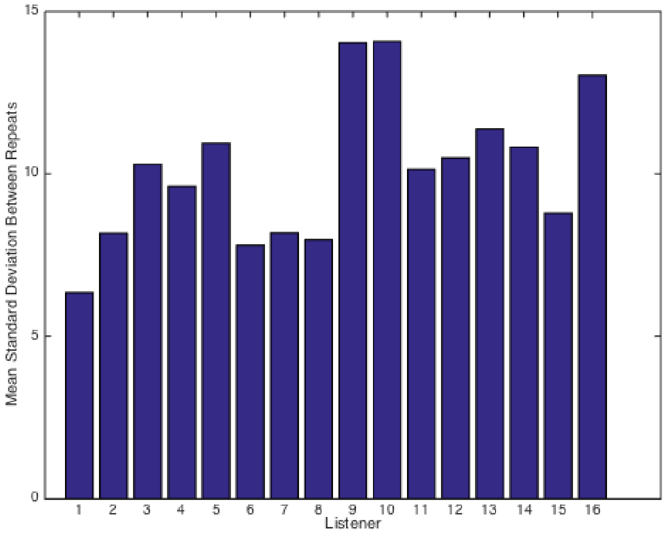


Figure 4.3: Consistency of listeners displayed as mean standard deviation between repeats..

The 3 least consistent listeners (9, 10 and 16) were notably less consistent and so were removed. The 3 repeats of the remaining listeners were averaged together and the resulting data are plotted in the box plot in figure 4.4.

Listeners 4 and 12 both have more than two outliers and so were also excluded from all further analysis. The remaining 11 listeners were the most consistent and congruent and their mean ratings are used in the remaining analysis.

4.6.2 Analysis and Discussion

The means for all the post-screened data are shown in figure 4.5. The following paragraphs look in more detail at specific stimuli to answer the research questions raised in section 4.2.

Visual inspection of the box plots shows the data to be close to normally distributed. Histograms for each stimuli showed no obvious multi modal behaviour. Therefore, a repeated measures ANOVA test was used to test for significant differences as well as significant interactions in cases where all variables are used in all possible combinations. Not

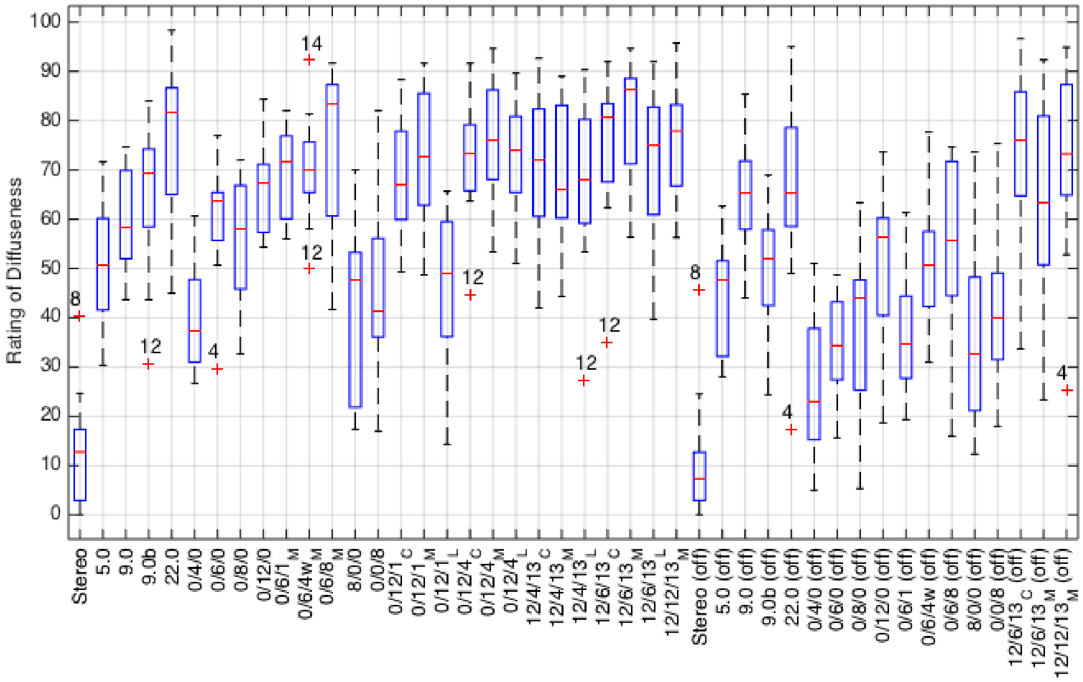


Figure 4.4: Box plot of data with the 3 least consistent listeners removed.

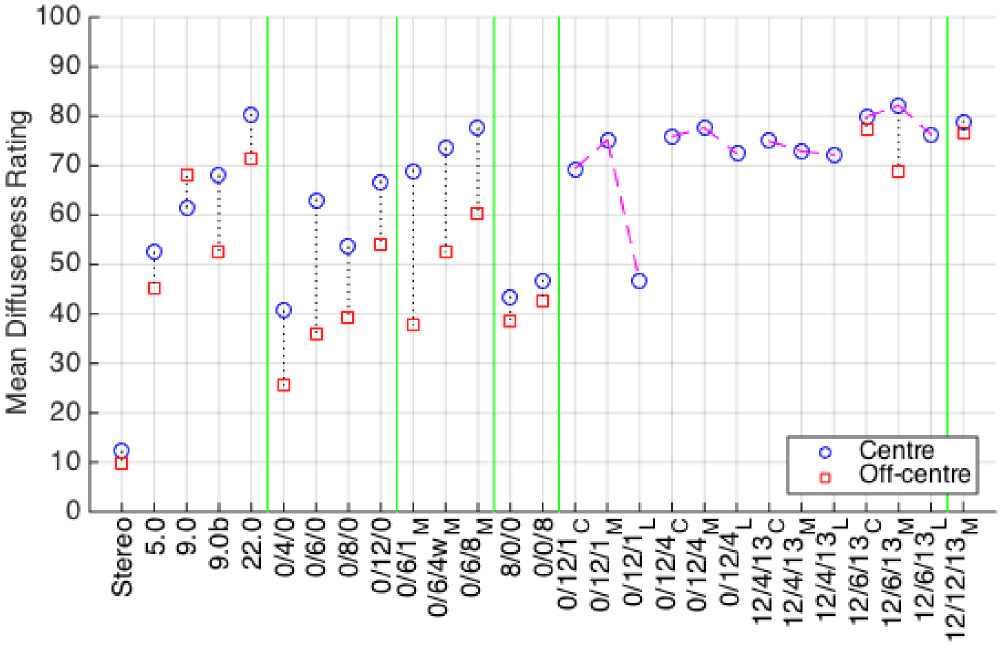


Figure 4.5: Means of all post-screened data, vertical green lines group similar stimuli, dotted black lines indicate the relative performance when moving off-centre, dashed pink lines show the effect of changing the ISLD. Of the stimuli investigating ISLD, only $12/6/13_C$ and $12/6/13_M$ were evaluated off-centre.

all combinations of variables were included in the test so the stimuli were split into groups to investigate the main effects and the interactions between specific variables to answer specific research questions.

Compare Standard Layouts The mean perceived diffusenesses of the standard layouts in both on-centre and off-centre positions are plotted in figure 4.6.

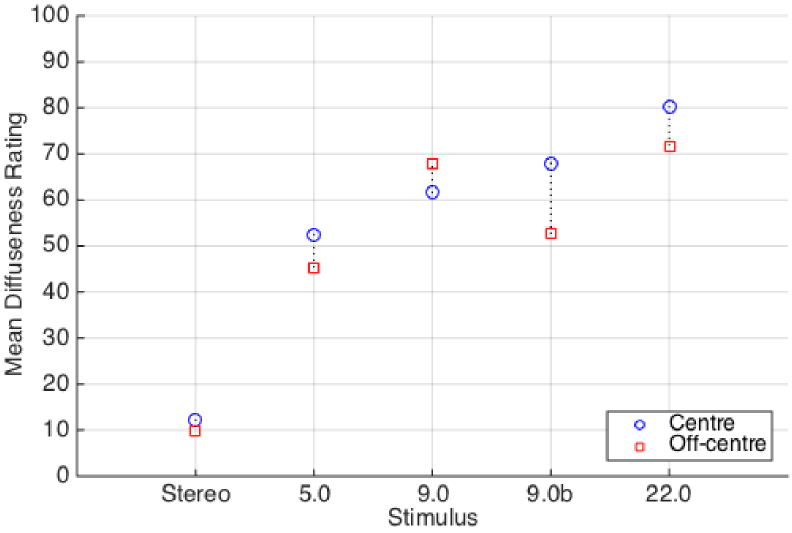


Figure 4.6: Means of post-screened data for standard layouts. Black dotted line indicates the relative performance change when moving off-centre.

These layouts all perform as expected with more loudspeakers providing greater perceived diffuseness. In the on-centre listening position, both 9.0 systems are similarly diffuse. When moving off-centre, 9.0 actually increases in diffuseness although, as stated earlier, the on and off-centre conditions cannot be compared as accurately as differences in the same listing position as they were not directly compared on a single trial of MUSHRA. The 9.0b layout performs a lot worse off-centre which may be for several reasons illustrated in figure 4.7. Firstly, the position of the loudspeakers in the room means that for the same movement distance, the listener is more off-axis from the elevated loudspeakers for the 9.0b stimulus, leading to more colouration. Also the angles of the loudspeakers change more for 9.0b. So for the same movement distance, the relative positions of the loudspeakers to the listener in 9.0b are distorted more with both loudspeakers now on the same side of the listener.

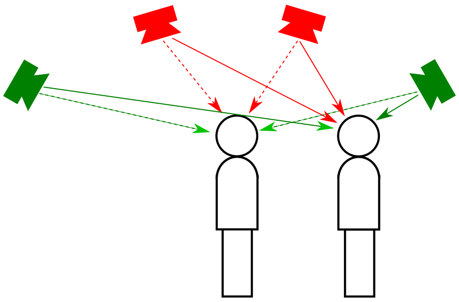


Figure 4.7: Difference between 9.0 and 9.0b when moving off-centre.

A repeated measures ANOVA run on the standard layout stimuli, reveals both the layout and the listening position to be statistically significant ($F(4,7) = 21.603$, $p < 0.0005$ and $F(1,10) = 15.671$, $p = 0.003$ respectively) with no significant interaction ($F(4,7) = 3.200$, $p = 0.086$). Pairwise comparison (table 4.2) shows 22.0 to be significantly more diffuse than all the other layouts. There is no statistical difference between the 9.0b layout and the 5.0 and 9.0 layout. Stereo is significantly less diffuse.

	5.0	9.0	9.0b	22.0
Stereo	**0.000**	**0.000**	**0.000**	**0.000**
5.0		**0.014**	0.078	**0.000**
9.0			1.0	**0.037**
9.0b				**0.038**

Table 4.2: Significance of pairwise comparisons when changing the loudspeaker layout. Significance has Bonferroni adjustment for multiple comparisons.

Number of Head-Height Loudspeakers The stimuli chosen to investigate the number of loudspeakers at head-height are shown in figure 4.8.

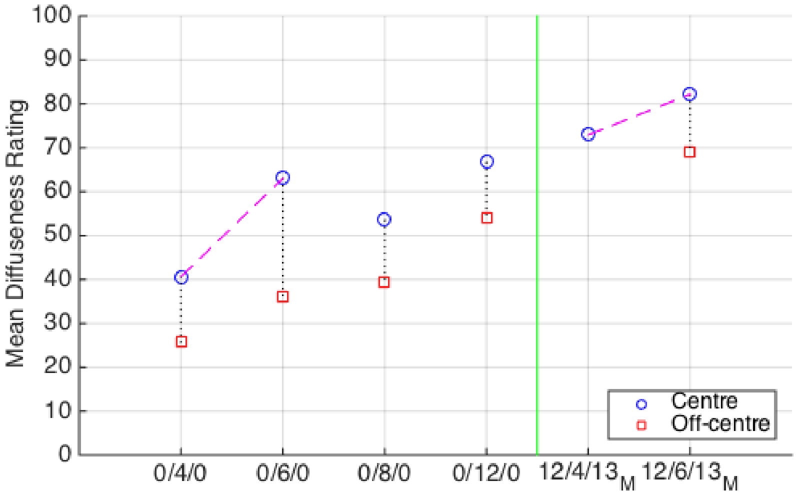


Figure 4.8: Post-screened mean diffuseness for stimuli that show the effect of the number of head-height loudspeakers. The vertical green line separates cases with different numbers of non-head-height loudspeakers. Dashed magenta lines show the effect of increasing the number of head-height channels from 4 to 6. $12/4/13_M$ was not evaluated off-centre.

The 0/12/0, 0/6/0 and 0/4/0 layouts perform as expected with more loudspeakers being perceived as more diffuse. Interestingly, 0/8/0 is less diffuse than 0/6/0 possibly due to the fact it is less uniformly distributed. Were 0/8/0 more evenly distributed it is expected that it would be rated between 0/6/0 and 0/12/0 as was found by Hiyama et al. (2002). Off-centre the 0/6/0 case performs poorly as the listener is very close to the 90° loudspeaker with relatively few other loudspeakers.

With 25 non-head-height loudspeakers, the relative improvement in diffuseness going from 4 to 6 head-height loudspeakers is smaller than when there are no non-head-height loudspeakers.

The effect of the number of head-height loudspeakers and the listener position are both statistically significant ($F(3,8) = 20.806$, $p < 0.0005$ and $F(1,10) = 45.974$, $p < 0.0005$ respectively). There is also a statistically significant interaction ($F(3,8) = 5.099$, $p = 0.029$). Table 4.3 shows the significant differences between the different numbers of head-height loudspeakers. The only stimuli not significantly different are the 0/6/0 and 0/8/0 stimuli.

	0/6/0	0/8/0	0/12/0
0/4/0	**0.001**	**0.036**	**0.000**
0/6/0		1.000	**0.050**
0/8/0			**0.006**

Table 4.3: Pairwise comparison significances (Bonferroni adjusted) between the number of head-height loudspeakers.

Number of Non-Head-Height Loudspeakers The stimuli that show the effect of changing the number of non-head-height loudspeakers are shown in figure 4.9.

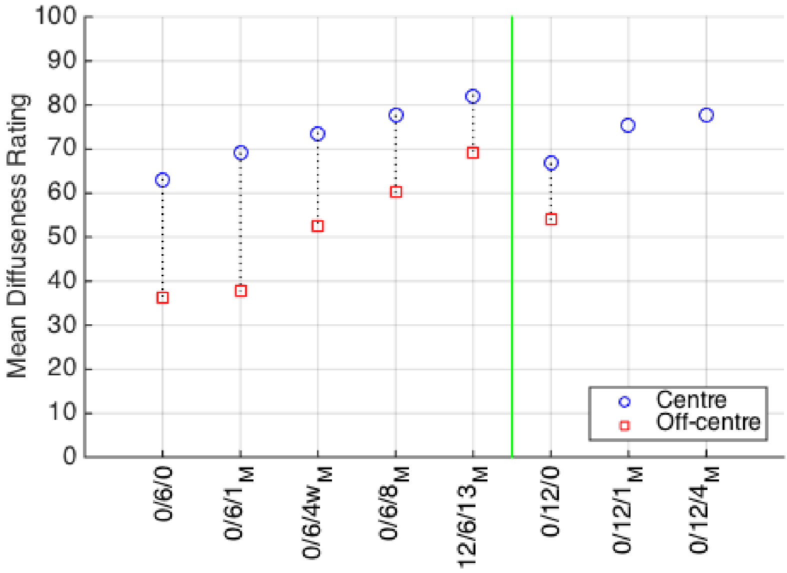


Figure 4.9: Means of post-screened data that shows the effect of changing the number of non-head-height loudspeakers. The green line separates different numbers of head-height loudspeakers. $0/12/1_M$ and $0/12/4_M$ were not evaluated off-centre.

Although greater numbers of non-head-height loudspeakers always increases the perceived diffuseness, there is a case of diminishing returns as the increase in diffuseness is not linearly proportional to the number of non-head-height loudspeakers. Interestingly, the relative improvements in diffuseness performance are very similar for both 6 and 12 head-height loudspeakers despite 12 loudspeakers supposedly being the maximum diffuseness in the 2-dimensional case (Hiyama et al., 2002).

The number of non-head-height loudspeakers and the position of the listener were investigated using a repeated measures ANOVA test. The 5 loudspeaker layouts, 0/6/0, $0/6/1_M$, $0/6/4w_M$, $0/6/8_M$ and $12/6/13_M$ in both on and off-centre locations were analysed. This revealed there to be a statistically significant difference between layouts with different numbers of non-head-height loudspeakers ($F(4,7) = 9.066$, $p = 0.007$) as well as statistically significant differences between the two listening positions ($F(1,10) = 53.913$, $p < 0.0005$). There was no statistically significant interaction ($F(4,7) = 3.106$, $p = 0.091$) between the listener position and the number of non-head-height loudspeakers despite figure 4.9 appearing to show layouts with more non-head-height loudspeakers being degraded less by moving off-centre.

The pairwise comparisons between the stimuli with 6 head-height loudspeakers and different numbers of non-head-height loudspeakers are shown in table 4.4.

	$0/6/1_M$	$0/6/4_M$	$0/6/8_M$	$12/6/13_M$
0/6/0	1.000	**0.014**	**0.009**	**0.000**
$0/6/1_M$		0.090	**0.042**	**0.001**
$0/6/4_M$			0.390	**0.006**
$0/6/8_M$				0.757

Table 4.4: Pairwise significance comparisons (Bonferroni adjusted) between stimuli with different numbers of non-head-height loudspeakers and 6 head-height loudspeakers.

An alternate selection of stimuli for analysis (considering there were no 0/12/8 or $12/12/13_C$ stimuli) are the non-head-height subsets $0/n/0$, $0/n/1_M$ and $0/n/4(w)_M$ with both $n = 6$ and $n = 12$ head-height subsets. In this combination, the arrangement of non-head-height loudspeakers is once again significant ($F(2,9) = 11.328$, $p = 0.003$) but there is also a significant effect of the number of head-height loudspeakers ($F(1,10) = 5.331$, $p = 0.044$). There is no significant interaction ($F(2,9) = 0.348$, $p = 0.715$). The pairwise significances are given in table 4.5 showing adding non-head-height loudspeakers increases the perceived diffuseness significantly, although changing from a single non-head-height loudspeaker to 4 non-head-height loudspeakers is not significantly different.

Positioning of a Layer of Loudspeakers Figure 4.10 shows that on-centre, a head-height layer of loudspeakers is more perceptually diffuse than placing the loudspeakers above or below head-height. The 0/8/0 layout would likely be rated even higher if it were

	$0/n/1_M$	$0/n/4(w)_M$
$0/n/0$	**0.030**	**0.029**
$0/n/1_M$		1.000

Table 4.5: Pairwise significance comparisons (Bonferroni adjusted) between stimuli with different number of non-head-height loudspeakers and either 6 or 12 head-height loudspeakers.

evenly distributed based on the ratings from this experiment for the 0/12/0 and 0/6/0 layouts and from the research of Hiyama et al. (2002).

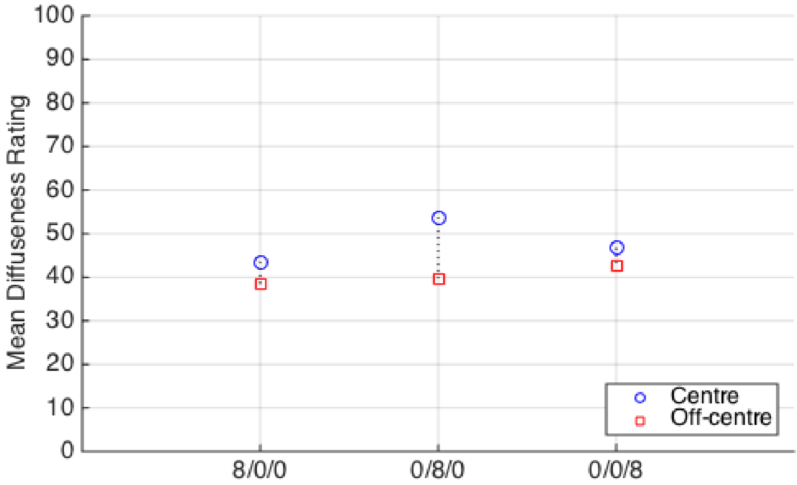


Figure 4.10: Means of post-screened data for stimuli demonstrating changing the height of a horizontal layer of 8 loudspeakers.

However, off-centre, they all perform similarly. This appears to be a combination of 0/8/0 being more diffuse but simultaneously more susceptible to degradation when moving off-centre. The head-height layer is closer to the off-centre listener and the loudspeaker at 90° becomes easier to localise than in the cases where the loudspeakers are above or below the listener with 8/0/0 and 0/0/8 seemingly less degraded by moving off-centre.

A repeated measures ANOVA test found listener position to be statistically significant ($F(1, 10) = 8.300$, $p = 0.016$) but the layer position to not be significant ($F(2, 9) = 0.863$, $p = 0.454$). There was also no significant interaction ($F(2, 9) = 1.142$, $p = 0.361$). Whilst these differences are not significant, it is predicted that the differences are slightly under represented by the non-uniform arrangement of 0/8/0 which was rated lower than the 0/6/0 stimulus despite the results of other authors (Hiyama et al., 2002).

Quantify Increase in Diffuseness Relative to 2-Dimensional Layouts Looking at figure 4.5, diffuseness can be increased by adding non-head-height loudspeakers, even when 12 head-height loudspeakers are used. Even a single “V.O.G.” loudspeaker (at an appropriate level e.g. $ISLD_{Mean}$) increases the perceived diffuseness, although this is not statistically significant (table 4.6). The difference between 0/12/0 and the most diffuse layout (12/6/13) is statistically significant ($p = 0.002$) adding evidence that 3D is more perceptually diffuse than 2D.

	$0/12/1_M$	$12/6/13_M$
0/12/0	0.096	**0.002**
$0/12/1_M$		0.077

Table 4.6: Bonferroni adjusted significance for pairwise comparison between layouts in the on-centre listening position.

Validate Experiment 1 Figure 4.11 shows the layouts tested with a range of ISLD options. The $ISLD_{Mean}$ is, in general, rated higher than the equal layer level and equal channel level conditions.

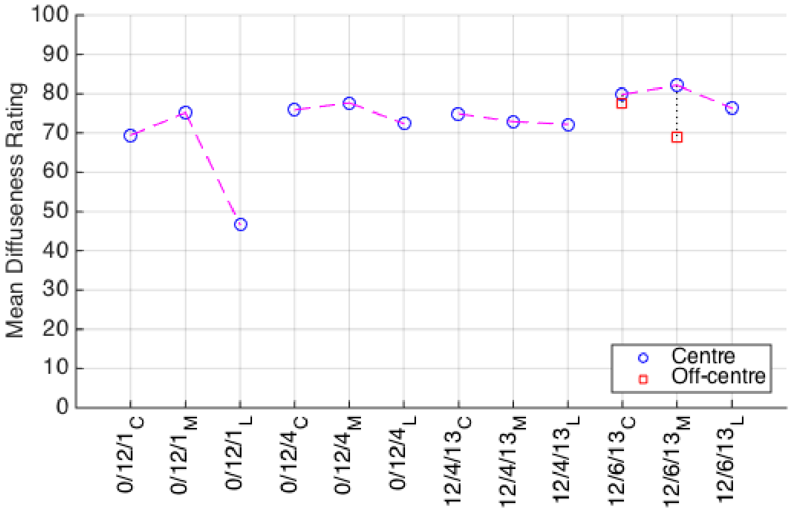


Figure 4.11: Means of post-screened data displaying the stimuli that test the effect of ISLD. Layouts that are the same are connected by dashed magenta lines. Only $12/6/13_C$ and $12/6/13_M$ were evaluated off-centre.

This variation is low in comparison to the variation between loudspeaker layouts. From the stimuli tested, the ISCLD=0 is a reasonable choice of ISLD. The particular case tested off-centre shows the optimised ISLD to be worse off-centre than maintaining ISCLD=0. With many more non-head-height loudspeakers than head-height loudspeakers, ISCLD=0 gives a low total level of the head-height layer. The optimised ISLD increases the level of the head-height layer and because the head-height layer has only 6 loudspeakers, the loudspeaker at 90° is close and loud when the listener is off-centre. However, the decrease in diffuse performance is still less than for other stimuli with 6 head-height layers but less non-head-height channels (e.g. $0/6/8_M$). A repeated measures ANOVA test was conducted on the factors of ISLD and loudspeaker layout to determine significant factors. Both the ISLD and the loudspeaker layout are statistically significant ($F(2, 9) = 32.517$, $p < 0.0005$ and $F(3, 8) = 19.282$, $p = 0.001$ respectively) as well as the interaction between them ($p = 0.005$).

Pairwise comparison (table 4.7) shows ISCLD=0 and ISLD=$ISLD_{Mean}$ to not be statistically different despite $ISLD_{Mean}$ stimuli appearing to be more rated slightly higher than the ISCLD=0 cases.

	ISLD=$ISLD_{Mean}$	ISLD=0
ISCLD=0	**1.000**	0.001
ISLD=$ISLD_{Mean}$		0.000

Table 4.7: Pairwise significance of ISLD with Bonferroni adjustment for multiple comparisons.

Although the $ISLD_{Mean}$appears to usually be rated slightly higher than the ISCLD=0 cases, this is not statistically significant. ISCLD=0, i.e. all loudspeakers the same level, appears to be a reasonable choice of ISLD for all 3D layouts.

4.7 Comparison with Objective Metrics

In this section different methods of measuring the diffuseness of a sound field are compared to the perceived diffuseness of the stimuli. In reverberation rooms, metrics of diffuseness are designed to validate the sound field is sufficiently diffuse to allow measurements of absorption, microphone directivity index or total power output of sound sources to be accurate (Cook et al., 1955).

In concert hall acoustics and reproduced sound, the sound field is not used for measurements and the advantages of a more diffuse sound field are the improvements in the perceived diffuseness relating to ASW and LEV.

Many of the metrics in this section were developed for concert hall acoustics or reverberation rooms. However the sound fields generated in rooms are different from those in reproduced audio although they might be perceived as equally diffuse. In concert halls the sound field comprises of the direct sound some early reflections and many late reflections (figure 4.12). All reflections are derived from the same source and so are not uncorrelated. Early reflections add up to give an interference pattern. Late reflections are more numerous and come from more directions, add up as if incoherent and can be considered approximately diffuse. Alternatively, in reproduced audio, the sound field is composed of a finite number of truly uncorrelated components. Although the number of loudspeakers is low in comparison to the number of reflections in a concert hall, the sources can be completely uncorrelated and so always add incoherently. These are two examples of approximately diffuse sound fields. The difficulty in measuring approximate diffuseness is that a sound field is either diffuse (both homogeneous and isotropic) or not diffuse (either not homogeneous and/or not isotropic) (Nélisse and Nicolas, 1997) and "diffuseness" is not defined. The most useful objective metric of diffuseness in these listening tests is one that correlates well with the perceived diffuseness. The differences between these two types of partially diffuse sound field are used throughout this section to explain some of the differences between the measured diffuseness and the perceived diffuseness. Particularly differences that appear in reproduced audio that may not appear when the equivalent metric is used in architectural acoustics.

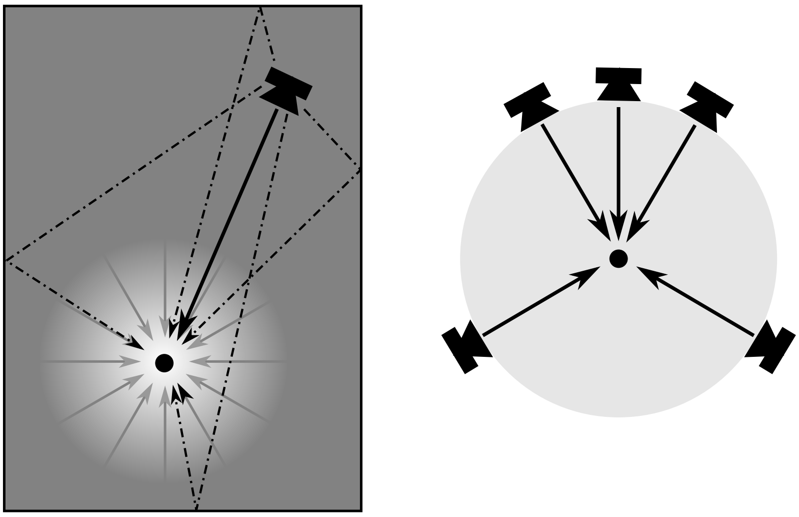


Figure 4.12: Difference in sound field of architectural acoustics (left) and reproduced sound (right).

The objective metrics fall broadly into three categories that relate to the reproduced sound field signal chain described in section 1.2.

1. Firstly, simple parameters of the reproduction system, in this case the number of loudspeakers, can be correlated with the perceived diffuseness.

2. Secondly measurements of the sound field using arrays of microphones can be related to the perceived diffuseness. These metrics test the sound field against the theoretical diffuse sound field (section 2.3) in terms of either homogeneity, isotropy or relationship between spatially separated measurement positions such as correlation and coherence.

3. Finally binaural recordings of the stimuli can be processed to derive psychoacoustically motivated predictions of the perceived diffuseness and compared to the elicited perceived diffuseness.

These second two methods are of particular interest as they are more likely to relate to the underlying psychoacoustic behaviour that determines diffuseness perception. Binaural is especially convenient as it is only a pair of signals.

Objective quantities of the sound field or listening system are developed from the literature reviewed in section 2.4 and are plotted against the means of the data to see which metrics correlate well. The adjusted r-squared value between a line of fit and the means of the data is used as an indication of the accuracy of the metric.

4.7.1 Number of Loudspeakers

The first metric is the number of loudspeakers. The perceived diffuseness is plotted against number of loudspeakers in figure 4.13.

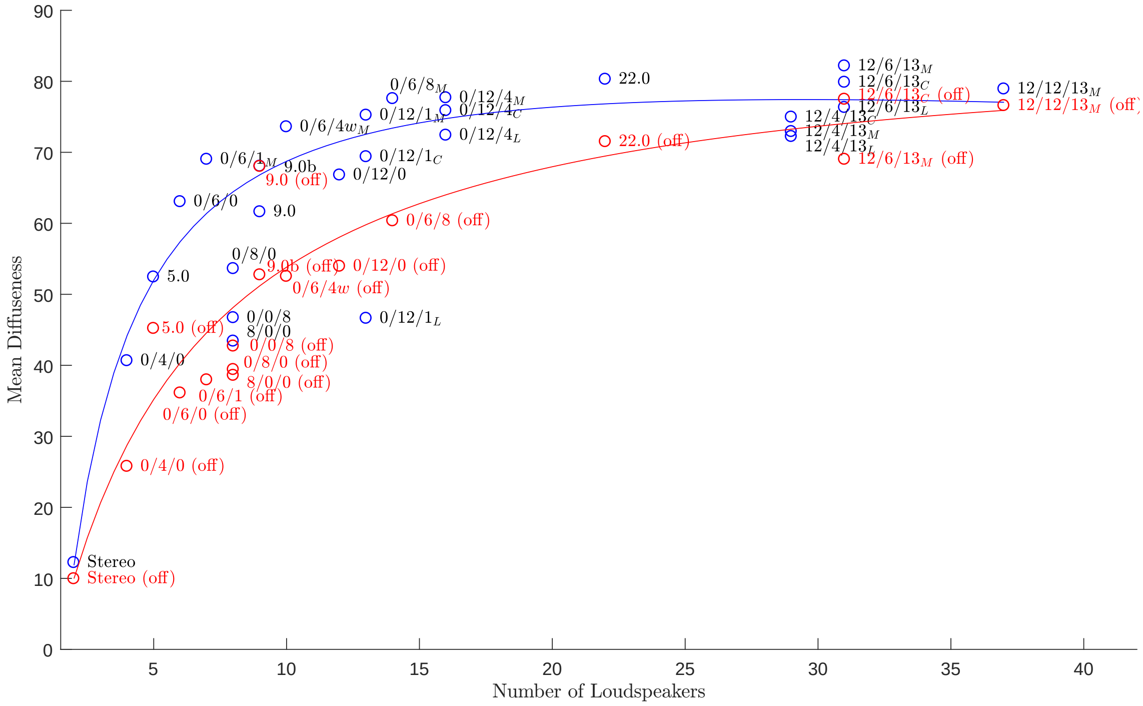


Figure 4.13: Diffuseness plotted against the number of loudspeakers. Blue points are on-centre, red points are for the off-centre listening position. Includes lines of best fit for on and off-centre listening positions. Layouts 0/0/8, 8/0/0 and $0/12/1_L$ are excluded when fitting curves.

The layouts (8/0/0, 0/0/8 and $0/12/1_L$) are most biased to loudspeakers not at head-height. They are notably less diffuse for the number of loudspeakers than layouts with more head-height loudspeakers. For this reason these layouts were excluded from the line fitting process. A pair of lines of best fit of the form,

$$f(x) = \frac{p_1 x^2 + p_2 x + p_3}{x + q_1} \tag{4.1}$$

for both the on-centre and off-centre positions give an indication of the number of channels of arbitrary layout required for a given diffuseness. A tentative suggestion for the required number of loudspeakers to produce a perceptually diffuse sound field are 10 loudspeakers for the on-centre listening position or around 20 for the off-centre listening position.

Possibly most interesting are the layouts that give a high diffuseness for the number of loudspeakers. Layouts such as $0/6/1_M$ on-centre and 9.0 off-centre which are very close to the maximum diffuseness for that listening position but with many fewer loudspeakers. Unfortunately, both of these perform poorly in the other listening position (below the curves) and so neither can be said to be definitively better than any other system.

The number of loudspeakers is a simple metric and it ignores factors such as the position of the loudspeakers with $0/12/1_C$, $0/12/1_M$ and $0/12/1_L$ estimated equally despite large perceptual differences. Measurements of the sound field will take these differences into account and therefore might lead to metrics that are more generally applicable.

4.7.2 Sound Pressure Level Uniformity

The sound pressure level uniformity is a measure of homogeneity. The SPL was measured in a horizontal 5×5 array measuring 2 m×2 m centred at the central listening position. A B&K free-field microphone type 4190 with B&K preamplifier type 2669 at head-height and pointed vertically upwards was moved around the room and recorded all the stimuli in all grid positions. The sound pressure level was calculated for the full bandwidth (unweighted) and also in octave bands.

The technique used previously by (Veit and Sander, 1985) involved specifying the area in which the sound field was diffuse by specifying the distance to the nearest loudspeaker before the sound pressure level began to change. This is less appropriate for the stimuli tested here as the arrangement of loudspeakers changes and so the nearest loudspeaker changes.

The non-uniformity of sound pressure levels is given as the standard deviation of the unweighted sound pressure levels across the array and is plotted against the perceived diffuseness in figure 4.14.

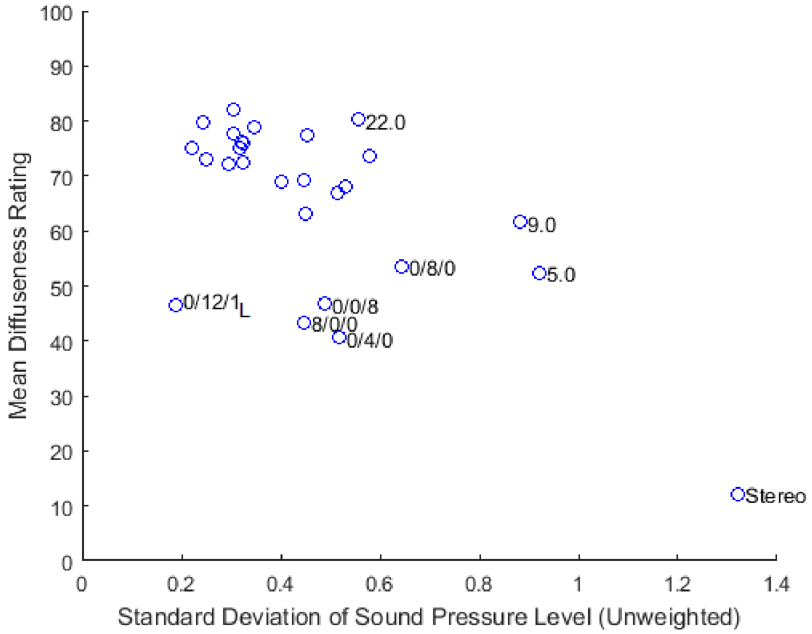


Figure 4.14: Standard deviation of the unweighted sound pressure level plotted against the mean rating for diffuseness.

We would expect a negative correlation with a smaller standard deviation in SPL associated with high homogeneity and therefore a more diffuse sound field perceptually. However, several stimuli do not follow this trend. The stimuli 0/0/8 and 8/0/0 both do not have any head-height loudspeakers and have a lower standard deviation of SPL than their head-height counterpart 0/8/0 despite being perceived less diffuse. The stimulus $0/12/1_L$ has a single loudspeaker above the listening position that is at a very loud level relative to any single head-height loudspeaker. In these cases a large proportion of the measured SPL comes from loudspeakers far from the 2D head-height microphone array. The layout 0/4/0 also has loudspeakers that are generally further from the array because the array is square whereas the room and loudspeaker array was rectangular. All these stimuli appear overestimated whereas the layouts of stereo, 5.0, 9.0 and 22.0 are all underestimated. These standardised layouts have more loudspeakers concentrated at the front than at the rear increasing the SPL variation despite the perceived diffuseness being seemingly unaffected by the unevenness of the loudspeaker array.

The adjusted r-squared values for lines fitted to the data in each octave band show the extreme high and low octave bands to correlate poorly with the subjective data (table 4.8). At low frequencies the modal response of the room is likely to play a part in biasing the objective measure. At high frequencies, the directivity of the microphone and the loudspeakers may bias the measure because the microphone moves off-axis of the loudspeakers in some stimuli more than others. The microphone was also always pointed vertically up-

Centre Frequency	Adjusted R-Squared
62.5	0.2002
125	0.0005
250	**0.6075**
500	**0.1695**
1000	**0.5437**
2000	**0.6098**
4000	**0.5769**
8000	**0.4660**
16000	-0.0124

Table 4.8: Adjusted R-Squared for lines fitted to the standard deviation of SPL in octave bands.

ward and becoming less omnidirectional at higher frequencies thus biasing the result for layouts with loudspeakers above head-height or below head-height.

Instead of using the unweighted SPL, the standard deviation of the SPL of octave bands from 250 Hz to 8 kHz were averaged and plotted in figure 4.15 against the mean diffuseness rating.

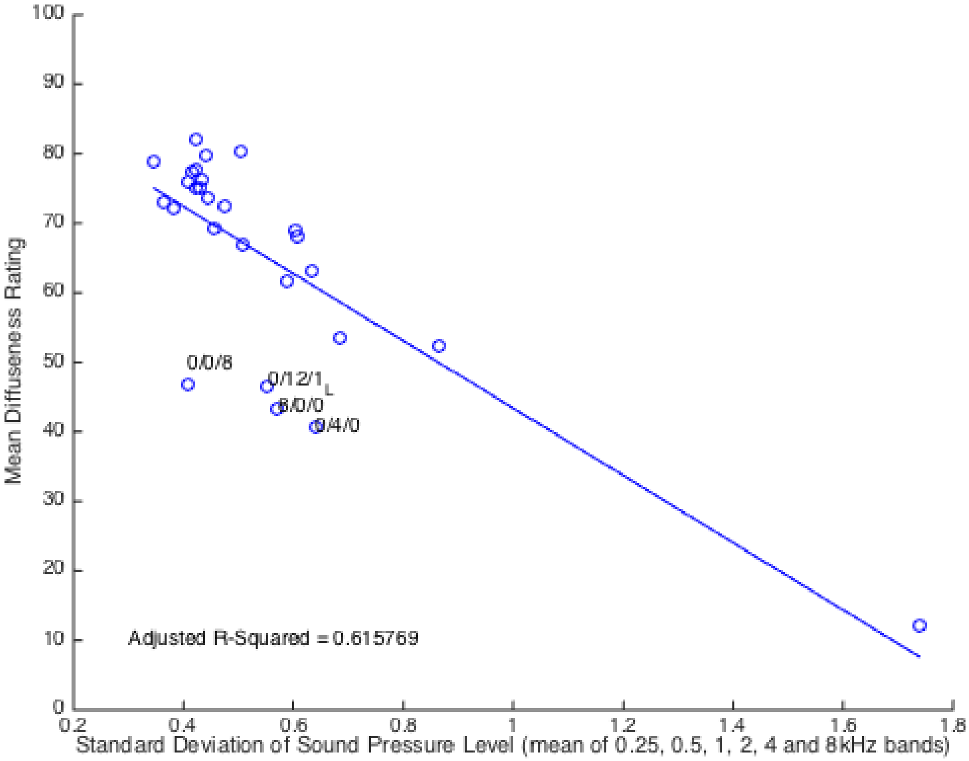


Figure 4.15: Standard deviation of the sound pressure level averaged across 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz and 8 kHz octave bands, plotted against the mean diffuseness ratings for all on-centre stimuli. A linear regression model is also plotted.

With an adjusted R-Squared of 0.6157969 the model is a fairly poor fit to the means of the data although (with the exception of the 4 unusual stimuli) the metric is able to differentiate between high and low perceived diffuseness.

The metric is also not based psychoacoustically. It can be assumed that the listener cannot perceive level variations outside the listening area so measuring the SPL over a 2 m×2 m grid is unlikely to address the psychoacoustic cause of diffuseness perception. A subset of the grid covering 1 m×1 m was also investigated but this gave very little contrast between the stimuli and fit the data poorly. The metric is also not usable off-centre. Most of the room is included in the 2 m×2 m sampling of the room and the room is not sufficiently big enough to centre the array at the off-centre listening position.

4.7.3 Sound Intensity

In a diffuse sound field the time averaged intensity vector must equal zero (section 2.4.2).

The diffuseness estimation $\psi(t, f)$ used in DirAC, uses B-format signals to estimate the time averaged intensity and time averaged energy density for each time window and frequency bin in the Short Term Fourier Transform (STFT) domain (Ahonen et al., 2008)(section 2.4.2). An A-format Tetramic was used to record the B-format signals calculate the diffuseness estimation $\psi(t, f)$ described in section 2.4.2. The diffuseness estimation was averaged across frequency bins and windows to give a single diffuseness value for each stimulus and these values are plotted in figure 4.16 versus the perceived diffuseness.

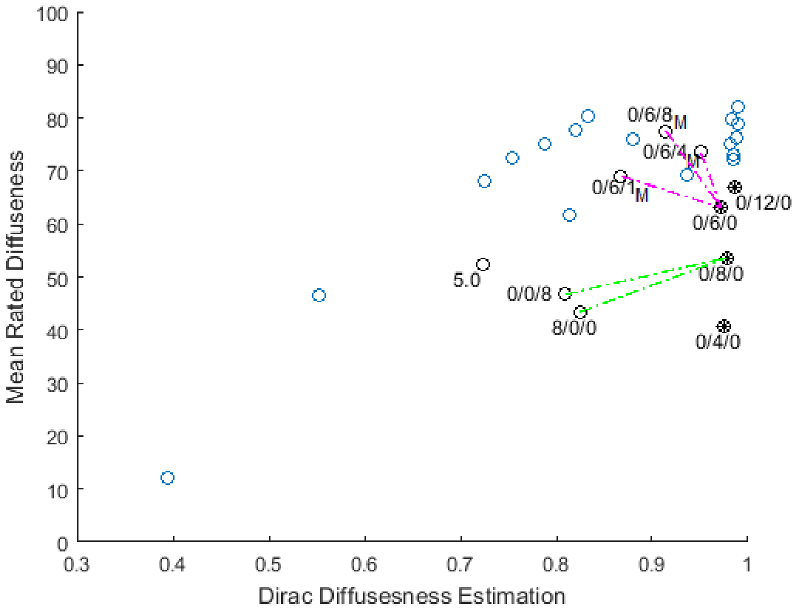


Figure 4.16: Diffuseness estimation for all stimuli based on DirAc. Pink lines show the effect of adding non-head-height loudspeakers to 6 head-height loudspeakers. Green lines show stimuli with 8 loudspeakers in a single layer.

The two dimensional layouts ($0/n/0$), all have a diffuseness close to 1 because the loudspeakers are positioned opposite the measurement position. With equal energy coming from opposite directions the net flow of energy is zero and this is independent of the number of loudspeakers. This means the layout 0/4/0 is predicted as diffuse as 0/12/0 despite a large perceptual difference.

The magenta lines shows the effect of adding elevated loudspeakers to a 2D layout. There are more loudspeakers and the array is now three dimensional increasing the perceived diffuseness (section 4.6.2) but more energy is coming from above the measurement position leading to some intensity in the Z dimension. This gives a lower diffuseness estimation despite being perceived as more diffuse. Whilst this might imply that some intensity in the Z dimension increases the perceived diffuseness, we also see that layouts such as 0/0/8 and 8/0/0, are rated as less diffuse than 0/8/0 (green lines) despite having a greater absolute intensity vector due to changing the height of the 2D 0/8/0. When comparing 0/4/0 to 5.0 we see this same bias but not in the Z dimension. The layout 5.0 has an additional loudspeaker which increases the perceived diffuseness but also increases the intensity as more energy comes from the front than from the back.

Calculating the diffuseness in octave bands was found not to improve the estimation accuracy.

These results show the intensity to be related to the uniformity of the arrangement of loudspeakers but this does not seem to correlate with the perceived diffuseness.

4.7.4 Interaural Cross-correlation Coefficient (IACC)

The InterAural Cross-correlation Coefficient (IACC) is a perceptually motivated measure of the diffuseness. It is a measure of the similarity between the two ear signals and is given by the maximum value of the normalised InterAural Cross-correlation Function (IACF) over the possible range of ITDs (equations 4.2 and 4.3) (ISO, 2009).

$$IACC_{t1,\,t2} = max|IACF_{t_1,t_2}| \tag{4.2}$$

for -1 ms< τ <+1 ms, where,

$$IACF_{t_1,\,t_2}(\tau) = \frac{\int\limits_{t_1}^{t_2} p_l . p_r(t+\tau)dt}{\sqrt{\int\limits_{t_1}^{t_2} p_l^2(t)dt \int\limits_{t_1}^{t_2} p_r^2(t)dt}} \tag{4.3}$$

and p_l and p_r are the impulse responses for the right and left ears respectively. It is usually computed in octave bands and performed on room impulse responses. Early reflections ($t_1 = 0$, $t_2 = 80$ ms) fuse with the direct sound and the late diffuse reflections ($t_1 = 80$ ms, t_2 <RT60) are perceived as part of the environment. Early and late IACC correlate therefore correlate with apparent source width (ASW) and listener envelopment (LEV) respectively. The IACC$_3$ of Hidaka et al. (1995) uses the mean IACC from the middle frequency bands 500 Hz, 1 kHz and 2 kHz to give a single value for the IACC$_3$. In reproduced sound there is no difference in the amount of diffuseness between the early and late reflections and so the IACC can be calculated using steady state full-bandwidth noise dummy head recordings. The measured IACC$_3$ is plotted versus the perceived diffuseness in figure 4.17.

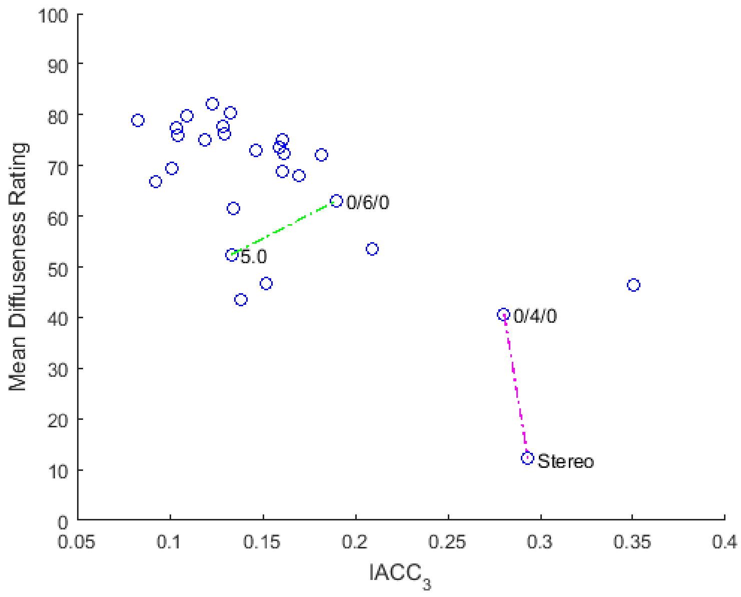


Figure 4.17: IACC$_3$ of Hidaka et al. (Hidaka et al., 1995) versus perceived diffuseness. The pink line highlights the difference between stereo and 0/4/0 which are rotationally symmetrical about the interaural axis. The green line compares 5.0 and 0/6/0 which highlight the differences between the number of cones of confusion (5 and 4 respectively) against the number of loudspeakers (5 and 6 respectively).

Interestingly, stereo and 0/4/0 have near identical IACC$_3$ values. The dummy head is static and in both cases, the sources lie in the same two cones of confusion at 60° and 120° from the interaural axis. However, listeners are able to separate loudspeakers in the same cone of confusion using head rotation. The same can be seen in 0/6/0 and 5.0. The layout 5.0 has the loudspeakers in the front at different angles to the loudspeakers in the rear and therefore has 5 cones of confusion (relating to the ITDs for each of the individual loudspeakers). The layout 0/6/0 has loudspeakers symmetrical about the interaural axis and therefore has 4 cones of confusion (as the ITDs relating to the front loudspeakers are the same as those in the rear). Therefore 5.0 has fewer loudspeakers but more cones of confusion leading to a lower IACC where 0/6/0 has more loudspeakers is perceived more diffuse and yet the IACC estimate is lower due to the fewer cones of confusion.

The quality of the fit was no higher in any of the individual octave bands.

4.7.5 Spatial Correlation

The narrow band cross-correlation function and coherence function between the signals at two points in a diffuse sound field is dependent only on the distance between them. The distance between the microphone capsules was chosen to be 20 cm. This is similar to the distance between the ears. Despite the measure not being an interaural measure, it makes sense that listeners cannot perceive the sound field outside the listening area as the listeners were asked not to move so it is logical to approximately measure the listening area. Unlike the interaural cross-correlation (which includes the effect of the head and torso), the spatial correlation function can be found analytically. One microphone was placed at the centre of the listening position and another 20 cm to the right. The maximum value of the normalised cross-correlation function is used in a similar manner to the calculation of IACC and is plotted against the mean perceived diffuseness in figure 4.18.

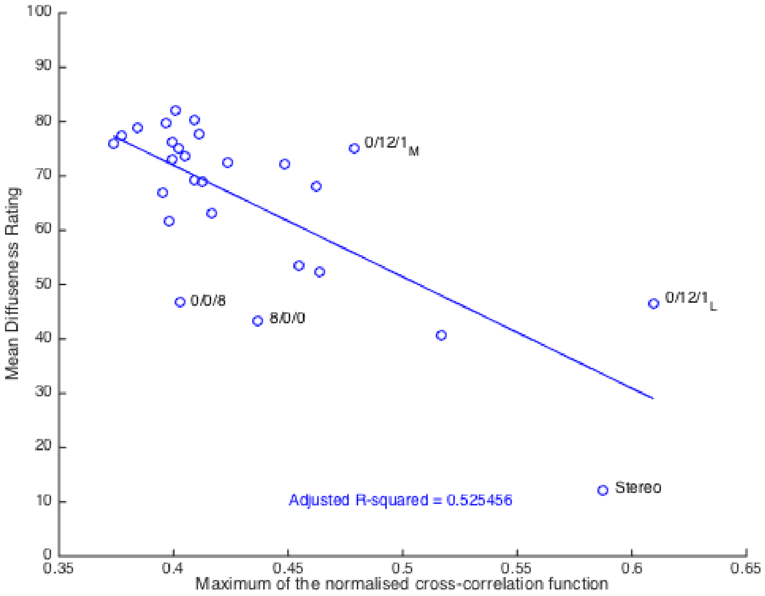


Figure 4.18: Maximum of the normalised cross-correlation function between 2 microphones 20 cm apart with a max lag of ± 1 ms.

The cross-correlation function treats all frequencies equally and is dependent on the frequency content of the two signals. The logarithmic nature human hearing frequency perception means that the cross-correlation is therefore likely to be biased towards HF. To avoid bias towards the high frequencies the correlation coefficient was calculated in octave bands. The adjusted r-squared values for lines of best fit to the data in separate octave bands is shown in table 4.9.

Centre Frequency	Adjusted R-Squared of linear fit
62.5	-0.0126
125	-0.0074
250	0.3115
500	0.2369
1000	**0.5359**
2000	**0.5376**
4000	0.1333
8000	0.3450
16000	0.2448

Table 4.9: Accuracy of linear regression model in different octave bands.

The lowest and highest frequency octave bands did not discriminate well between stimuli. The maximum of the normalised cross-correlation functions for the octave bands at 1 kHz and 2 kHz were averaged and the result is plotted against the mean diffuseness rating in figure 4.19.

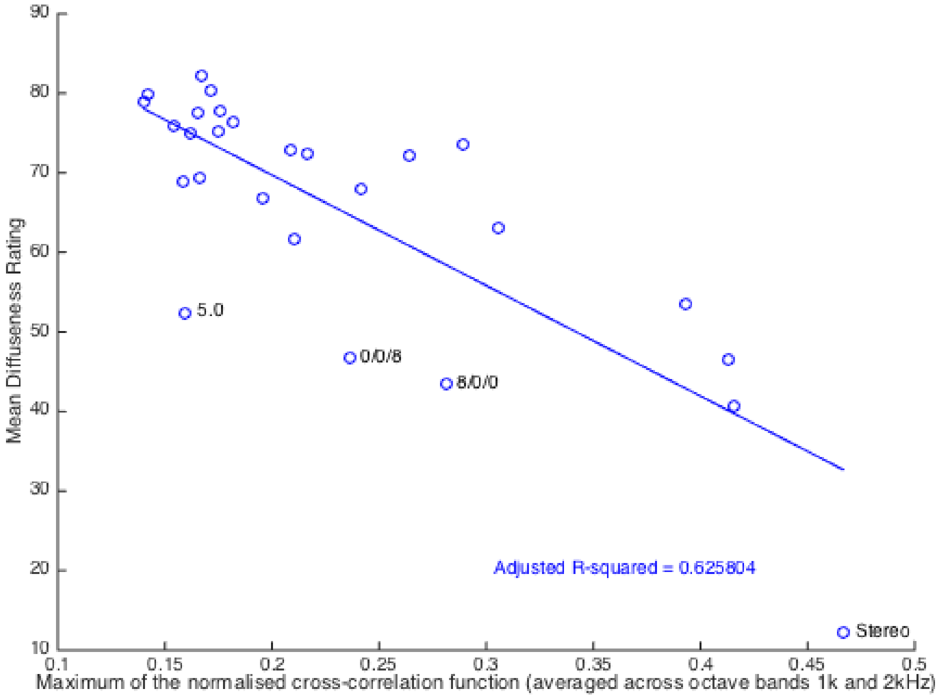


Figure 4.19: Correlation coefficient averaged across the 1 and 2 kHz octave bands.

A notable bias of the correlation, as was the same for the IACC, is that sources in the same cone of confusion are essentially summed. As with the IACC the stereo and 0/4/0 are poorly separated. Both have loudspeakers at ± 30° but 0/4/0 also has loudspeakers at ± 150°. The correlation functions of both these stimuli are the same with two peaks one at + and - the time difference of the +30° loudspeaker (figure 4.20).

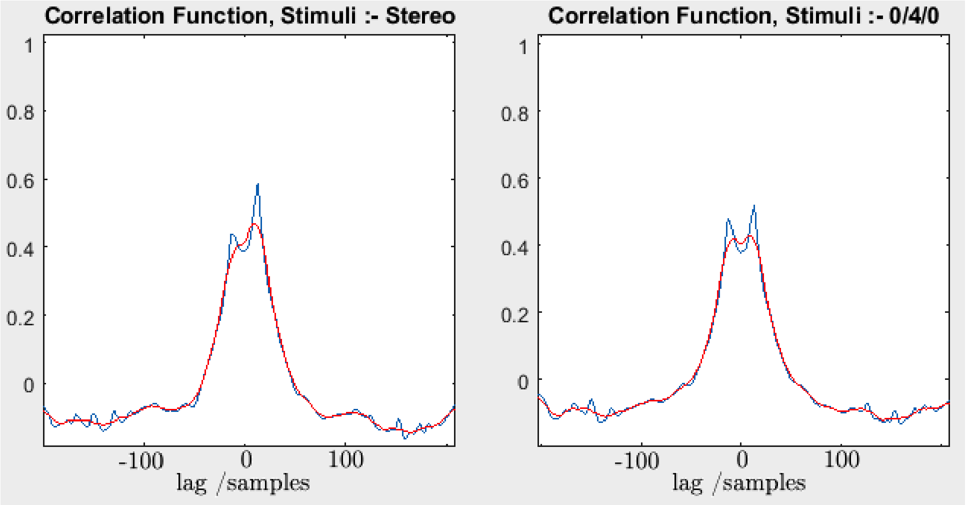


Figure 4.20: Normalised cross-correlation functions for stereo (left) and 0/4/0 (right). The red lines are smoothed.

A suggested adaptation was to recalculate the cross-correlation function in the front to back dimension (the y-dimension). This would identify the differences between stereo and 0/4/0. For thoroughness, it was also decided to use the z-dimension as this would identify differences between stimuli with and without non-head-height loudspeakers. Therefore, the maximum of the normalised cross-correlation function was calculated in the x, y and z dimensions using an array of microphones. One centre microphone, one microphone 20 cm to the right (for the x-dimension) one 20 cm forwards of the centre microphone (for the y-dimension) and one microphone 20 cm above the centre microphone (for the z-dimension). These maximum values of the normalised cross-correlation function, x, y and z were calculated in octave bands and then the 1 kHz and 2 kHz values were averaged and plotted in figure 4.21.

This shows a notable improvement in the estimation whilst still not perfect. It is not clear why the 0/6/0 stimuli is underestimated although the stereo, 8/0/0 and 0/0/8 stimuli could be biasing the line of best fit. This dimension averaging could likely be improved further by weighting the x, y and z components in a more perceptually appropriate way.

4.7.6 Spatial Coherence

The spatial coherence also uses the separated microphones used for the spatial correlation, one microphone on-centre and one 20 cm to the right. The coherence was calculated in Matlab using a window length of 2^{16}. This coherence function was then averaged across

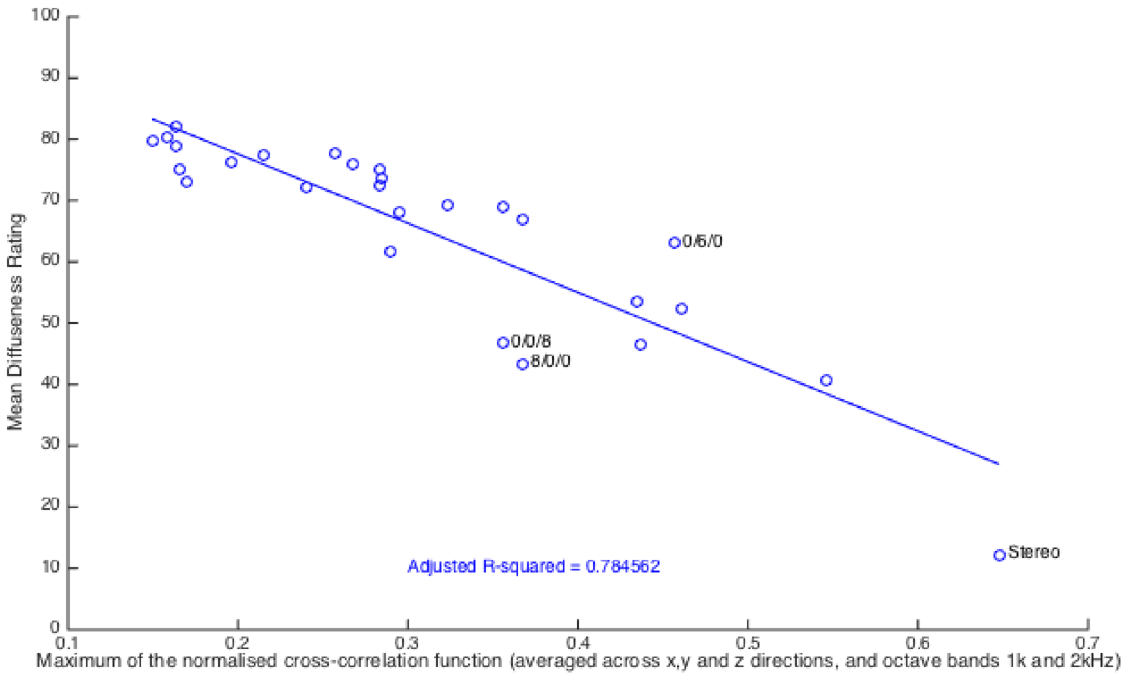


Figure 4.21: Correlation coefficient averaged across 1 and 2 kHz octave bands and across x, y and z directions.

the entire range of frequencies to give a single coherence value for each stimulus. A long FFT was used as the reverberation in the room will tend to increase the coherence and using a window longer than the reverberation time should give more repeatable results. This mean coherence can be seen plotted against mean diffuseness rating in figure 4.22.

The coherence is different to the cross-correlation in that the coherence is independent on the frequency content of the source signal. The result is the same for pink or for white noise. This makes it analytically easy to predict but may be biased by the lack of information about the source. Calculating the correlation in octave bands will have a similar effect on ignoring the frequency content of the two signals. Octave band averaging of the coherence is only used to avoid influencing the result more by high frequencies where there are more frequency bins per octave than low frequencies. The coherence was calculated in octave bands and only the octave bands that discriminated the stimuli well were averaged together. In this case this was the octave bands from 500 Hz-8 kHz (table 4.10).

The resultant mean coherence averaged across the given frequency bands is plotted in figure 4.23.

The coherence suffers from some of the same biases as the IACC (inability to separate sources in the same cone of confusion) and so the coherence was calculated in the x, y, and z directions and then these values were averaged. These mean coherence values averaged

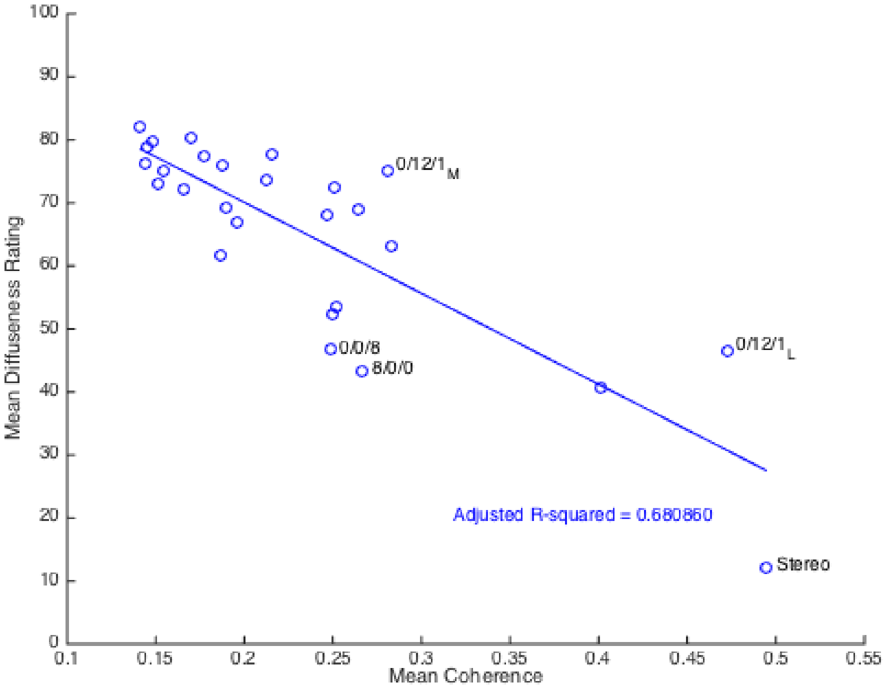


Figure 4.22: Mean Coherence plotted against stimulus with associated linear regression line (on-centre only).

Lower Frequency/Hz	Centre Frequency/Hz	Upper Frequency/Hz	Adjusted R-Squared of linear fit
44	63	88	-0.0355
88	125	177	0.1783
177	250	335	0.2865
335	500	710	**0.5596**
710	1000	1420	**0.7596**
1420	2000	2840	**0.7886**
2840	4000	5680	**0.5735**
5680	8000	11360	**0.7524**
11360	16000	22720	0.5548

Table 4.10: Adjusted r-squared values for linear model fitting the data from the mean coherence in different octave bands.

in octave bands and in x, y and z dimensions are plotted against the perceived diffuseness plotted in figure 4.24.

It is worth emphasising the mean coherence values for the x, y and z dimensions are averaged and not the spectra of the two signals for each position. Averaging of the spectra has been shown to cause even ordinary rooms to approach the coherence function for a theoretical diffuse sound field (Jacobsen and Roisin, 2000). Therefore the averaging must be done after the coherence calculation. Averaging in the x, y, and z dimensions is a holistic method of separating the stereo and 0/4/0 stimuli that have the same number of cones of confusion.

The coherence also appears to work off-centre where the cross-correlation does not. A separate regression line is plotted in figure 4.24 for the off-centre position as mentioned

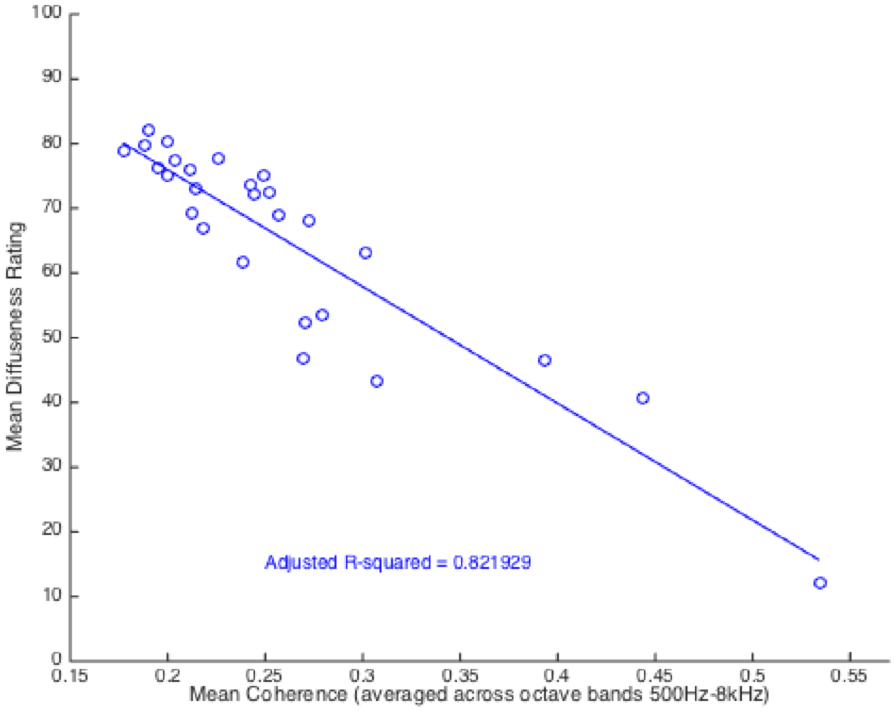


Figure 4.23: Average mean coherence across octave bands 500 Hz-8 kHz.

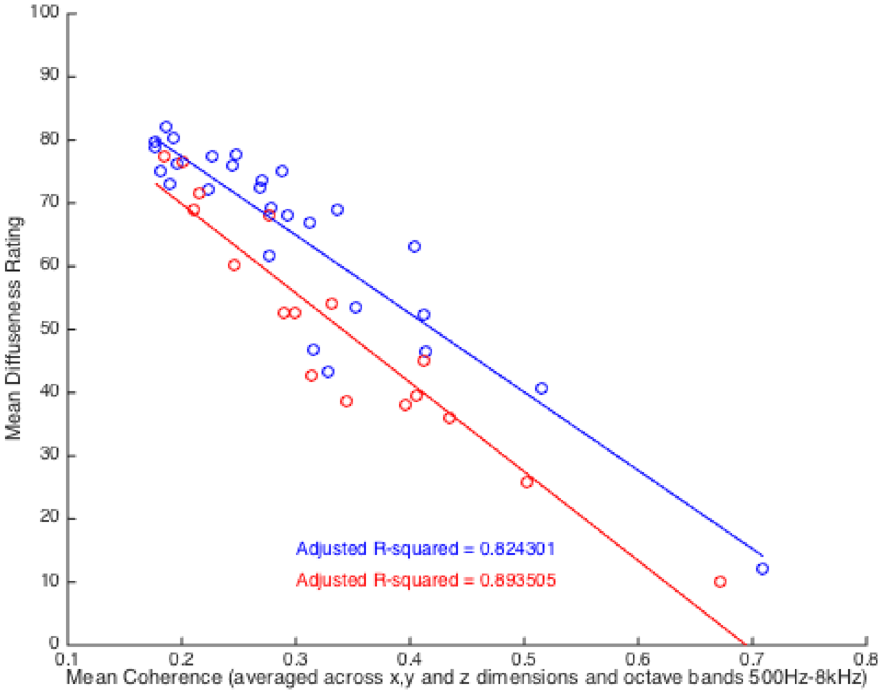


Figure 4.24: Mean coherence averaged across octave bands 500 Hz-8 kHz and dimensions x, y and z.

before, the two positions were evaluated separately so direct comparisons between positions are unlikely to be as accurate as comparisons in the same position.

Of all the linear regression metrics so far this has the highest adjusted R-squared (0.8243). There is still room for further development including finding a more optimal way to combine the x, y and z components. This metric also works across a wider frequency range than many of the metrics mentioned earlier implying greater robustness.

An analytical model of the coherence has also been found to match well to the measured data. In section 2.4.3 a plane wave model was described for measuring the narrow band cross-correlation coefficient between two points. This was then integrated for all directions to find the cross-correlation function and the coherence function $\gamma_{xy}(f)$ between two points for a theoretical diffuse sound field. If the infinite integral is replaced by a finite sum, then the coherence between points for an arbitrary arrangement of plane waves can be found. As the listener is fairly far from the loudspeakers and there is not much reverberation, the analytical model appears to fit fairly well and is given as follows,

$$\gamma_{xy}^2(f) = \left(\frac{1}{\sum A}\sum_{n=1}^{N} A_n \cos(kR\sin(\theta_n))\right)^2 + \left(\frac{1}{\sum A}\sum_{n=1}^{N} A_n \sin(kR\sin(\theta_n))\right)^2 \tag{4.4}$$

Where k is the wavenumber, R the separation between the measurement points, N is the number of loudspeakers A_n is the linear gain of the n-th loudspeaker and θ_n is the angle of the n-th loudspeaker to the median plane. The fit to the measured data is good up to 10 kHz for all stimuli (figure 4.25).

4.8 Summary

In this second experiment a variety of different loudspeaker arrangements and loudspeaker gain options were rated by listeners in terms of perceived diffuseness. The rated diffuseness correlates well with the number of loudspeakers although the position of the loudspeakers also has some effect on the perceived diffuseness. Moving off-centre has a strong effect on the perceived diffuseness and appears to be related to the movement towards the nearest loudspeaker as layouts without loudspeakers near the off-centre listening position appear more robust to listener movement. Objective metrics of the sound field have several biases. Some of these biases arise in reproduced audio but may not appear in the architectural acoustics where the metrics were developed. The spatial coherence appears to correlate best with the results and an analytical model for the coherence is presented. Up to this point the stimuli have been completely uncorrelated and this is not the case in real world situations. This is addressed in the following chapters.

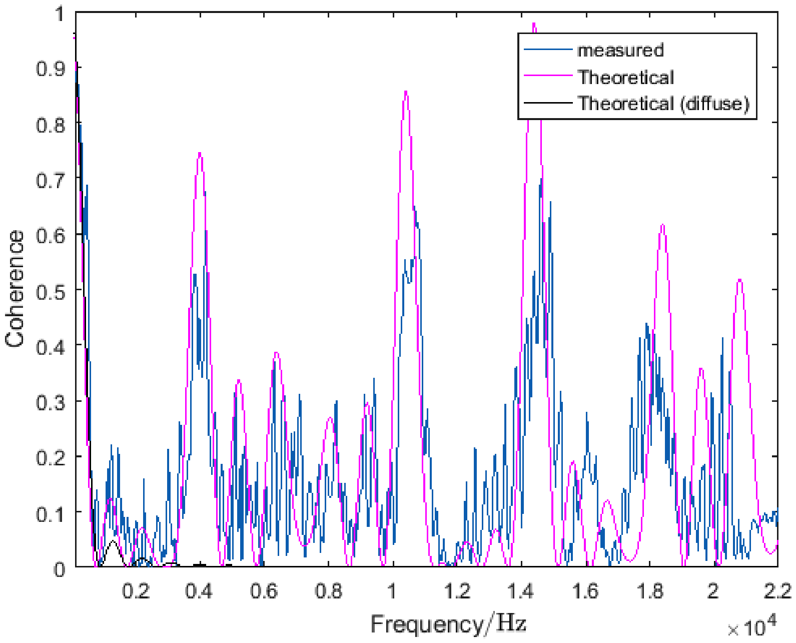


Figure 4.25: Coherence function between two microphones separated by 20 cm. Measured response, theoretical response for the arrangement of loudspeakers and theoretical response for a diffuse sound field.

Chapter 5

Experiment 3: Inter-Channel Correlation Coefficient

5.1 Overview

Experiments 1 and 2 both used completely uncorrelated signals. Real world signals that are not synthesised are likely to have some degree of correlation. Microphone techniques with limited directivity (for coincident techniques) or limited separation (for spaced techniques) or panning sources between loudspeakers will lead to some correlation between different loudspeakers.

The purpose of this experiment was to investigate the relationship between the Inter-Channel Correlation Coefficient (ICCC) between loudspeakers and the perceived diffuseness for different loudspeaker layouts.

The stimuli in the experiment were four loudspeaker layouts and six ICCCs in all combinations (section 5.2). The first part of this experiment was a pretest to ensure that all the stimuli were at the same loudness (section 5.4). Once aligned, all the stimuli were rated for their perceived diffuseness using the MUSHRA methodology (section 5.5). The results are discussed in section 5.7.

5.2 Stimuli

5.2.1 Loudspeaker Layouts

The number of loudspeakers appeared to be the main factor in experiment 2. Including a range of loudspeaker layouts allows the effect of varying the ICCC to be directly compared to the effect of varying the number of loudspeakers.

The layouts were chosen based on the following considerations

- A range of diffusenesses from medium to very high.
- Include some commercial layouts that are popular.
- Include both 2D and 3D.

The layouts chosen were, 5.0, 9.0, 0/12/0 and 12/12/13. In experiment 2, when uncorrelated, these layouts were rated in approximately even steps at 52.4, 61.6, 66.8 and 78.9 respectively ranging from medium perceived diffuseness to high perceived diffuseness. There are two 2D layouts and two 3D layouts; two standardised layouts and two layouts distributed evenly in azimuth around the listener. Whilst 12/6/13 was the most diffuse layout from experiment 2 and was rated more diffuse than 12/12/13, this was only by a small amount and using 12/12/13 allows for direct comparison with the 2D layout 0/12/0. Because the ISLD made little difference to the perceived diffuseness from using equal loudspeaker levels in experiment 2, the ISLD has been avoided in this experiment. For simplicity and the ISCLD=0 option where all loudspeakers are at equal level (i.e. $12/12/13_C$) is used in this experiment.

5.2.2 Method of decorrelation

There are several ways in which to generate partially correlated loudspeaker signals.

1. Adding a common signal to all the uncorrelated loudspeakers in a given ratio to increase correlation.
2. Sampling a theoretical diffuse sound field using coincident techniques with limited directivity.
3. Sampling a theoretical diffuse sound field using spaced techniques with limited separation.
4. The panning of uncorrelated components between loudspeakers.

These different methods give different correlation coefficient matrices. Methods 2, 3 and 4 would have different correlation coefficients between different loudspeakers. The interchannel correlation coefficient matrix for method 3 would be also be frequency dependent. Whilst these more “real world” cases may be more realistic, there is no single value for the correlation between loudspeakers and therefore there are additional variables. The first method is the simplest and is the one that was chosen for this experiment as the ICCC is the same between any two different loudspeakers.

Equation 5.1 was used to generate a range of correlations between loudspeaker signals. Y_n is the signal for the n-th loudspeaker and is given by,

$$Y_n = (\sqrt{(1-i)} \times X_n) + (\sqrt{i} \times W) \tag{5.1}$$

Where X is a set of N uncorrelated pink noise signals and W is a different pink noise signal common to all loudspeakers. The variable i is the inter-channel correlation coefficient between 0 and 1 and sets the ratio between the uncorrelated signals and the correlated signals. The square root maintains the r.m.s level for each loudspeaker. This gives the ICCC matrix shown in equation 5.2.

$$ICCC = \begin{bmatrix} 1 & i & i & i & & i \\ i & 1 & i & i & & i \\ i & i & 1 & i & & i \\ i & i & i & 1 & & i \\ & & & & \ddots & i \\ i & i & i & i & i & 1 \end{bmatrix} \tag{5.2}$$

If $i = 0$ the signals are completely uncorrelated and the inter-channel correlation coefficient matrix is an identity matrix. If $i = 1$, all loudspeakers are fully correlated and the ICCC matrix is an all-ones matrix.

Uncorrelated pink noise was generated using the dsp.ColoredNoise object in Matlab. These signals are generated using independent random sequence generators to ensure a correlation coefficient of 0 (for a long sequence).

5.2.3 Inter-Channel Correlation Coefficient (ICCC) values

There was a fear that high ICCC would sound “phasey”. The interference between loudspeakers would cause comb filtering and these frequency variations would be easier to distinguish than differences in diffuseness, causing bias. However, in an informal pretest, listeners were not distracted by the phasiness when rating the diffuseness and so the full range of ICCC was tested with six ICCC values from 0 to 1 with even intervals of 0.2.

5.2.4 Listening Positions

As in the previous experiment a pair of listening position are investigated, on-centre and a position 80 cm to the right. This is especially interesting in this experiment as the correlated signals from the different loudspeakers will interact differently at the different listening positions, adding coherently in the centre where the loudspeakers are time aligned and adding based on the propagation delays when off-centre.

5.2.5 Summary

The combination of four loudspeaker layouts, 6 ICCC levels and 2 listening positions and multiple repeats would lead to a fairly long experiment. To investigate all these was deemed unnecessary and therefore, in the off-centre listening position, only the most and least diffuse layouts were tested (5.0 and 12/12/13). This led to a total of 36 stimuli to be judged 3 times by each listener.

The introduction of the ICCC as a variable means that in this experiment extra care has to be taken when loudness matching the stimuli. This is discussed in section 5.4.

5.3 Reproduction System

The loudspeakers and listening room are the same in for the previous two experiments as described in section 3.4. The Audio Lab at the University of Southampton with T_{60} of 0.12 s $\pm$0.02 s in 1/3 octave bands between 125 Hz and 8 kHz and 39 Kef HS3001SE loudspeakers mounted to the walls and ceiling mounting system.

However, experiment 3 uses slightly different equalisation; 1/3 octave bands instead of 1/6 octave bands; the approximation for the roll-off of the loudspeakers is replaced by a mathematically simple -24 dB per octave roll off below 95 Hz. Also, if the roll-off goes below the noise floor, the target curve follows the noise floor. The iterative process has been replaced by a method of checking the equalisation is within $\pm$1 dB in all bands to avoid compounding errors in difficult bands (for example below the noise floor at low frequencies).

5.4 Loudness Alignment: Pretest

5.4.1 Overview

Loudness is a common source of bias in listening tests and alignment of loudness is important. However, in listening tests that are dealing with many loudspeakers and varying correlation between loudspeakers the loudness will vary.

At a central measurement position in anechoic, optimal conditions, uncorrelated signals add incoherently whereas correlated signals add coherently. However, in real situations there are several variables other than the correlation of the loudspeaker signals that affect the perceived loudness. The SPL is position dependent and is known to not always correlate well with the perceived diffuseness. With high correlation between loudspeaker signals the signals at the ears of a listener are complex and not commonly found in the real world. These complex binaural signals are likely to be difficult for objective metrics of loudness to predict accurately. The easiest way to get a “correct” value for the loudness of the various stimuli is to use a listening test. Loudness matching experiments have been found to not require many listeners as listeners are generally very consistent. Although this takes more time, with complex sound fields it is more accurate and therefore most likely to reduce bias in the main experiment testing the perceived diffuseness of the stimuli.

5.4.2 Listener Response Methodology

The subjective loudness alignment test was an adjustment task in which listeners were asked to match the loudness of a test stimulus to the loudness of a reference stimulus. Figure 5.1 shows the user interface implemented in Matlab.

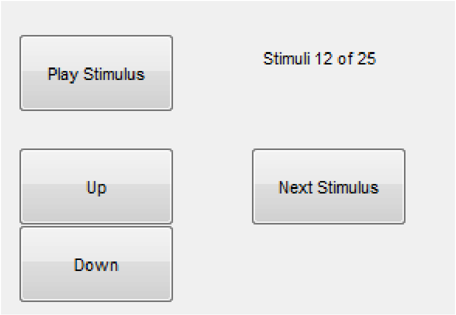


Figure 5.1: UI used for loudness alignment task.

Two second bursts of pink noise were played one after each other. The first being the reference and the second being the test stimulus. The listener could either increase the gain of the test stimulus by 1 dB or decrease the level of the test stimulus by 1 dB. There was also an option to replay the reference and test stimuli with the current gain setting. Changing the gain of the test stimulus would automatically replay the reference/test stimulus pair with the new level adjustment. Listeners were asked to use the up and down buttons to adjust the loudness of the second noise burst to match the loudness of the first noise burst. When the listener was happy they could click next and the next test stimulus would be selected.

The reference was always 12/12/13 with ICCC=0. This stimulus is the most consistent across the listening area in terms of colouration and SPL. It was also to be used in the next part of the test as the reference for the most diffuse stimulus.

All 25 stimuli (all combinations of 6 ICCC values and 4 layouts plus uncorrelated stereo) were adjusted against the reference.

5.4.3 Subjects

Each loudness comparison was made twice by each of the 4 listeners who were all post-graduates at the University of Southampton with self reported normal hearing.

5.4.4 Results

Post-screening

Table 5.1 shows the mean absolute difference between repeats of the same stimulus for each listener.

Listener	mean(abs($x_1 - x_2$))
1	**0.76**
2	**0.48**
3	**0.76**
4	3.28

Table 5.1: Mean absolute difference between repeats of adjustments in dB of the same stimulus.

Listener 4 has a large average difference between repeats (3.28 dB) and so was excluded. It is possible they moved significantly between repeats but this is not clear. The data from the other listeners all followed the same trends and were very consistent (<1 dB mean absolute difference between repeats).

Analysis and Discussion

The means from the 3 listeners that passed post-screening are plotted in figure 5.2 along with the standard deviation.

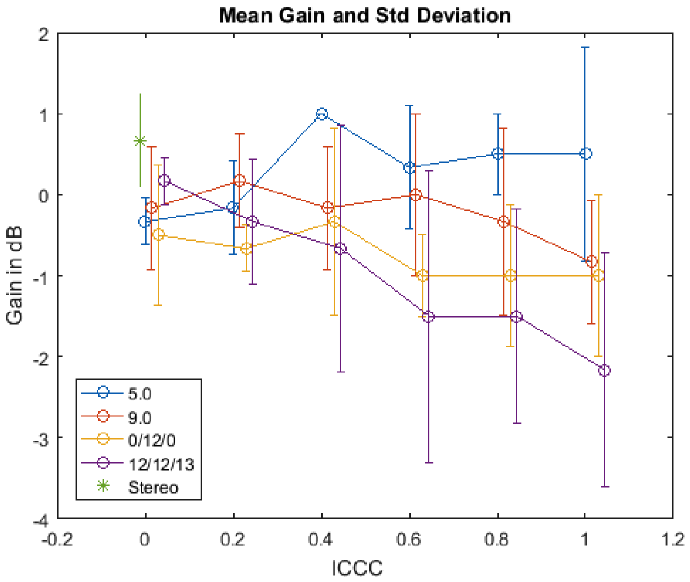


Figure 5.2: Mean and standard deviation of the adjustment gains to match a 12/12/13 ICCC=0 reference loudness.

Overall we see as expected that layouts with more loudspeakers get louder as the ICCC increases relative to layouts with fewer loudspeakers. At 0 ICCC the uncorrelated signals are adjusted to within around ±0.5 dB. We see 12/12/13, ICCC=0 is adjusted to very similar to the reference implying there is no large bias in level between the first sound played (the reference) and the second (the test). Interestingly listeners increased the level of 5.0 as the inter-channel correlation increased. This might be a bias of the listening test. Listeners expect that some stimuli may be louder than the reference and others quieter. Because the reference is highly uncorrelated, all other stimuli are, in theory, the same loudness or louder. When listeners are given the option to either increase the gain or decrease the gain they assume that either is equally likely and may end up increasing the gain when in doubt about the loudness differences but when there are obvious spectral and spatial differences. However seeing as the maximum gain is 0.5 dB and listeners were consistent, it is probably not a large enough bias to warrant any further action.

The variation between listeners is fairly large although individual listeners appear fairly consistent. All the adjustments are smaller than the theoretical differences. Theoretically, there is a 10×log(n) difference in SPL between ICCC=0 and ICCC=1 where n is the number of loudspeakers in the layout. However the results of the adjustment task show the subjective adjustments to be much lower. 12/12/13 is adjusted by -2 dB when correlated rather than the theoretical -16 dB. And 5.0 is adjusted to +0.5 dB instead of the theoretical -7 dB. Measuring the SPL at different distances from the alignment position shows one partial cause of this difference (figure 5.3).

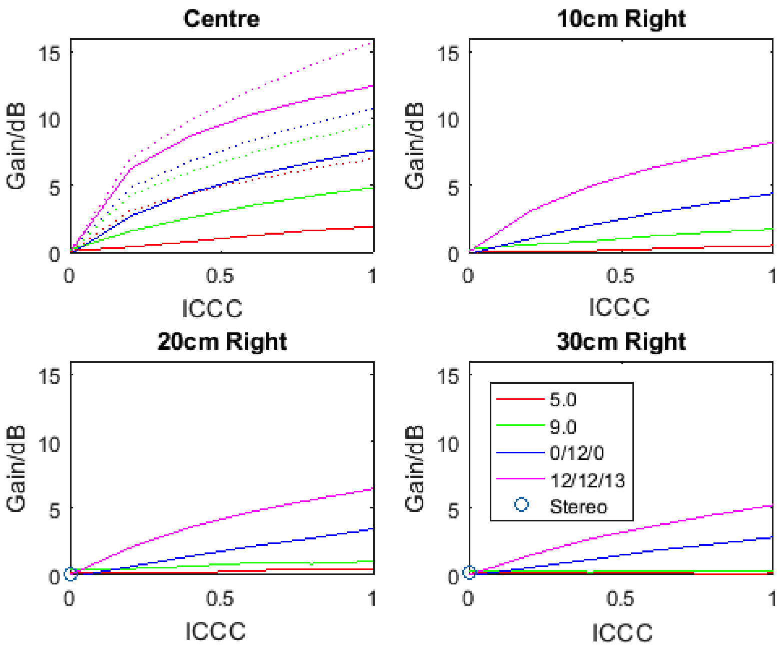


Figure 5.3: ICCC plotted against the difference in unweighted SPL between the reference and the test stimulus for various distances from the alignment position. Dotted line shows the theoretic level differences.

Firstly the SPL increase caused by increasing the ICCC is smaller as the listener moves away from the alignment position. And secondly, at the alignment position the SPL at ICCC=1 is lower than the predicted theoretical SPL difference. However, despite these measurements agreeing the subjective loudness adjustment are unlikely to be as large as the theoretical loudness adjustments, the subjective adjustments are still lower than the measured adjustments.

When the ICCC is high the loudspeaker signals will interfere creating an interference pattern. At the alignment position all signals arrive at the same time and are in phase. They should therefore sum coherently and increase the loudness. As soon as the measurement point is not in the exact centre, some frequencies will add in phase (causing summation) and others out of phase (causing cancellation). This will mean a lower SPL than at the centre. The microphone was not moved in between time alignment (nearest 1/48000 s) and the measurement of SPL. Therefore the difference between the measured and theoretical loudnesses at the on-centre position has some other cause. This could be related to the reverberation of the listening room. Whilst the reverberation time is low, the room modes will have an effect on the SPL and this will be position and room geometry specific. This may account for some of the difference between the theoretical SPL difference and the measured SPL difference. Further variation can be attributed to listener position. The distance of 10 cm is reasonable for the position for the ears of a listener in the exact adjustment position due to the interaural distance. And 30 cm is not unreasonable difference for a listener who slouches in the chair rather than sitting upright.

Whilst room acoustics and listeners not being exactly at the centre explains some of the variation between the theoretical level differences and the results of the listening test (figure 5.2), on-centre there is still a deviation between the theoretical and measured levels, and the listeners adjustments are still smaller than the measured gains (even far from the measurement position). The following paragraphs propose some of the possible causes of these differences.

Frequency Response of the Loudspeakers Small deviations in the response of the individual loudspeakers mean that the signals are no longer exactly the same and the 1/3 octave band equalisation is only ideal at the exact centre.

Spatial impression As the ICCC increases the interaural cues vary strangely and for high ICCC the sound appears almost as if monophonic. In this case the test stimulus (ICCC=1) is very different from the reference stimulus (ICCC=0) which is very diffuse. There is a possibility that this too is a factor. Maybe the smaller perceived size of the source is also perceived as less loud.

Frequency content The unweighted SPL assumes that all frequencies are perceived equally loud. In the theoretical case the spectrum does not change regardless of coherent or incoherent summing, all frequencies are treated the same. However, as soon as the listener does not have both ears at the exact centre, there will be colouration and this will affect the loudness. The unweighted SPL is independent of frequency with all frequencies contributing equally to the total SPL. Consider a single frequency at 100 Hz and a single frequency at 1 kHz both at 80 dBSPL(unweighted). These will incoherently sum to give an SPL that is equally due to the low and high frequency components (83 dB). If this signal was A-weighted then the low frequency signal is attenuated by 20 dB. Therefore the total SPL is dependent more on the SPL of the high frequency than the SPL of the low frequency (incoherent summing of 80 dB and 60 dB is 80.043 dB). Therefore, because listeners are more sensitive to mid and high frequencies which are also less in phase off-centre due to the short wavelengths, the loudness is likely to change by less than the predicted, unweighted SPL.

5.4.5 Gains Used for the Main Experiment

The best guess for the correct alignment gains for the main part of the test is the means gains from this adjustment experiment. However, because listeners move a little and some may be closer to the alignment position than others, there is likely to be some variation in the loudness that may be listener specific and could be biasing. For this reason an additional ±2 dB of uniform level roving was also added randomly to all the stimuli. The randomisation of the loudness should help any residual bias to average to zero.

In the test the inclusion of the off-centre position is to see how a system that is aligned to the centre is judged off-centre. The additional gain associated with the near loudspeaker is one cue that affects how the sound has changed from the central listening position. For this reason the mean gains on-centre plus the randomised roving were used for both listening positions.

5.5 Listener Response Methodology

A MUSHRA interface implemented in Matlab (figure 5.4) allowed listeners to replay a 2 second burst of a pink noise stimulus. Sliders allowed the stimuli to be rated and a next button allowed the next page of stimuli to be presented. A reference was included of 12/12/13 with ICCC=0. Uncorrelated (ICCC=0) stereo was used as a hidden low anchor. This was chosen as it was used in the previous experiment; was not usually rated 0 for diffuseness (allowing some scope for less diffuse stimuli when introducing the variable of ICCC); and also is a sensible lower limit for the diffuseness (with sound less diffuse than stereo considerably not diffuse). As a hidden anchor (unlike a reference), it does not limit the range of the given results. The 36 stimuli were split between 6 pages of 6 stimuli. With the anchor and reference on every page, a total of 6 pages of 8 stimuli were rated 3 times by each listener. The head was not locked in place.

The definition of perceived diffuseness in this experiment is slightly modified from the previous experiments. High correlation between loudspeakers is known to sound “phasey”. This is an unnatural effect caused by the interference of the sound field and an sound differently to different people. Common descriptors are “inside the head” or moving/unstable sources. In both these cases the signal is not perceived diffuse but also not easy to localise. Therefore the perceived diffuseness and the ability to localise are not antonymous and therefore the phrasing of the description of diffuseness used in the first two experiments was changed. The new instructions for listeners were,

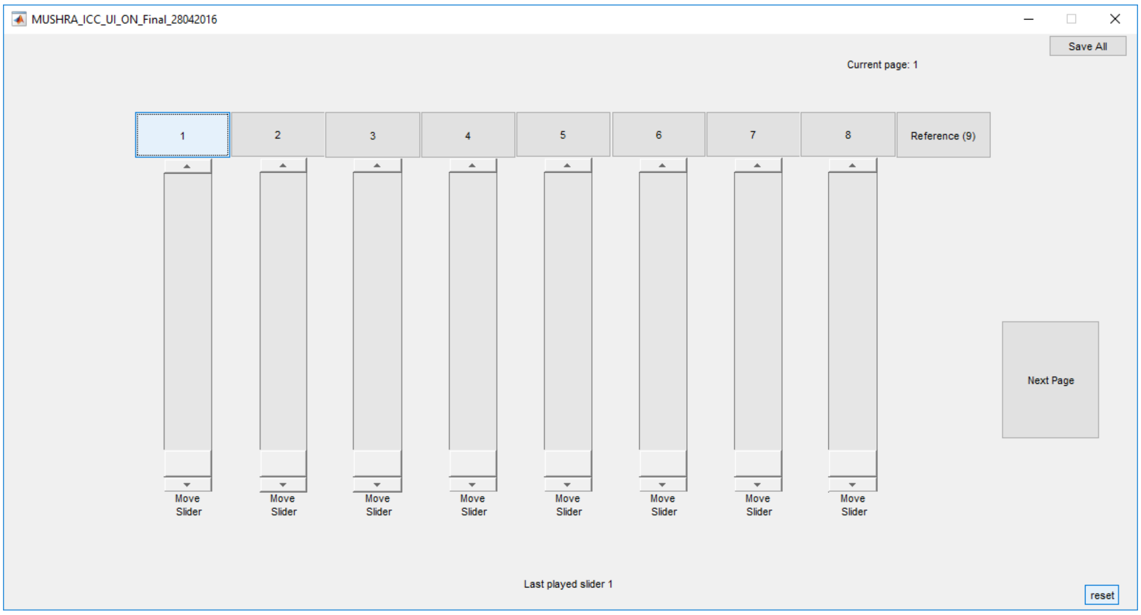


Figure 5.4: MUSHRA-style user interface used in experiment 3 implemented in Matlab.

> “You are to rate the how diffuse and surrounding you perceive each stimulus. Diffuseness is the degree of being surrounded/enveloped by the sound field. This may be heard when standing and listening to the rain hitting the pavement; applause in a concert hall; atmosphere or air conditioning (room tone). Being able to localise the source of the sound will decrease diffuseness. Holes (an absence of sound from a certain directions) would normally reduce envelopment. Feeling the sound inside your head or as moving/unstable sources would also usually be less diffuse. Try to ignore any differences you hear in terms of loudness, colouration, distortion, frequency content and focus purely on where the sound is coming from."

5.6 Subjects

The listening test was approved by the ethics and research governance committee (ID: 18709). The listening test was sat by 12 undergraduate and postgraduate students at the University of Southampton with self reported normal hearing. Of the listeners 7 could be considered experienced listeners with previous experience in more than one other listening test.

5.7 Results

5.7.1 Post-screening

All the listeners were consistent between repeats. Each listener had their repeats for the same stimulus averaged and the results are shown in the box plot in figure 5.5. All of the listeners seem to be congruent.

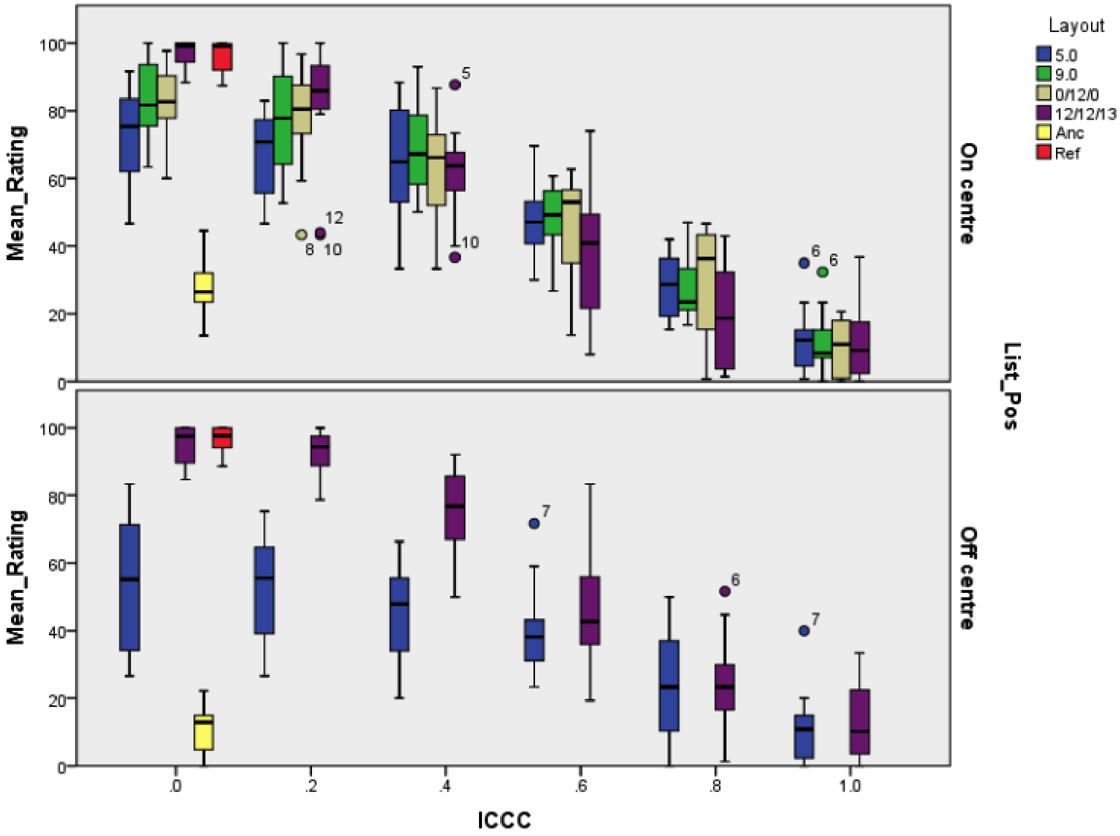


Figure 5.5: Box plots of mean ratings from each listener.

5.7.2 Analysis

The mean ratings for each stimulus are plotted in figure 5.6. On and off-centre listening positions are plotted on the same graph. In this experiment the question was rate the stimuli relative to the reference. Therefore off-centre the reference is the off-centre diffuseness of 12/12/13 whereas on-centre the reference is the on-centre diffuseness of 12/12/13. In experiment 2 (where there was no explicit reference and listeners were asked to be consistent between listening positions both the on and off-centre 12/12/13 stimuli were rated similarly so the graph plotted here should remain representative although differences between layout or ICCC are likely to be more reliable than differences between listening position where the listener could not directly compare stimuli on a single trial of MUSHRA.

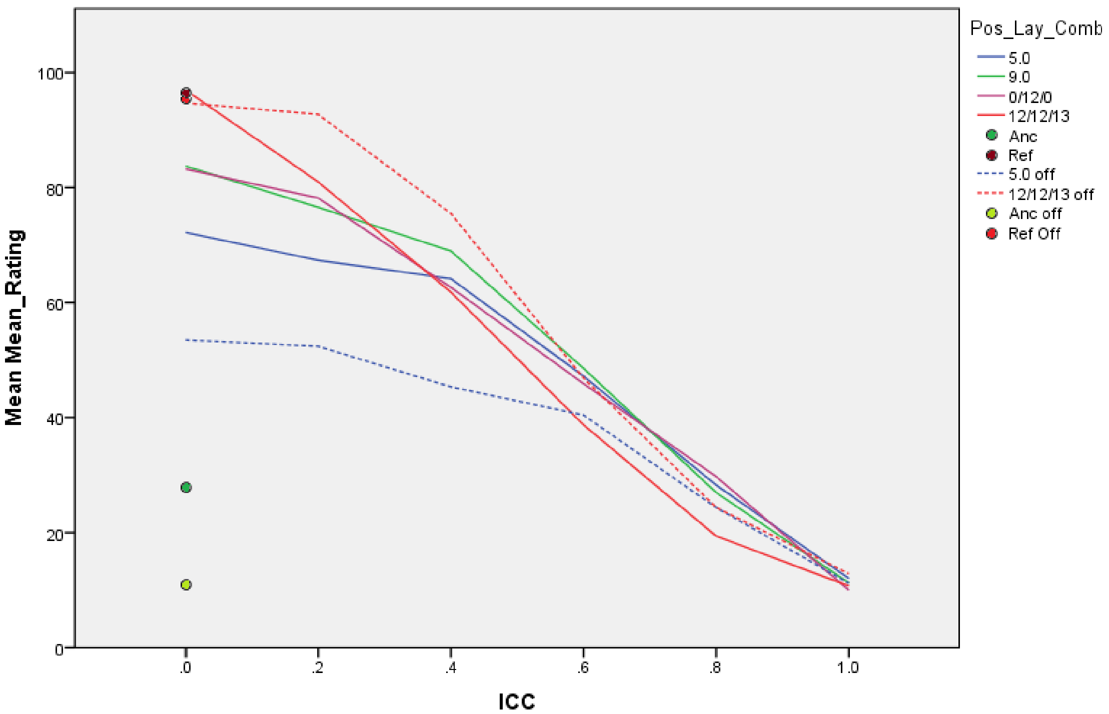


Figure 5.6: Mean diffuseness ratings for all stimuli.

Layout/Position	X^2	df	p	Kendall's W
5.0 On-centre	50.517	5	<**0.0005**	0.842
9.0 On-centre	57.337	5	<**0.0005**	0.956
0/12/0 On-centre	54.021	5	<**0.0005**	0.900
12/12/13 On-centre	58.029	5	<**0.0005**	0.967
5.0 Off-Centre	40.55	5	<**0.0005**	0.674
12/12/13 Off-Centre	58.385	5	<**0.0005**	0.973

Table 5.2: Friedman tests for effect of ICCC for each Layout/Position combination.

The reference at the top of the scale leads to skewed, non-normal rating distributions for stimuli at the top of the scale. Therefore Friedman testing was used to investigate significant differences for each variable in turn. Table 5.2 shows ICCC to be a significant factor for all layouts and listening positions.

Wilcoxon signed rank tests (table 5.3) show that in the on-centre position, the loudspeaker layout was a significant factor, but only for the uncorrelated case (ICCC = 0).

Off-centre, Wilcoxon signed rank tests (table 5.4) show significant differences at the $\alpha = 0.05$ level up to ICCC = 0.4 .

In experiment 2, the loudspeaker layout and specifically the number of loudspeakers was shown to be significant in the perceived envelopment of the stimulus. The results from this experiment show that the loudspeaker layout is only significant if the input signals have a very low ICCC. Although, the differences between loudspeaker layouts are more

ICCC	X^2	df	p	Kendall's W
ICCC = 0	24.399	3	<**0.0005**	0.842
ICCC = 0.2	7.667	3	0.053	0.213
ICCC = 0.4	1.900	3	0.593	0.053
ICCC = 0.6	5.521	3	0.137	0.153
ICCC = 0.8	7.301	3	0.063	0.203
ICCC = 1	2.745	3	0.433	0.076

Table 5.3: Friedman tests for effect of Loudspeaker Layout on-centre.

ICCC	Z	p
ICCC = 0	-3.059	**0.00**2
ICCC = 0.2	-3.059	**0.002**
ICCC = 0.4	-2.944	**0.003**
ICCC = 0.6	-1.413	0.158
ICCC = 0.8	-0.401	0.689
ICCC = 1	-0.847	0.397

Table 5.4: Wilcoxon signed rank tests for effect of Loudspeaker Layout off-centre.

pronounced off-centre than on-centre. In all cases the ICCC is a very significant factor that determines the perceived diffuseness of a system.

5.7.3 Discussion

Comparison of Results with Experiment 2

At ICCC=0 the pink noise is completely uncorrelated. Therefore the stimuli are the same as were used in experiment 2. The layout 12/12/13 has a different ISLD in the two experiments but this is unlikely to change the rated values notably for reasons discussed in section 4.6.2. The mean ratings for the stimuli that are the same in both experiments are displayed in table 5.5.

Layout	Mean (Experiment 2)	Mean (Experiment 3)
Stereo	12.2	27.8
5.0	52.4	72.2
9.0	61.6	83.8
0/12/0	66.8	83.2
12/12/13	78.9	97.0

Table 5.5: Comparison of the results of experiment 2 with ICCC=0 in experiment 3

In both experiments stereo is not diffuse, 12/12/13 is very diffuse and 5.0 is fairly diffuse. The layouts 9.0 and 0/12/0 are somewhere in between 5.0 and 12/12/13. The main difference is a translation of the ratings. The range has changed from 12.2-78.9 to 27.8 to 97.0. This is a mean increase of 18.4. The lowest rated stimuli in experiment 2 was stereo at 12.2 but in experiment 3 ICCC=1 (0/12/0) was rated the least diffuse at 10.0.

The addition of the reference appears to have increased ratings at the top end of the scale with better use of the 100. The addition of stimuli less diffuse than stereo appears to have increased the rating of stereo to make room for the less diffuse stimuli. These two factors appear to combine to result in a translation of the ratings and not a scaling of the ratings. In both the tests the lowest rating is around 10 but having a reference at the top end of the scale appears to spread the results better across the range (100 instead of 80).

The difference between 9.0 and 0/12/0 appears to have reduced in this experiment. It is possible that the use of a 3D anchor may make 9.0 appear more diffuse than before (more similar to the reference) where the 2D layout 0/12/0 may appear less diffuse.

On-centre versus Off-centre for ICCC=0

As in experiment 2, stereo and 5.0 become less diffuse whereas 12/12/13 remains just as diffuse. However, in this experiment the wording of the task was slightly altered. Here listeners were asked to rate everything relative to the reference. In the previous experiment listeners were asked to be consistent between pages. In the previous experiment 12/12/13 was rated the same in both listening positions and so the data here are probably the same. However one listener did feel that the diffuseness off-centre was a lot lower than for on-centre and so effectively used two different scales for the on and off-centre listening positions. However, as there was no large difference in the previous test, it is not unreasonable to plot these different positions on the same axes.

Relative Effect of ICCC and Layout

There is a strong dependence on the ICCC. All arrangements at coherence of 1 give a perceived diffuseness less than stereo. As the ICCC increases the dependence on the number of loudspeakers decreases. The difference between the four layouts is around 30 for ICCC=0. By the time ICCC is 0.3 the range is reduced to 10.

Off-centre for Different Layouts

The small differences between loudspeaker layouts on-centre are increased when moving off-centre. Moving off-centre appears to increase the perceived diffuseness when there is a small amount of correlation and many loudspeakers. Informal listening suggests there are two factors at play when moving off-centre. The listener is simultaneously moving towards some loudspeakers but also away from the position at which the signals are time aligned. The former effect is the same as was found in experiment 2. For layouts with fewer loudspeakers, the near loudspeaker appears very loud and localisable, reducing the perceived diffuseness. With many loudspeakers the effect is reduced. However, with ICCC moving off-centre also affects the interference pattern at the ears of the listener. The frequency at which all interference is constructive is now lower. There is some constructive interference at low frequencies but there is also some destructive interference and the hypothesis is that this sounds more diffuse as if the loudspeaker signal were slightly less correlated than they are.

Unexplained Data

By ICCC=0.5 three of the layouts have similar diffuseness regardless of the number of loudspeakers but the layout 12/12/13 is rated less diffuse than the others for ICCC 0.6 to 0.9. The possible cause is when the number of loudspeakers is extremely high the diffuseness saturates. More loudspeakers will not increase the perceived diffuseness but will increase the amount of colouration when the ICCC is high.

5.8 Comparison with Objective Metrics

These results show that the isotropy of the sound field is not the only condition for high perceived diffuseness. The loudspeaker arrangement and loudspeaker gains determine the isotropy of the sound field, however, the perceived diffuseness is also dependent on the ICCC. Therefore any objective metrics for predicting the perceived diffuseness based solely on isotropy such as rotating a directional microphone or the diffuseness estimate used in DirAC will be unable to predict the results found in this experiment. The sound field in spatial audio can be unnatural and isotropy-based metrics are designed for natural scenes such as reverberation or complex sound scenes.

In this section the data are plotted against the IACC and the spatial coherence. The IACC is commonly used and the coherence good correlation with the subjective data in experiment 2.

IACC

To calculate the IACC, all stimuli were recorded using a Neumann dummy head (KU 100). The recorded signals were windowed and the IACC calculated in octave bands for each time window. The IACC values were averaged to give a single value for the IACC for each stimulus. This averaging was done either over the entire frequency range (as in figure 5.7) or across the central frequency band of 500 Hz, 1 kHz and 2 kHz (as is shown in figure 5.8). The IACC values are also reported in tables 5.6 and 5.7.

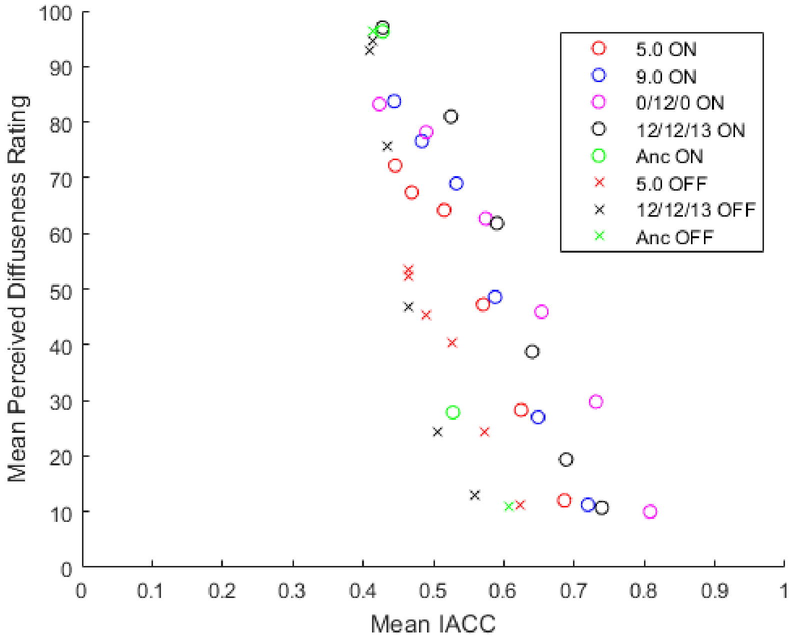


Figure 5.7: IACC for all stimuli averaged across all frequency bands.

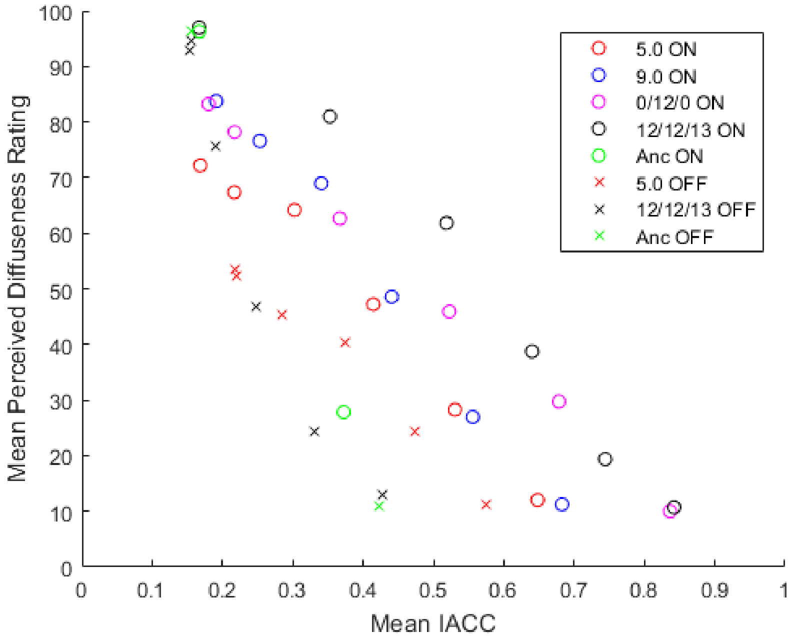


Figure 5.8: $IACC_3$ for all stimuli. The IACC is averaged across the middle frequency bands (500-2000 Hz).

Whist there is some degree of correlation in figure 5.7, it is small with different loudspeaker layouts and listener positions poorly estimated. Within a single loudspeaker layout

Listener Position	Loudspeaker Layout	ICCC					
		0	0.2	0.4	0.6	0.8	1
On-centre	Stereo Anchor	0.53					
	5.0	0.44	0.47	0.52	0.57	0.63	0.69
	9.0	0.44	0.48	0.53	0.59	0.65	0.73
	0/12/0	0.42	0.49	0.57	0.65	0.74	0.82
	12/12/13	0.43	0.53	0.59	0.65	0.69	0.75
	12/12/13 Reference	0.42					
Off-centre	Stereo Anchor	0.61					
	5.0	0.46	0.46	0.49	0.53	0.57	0.62
	12/12/13	0.42	0.41	0.44	0.46	0.51	0.56
	12/12/13 Reference	0.42					

Table 5.6: IACC values averaged across all octave bands.

Listener Position	Loudspeaker Layout	ICCC					
		0	0.2	0.4	0.6	0.8	1
On-centre	Stereo Anchor	0.37					
	5.0	0.17	0.22	0.31	0.42	0.53	0.65
	9.0	0.19	0.25	0.35	0.44	0.56	0.68
	0/12/0	0.18	0.22	0.37	0.52	0.68	0.84
	12/12/13	0.16	0.35	0.52	0.64	0.75	0.85
	12/12/13 Reference	0.16					
Off-centre	Stereo Anchor	0.43					
	5.0	0.22	0.22	0.29	0.38	0.48	0.58
	12/12/13	0.15	0.15	0.19	0.25	0.33	0.43
	12/12/13 Reference	0.15					

Table 5.7: IACC values averaged across octave bands 500 Hz, 1 kHz and 2 kHz (IACC_3).

and listening position, the fit to the data is good with stimuli with different ICCCs well distinguished and ranked relative to each other. However, the IACC is not able to make accurate predictions between both changes to loudspeaker layout or listener position and changes in ICCC.

When using the full frequency range to calculate the IACC, the mean IACC is higher than when using only the middle frequencies. This is due to low frequencies always having a high IACC and therefore raising the average when included in the averaging stage. When only the centre frequencies are used, a wider range is shown but the correlation is not notably increased .

Whilst the fit looks poor, it is once again worth mentioning that the on-centre and off-centre listening positions are not compared directly on the same page of MUSHRA. Therefore plotting and comparing between both listening positions on the same axis may be misleading. Looking at both listening positions independently, the correlation to the data is improved. However there is still a clear dependence on the loudspeaker layout that is not captured using the IACC.

Spatial Coherence

To calculate the spatial coherence, all the stimuli were recorded using a pair of microphones separated by 17 cm in both listening positions. The coherence was calculated between these signals and the coherence averaged across all frequency bins. These average coherence values for each stimulus are plotted against the perceived diffuseness for each stimulus in figure 5.9.

The spatial coherence shows a better fit to the data than the IACC. There is a close grouping between both the different loudspeaker layouts and the different listening positions. As with the IACC, the spatial coherence separates well the different IACC values and places them in order from most diffuse to least diffuse. However, the spatial coherence also predicts clearly 12/12/13 to be more diffuse than 5.0.

In experiment 2, the spatial coherence appeared to have a linear relationship with the perceived diffuseness. In this experiment we se a slight curvature at the bottom of the scale. In this experiment there are stimuli that are less diffuse than the least diffuse stimulus in experiment 2.

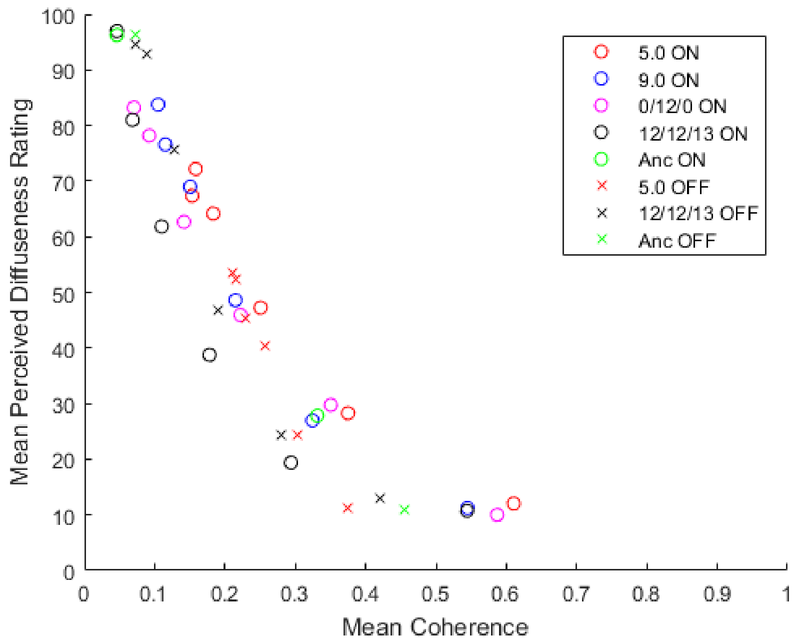


Figure 5.9: Mean value of the coherence averaged across all frequency bins and plotted against the perceived diffuseness.

5.9 Summary

This experiment looked at the effect of the inter-channel correlation coefficient between loudspeaker signals on the perceived diffuseness. The ICCC has a large effect with high ICCC perceived as not diffuse regardless of the number of loudspeakers in the layout (the main factor found in experiment 2). The effect of ICCC is also interesting off-centre where the perceived diffuseness decreases for layouts with few loudspeakers at low ICCC but increases for layouts with many loudspeakers and low ICCC. At high ICCC there appears no effect of moving off-centre as once again all layouts are perceived as equally not diffuse. The spatial coherence was shown to be a good predictor of perceived diffuseness in this experiment as well as for experiment 2. Whilst these results are interesting, they are limited to the material tested. Experiments to investigate the ICCC further are described in the next section and will hopefully lead to a greater understanding of the perception of diffuseness allowing for more generalised models based on the sound field and binaural signals than on the specific signals fed to the loudspeakers.

Chapter 6

Experiment 4: Interaction Between the ICCC and Frequency

6.1 Overview

Experiment 3 investigated the effect of inter-channel correlation coefficient (ICCC) on the perceived diffuseness. This was found to be highly significant however the ICCC has several different effects on the sound field. When the ICCC is high the coherent components of the sound field form an interference pattern by adding either constructively or destructively. This will depend on the positions of the loudspeakers, the measurement position(s) and the frequency.

The various effects are demonstrated in figure 6.1 using a plane wave simulation of the 0/12/0 layout.

For low ICCC the frequency response is flat with all frequencies adding incoherently. When the ICCC increases, an interference pattern is produced. There is a bass boost at low frequencies and a set of peaks and notches related to the coherent cancellation or summation of higher frequencies. Exactly on-centre all frequencies would add coherently leading to a flat frequency response, however, for a pair of receivers separated by 17 cm not all frequencies are in phase and so there are cancellations even in the on-centre position. As the listener moves off-centre, the frequencies of these features decreases. The cut-off of the low frequency boost is at a lower frequency as are the cancellations. Additionally, the limited spacing of the human ears leads to a high correlation between the ear signals at low frequencies regardless of the ICCC.

Using narrow frequency bands was chosen as the method to separate out these multiple effects. It was predicted that for low frequencies the ICCC would have less effect on the

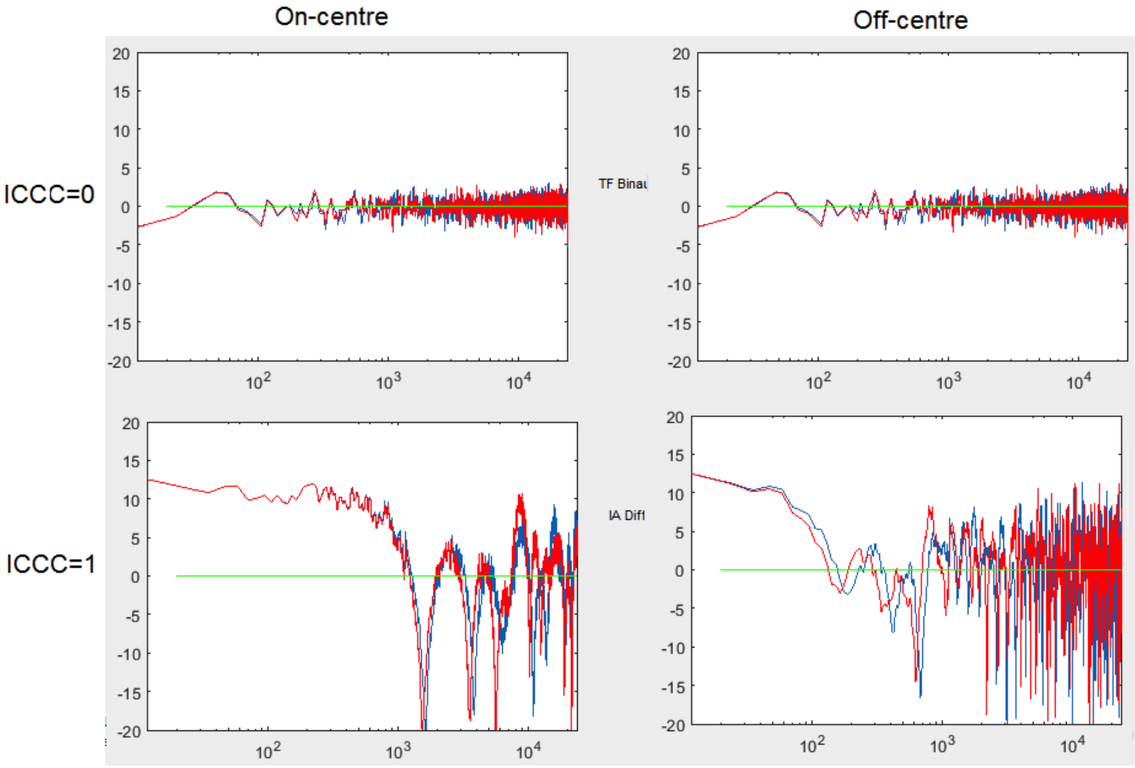


Figure 6.1: Effect on the frequency response when moving off-centre or changing the ICCC at two measurement positions separated by 17 cm and simulated plane waves for 12 evenly spaced head-height loudspeakers.

perceived diffuseness where the long wavelength relative to the interaural distance would make differentiating conditions more difficult. Also it was predicted that moving off-centre would have the same effect as decorrelating the signals and therefore reduce the effect of increasing the ICCC. This would be more noticeable at high frequencies where the highly detailed time structure is effectively decorrelated with shorter time shifts than at low frequencies. These predictions and observations led to the hypotheses outlined in the following section.

6.2 Hypotheses

The dependence of the frequency response at the ears of a listener on the ICCC and listener position led to the following hypotheses,

- At LF the ICCC only affects the level of both ear signals and does not affect the interaural differences. Therefore ICCC affects low frequencies less than higher frequencies.

- Moving off-centre increases the number of cancellations in a given frequency range. This gives a more complex frequency response at HF. If the frequency response is complex enough the listener cannot distinguish it from a flat frequency response. For high enough frequencies (likely above 6 kHz), far enough off-centre, the ICCC will have little effect on the perceived diffuseness.
- And finally these effects above are on the assumption of a plane wave model and in reality moving towards the nearest loudspeaker may outweigh these effects.

These hypotheses led to the null hypothesis that the ICCC affects the perceived diffuseness of a narrow frequency band equally regardless of the interaction between centre frequency and listener position.

6.3 Stimuli

Narrow 1/3 octave bandwidth noise signals were generated for three centre frequencies (125 Hz, 1 kHz and 8 kHz). The low frequency band was as low as possible considering the frequency response of loudspeakers. The high frequency band was chosen to allow a high modal density when off-centre. The bandwidth was wide enough to be perceived as noise like for all centre frequencies without exciting much more than a single critical bandwidth.

The off-centre listening position was changed from previous experiments to 1 m to the right of centre. This was done to hopefully increase any effects of moving off-centre by increasing the variation in propagation delays between the various loudspeakers.

The layout 12/12/13 was chosen as the loudspeaker arrangement for this experiment as this had a small difference in perceived diffuseness when moving off-centre.

The ICCC values were the same as for the previous experiment (0, 0.2, 0.4, 0.6, 0.8 and 1) and were generated in the same way, using a summation between a component uncorrelated between loudspeakers and a component common to all loudspeakers (equation 5.1 in section 5.2.2).

6.4 Reproduction System

The listening test was performed in the Audio Lab at the University of Southampton with T_{60} of 0.12 s ±0.02 s in 1/3 octave bands between 125 Hz and 8 kHz. The 39 Kef HS3001SE loudspeakers mounted to the walls, floor and ceiling were equalised in 1/3 octave bands from 95 Hz to 20 kHz. Below 95 Hz a -24 dB per octave roll off followed the approximate frequency response of the loudspeakers. Digital delay compensated for the differing propagation delays between the loudspeakers and the central listening position to the nearest sample (the sampling frequency was 48 kHz throughout).

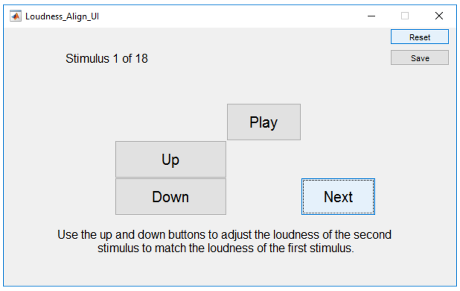
6.5 Pretest: Loudness Alignment

6.5.1 Overview

As in experiment 3, the ICCC will affect the loudness. In this experiment, there are perceptual differences in loudness at different frequencies –at a given sound pressure level– in addition to the differences in loudness that depend on coherent vs. incoherent summing when varying the ICCC. With loudness as a common source of bias in subjective listening tests, the stimuli were normalised by perceptual loudness and, as with experiment 3, this was done using the subjective adjustment task described in this section.

6.5.2 Listener Response Methodology

The listening test was an adjustment task in which listeners were asked to match the loudness of a test stimulus to the loudness of a reference stimulus. Figure 6.2 shows the user interface implemented in Matlab.



Use the up and down buttons to adjust the loudness of the second stimulus to match the loudness of the first stimulus.

Figure 6.2: UI used for loudness alignment task.

The presentation method was identical to that in experiment 3 with a pair of 2 second bursts of pink noise played one after each other. The first being the reference and the second being the test stimulus. The listener could either increase the level of the test stimulus by 1 dB or decrease the level of the test stimulus by 1 dB. There was also an option to replay the reference and test stimuli with the current gain setting. Listeners were asked to use the up and down buttons to adjust the loudness of the second noise burst to match the loudness of the first noise burst. When the listener was happy they could click next and the next test stimulus would be selected.

The reference was always 1 kHz with an ICCC of 0 (uncorrelated).

All 18 stimuli (all combinations of 6 ICCC values and 3 frequencies) were adjusted twice against the reference by each listener.

6.5.3 Subjects

Each stimulus was rated twice by each of the 3 listeners who were all postgraduates at the University of Southampton with self reported normal hearing.

6.5.4 Gains Used for the Main Experiment

The mean alignment gains for each of the stimuli are shown in figure 6.3.

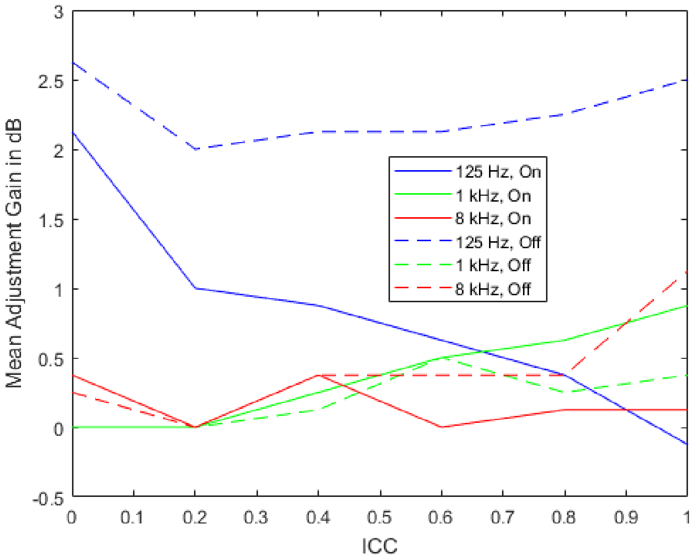


Figure 6.3: Mean adjustment gains to match the loudness of 12/12/13 at 1 kHz with ICCC=0 (uncorrelated).

For the 1 kHz and 8 kHz stimuli there is very little adjustment required. At low frequencies on-centre, the low frequency stimulus was 2 dB quieter than the other frequencies when uncorrelated.

When the ICCC is high the coherent summing of signals increases the loudness and therefore less gain is required to match the loudness of the other frequencies.

Conversely, when the listener is off-centre, the signals are no longer time aligned and so the required adjustment is close to that of the uncorrelated value of 2 dB. This 2 dB relates only to the difference in loudness between frequencies rather than the differences in loudness due to either coherent or incoherent summing.

These mean adjustment values were used in the main experiment with an additional ±2 dB of randomised gain.

			Test Stimulus																	
	Centre		125 Hz						1 kHz						8 kHz					
	Frequency	ICCC	0	0.2	0.4	0.6	0.8	1	0	0.2	0.4	0.6	0.8	1	0	0.2	0.4	0.6	0.8	1
	125 Hz	0	✓	✓	✓	✓	✓	✓	✓A						✓B					
Reference	1 kHz	0							✓	✓	✓	✓	✓	✓	✓C					
	8 kHz	0													✓	✓	✓	✓	✓	✓

Table 6.1: Comparisons made by each listener in experiment 2. All stimuli use the 12/12/13 loudspeaker arrangement. ✓ABC are the comparisons between different frequencies (inter-frequency).

6.6 Listener Response Methodology

In this main experiment there is a different focus than in previous experiments. The idea is to test whether different frequencies are robust or not to changes in ICCC for the reasons stated in the hypotheses section. Additionally in previous experiments all the stimuli were pink noise allowing for simple comparison of the spatial differences between the stimuli. In this experiment the different frequency bands will inherently sound different and this could easily lead to bias. It was therefore decided to avoid bias using a slightly different methodology to that used in previous experiments. Firstly, instead of comparing all stimuli to each other with a common reference, each stimulus is compared to an appropriate reference. In this case the reference is always the uncorrelated version for that frequency. In this way the listeners were always comparing between two stimuli at the same frequency. In addition to this changing of references, the MUSHRA type test is avoided in preference for an A/B type experiment with slider. When the listener is provided many stimuli on a single page they know the range of the stimuli and range equalisation bias and stimulus spacing bias lead the listener to use the whole scale as a linear scale rather than an absolute scale of what they perceive (Zielinski et al., 2007). Getting Absolute data rather than relative data is especially important to allow comparison between different frequencies and listener positions when they cannot be easily compared to each other directly. By presenting the listener with only 2 stimuli at a time this is minimised and unbiased absolute differences can be obtained. The disadvantage of only comparing stimuli with stimuli of the same frequency is that there are no data to compare the relative diffusenesses of the different frequency bands. Therefore additional comparisons were added that compare between the references for each frequency band. Because the references are completely uncorrelated the task is simplified relative to comparing differences in both frequency and ICCC at the same time. All the comparisons are shown in the table 6.1.

To perform these comparisons, the user interface show in figure 6.4 was used featuring two buttons labelled A and B which would play 2 s bursts of the two stimuli respectively (with 10 ms fade in and out to avoid clicking). One stimulus would be uncorrelated

(ICCC=0) as the reference and the other would be the test stimulus with either an ICCC between 0 and 1 but with the same centre-frequency (intra-frequency comparisons) or with an ICCC of 0 but a different centre-frequency (inter-frequency comparisons). The listener would then decide which of the two was most diffuse and then move the horizontal slider in that direction depending on how much more diffuse they perceived that stimulus. The slider was labelled “A is much more diffuse than B” (slider fully left), “A and B are equally diffuse” (slider in the centre) or “B is much more diffuse than A” (slider fully right).

The order of all the comparisons was randomised to avoid bias. Similarly, the pairwise presentation of the stimuli minimises the results of one comparison influencing the results of another comparison as can be the case with MUSHRA-style tests.

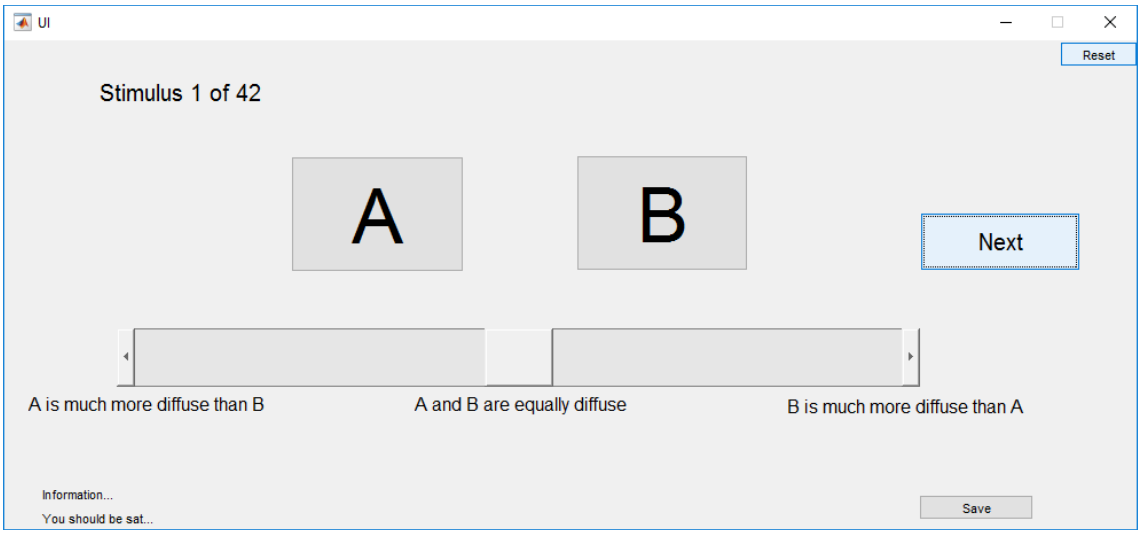

UI
Reset
Stimulus 1 of 42
A
B
Next
A is much more diffuse than B
A and B are equally diffuse
B is much more diffuse than A
Information...
You should be sat...
Save

Figure 6.4: A/B with slider style user interface used in experiment 4 and implemented in Matlab.

6.7 Subjects

The listening test was approved by the ethics and research governance committee (ID: 18709). Each comparison was completed three times by 9 postgraduate listeners at the University of Southampton with self reported normal hearing.

6.8 Results

In this experiment there is neither a common reference for all the comparisons nor comparisons between every possible stimulus pair. Each frequency and listening position is rated relative to its own reference and then, additionally, there are comparisons between the references in both the listening positions. The results–following post-screening–are therefore

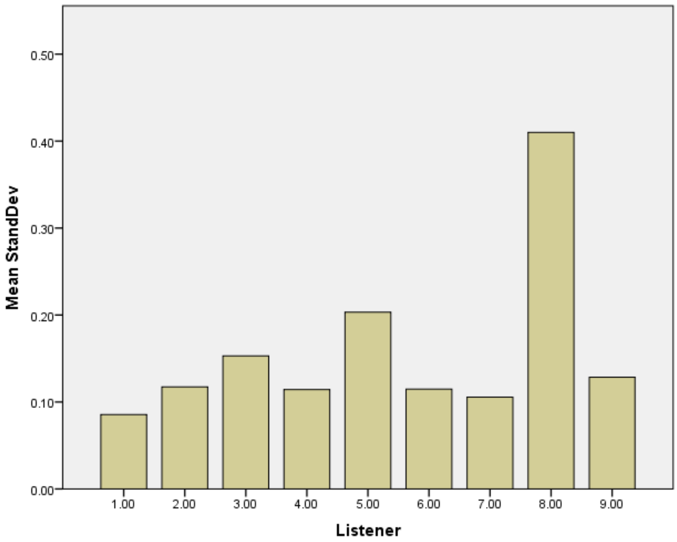


Figure 6.5: Mean of the standard deviation between repeats of the same stimulus for each listener.

shown in two parts. The first part has each listening position and frequency rated relative to its own reference (intra-frequency comparisons). In the second part, the comparisons between the frequency references (inter-frequency comparisons) are investigated.

6.8.1 Post-screening

Firstly the data were screened using the standard deviation between repeats of the same stimulus as a metric of listener consistency as show in figure 6.5.

Listener 8 was found to be very inconsistent between repeats. They also often moved the slider when the test stimulus was identical to the reference stimulus. For these reasons, listener 8 was removed from the remaining analysis.

The three repeats from the remaining 8 listeners were then averaged to give a single value as the rating for that particular comparison for each listener. These results are shown in the two box plots in figures 6.6 and 6.7. A value of zero represents that both stimuli were equally diffuse, a negative value indicates the reference was more diffuse and a positive value indicates the test stimulus was more diffuse. In the case of inter-frequency stimuli, the reference is arbitrarily the lower frequency. Listener 9 has a few outliers and tended to use more of the bottom of the scale, but in general the listeners were fairly congruent between each other and so only listener 8 was excluded from further analysis.

6.8.2 Analysis

Friedman tests (tables 6.2 and 6.3) are used to identify statistically significant factors.

Position	Centre Frequency	X^2	df	p	Kendall's W
On-centre	125 Hz	29.036	5	<**0.0005**	0.726
	1 kHz	34.600	5	<**0.0005**	0.865
	8 kHz	35.276	5	<**0.0005**	0.882
Off-centre	125 Hz	21.454	5	**0.001**	0.536
	1 kHz	29.549	5	<**0.0005**	0.739
	8 kHz	30.562	5	<**0.0005**	0.764

Table 6.2: Friedman Tests showing the effect of ICCC for the different listening positions and centre frequencies.

Position	ICCC	X^2	df	p	Kendall's W
On-centre	0	3.391	2	0.183	0.212
	0.2	7.000	2	**0.030**	0.438
	0.4	3.250	2	0.197	0.203
	0.6	14.25	2	**0.001**	0.891
	0.8	10.516	2	**0.005**	0.657
	1	6.258	2	**0.044**	0.391
Off-Centre	0	1.368	2	0.504	0.086
	0.2	3.364	2	0.186	0.210
	0.4	7.750	2	0.484	0.021
	0.6	1.750	2	0.417	0.109
	0.8	3.250	2	0.197	0.203
	1	3.000	2	0.223	0.188

Table 6.3: Friedman Tests showing the effect of centre frequency for the different listening positions and ICCCs.

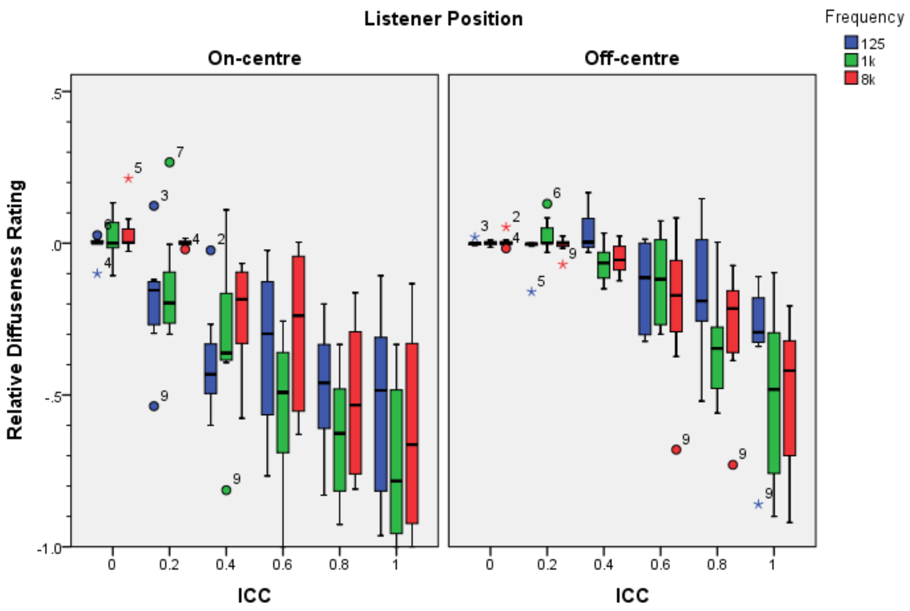


Figure 6.6: Box plots of mean ratings from each listener for intra-frequency stimuli. A rating of zero equates to the test ICCC being equally diffuse as the ICCC=0 for the same frequency. Negative values show the test stimulus to be less perceptually diffuse than the reference. Outliers are labelled with listener IDs.

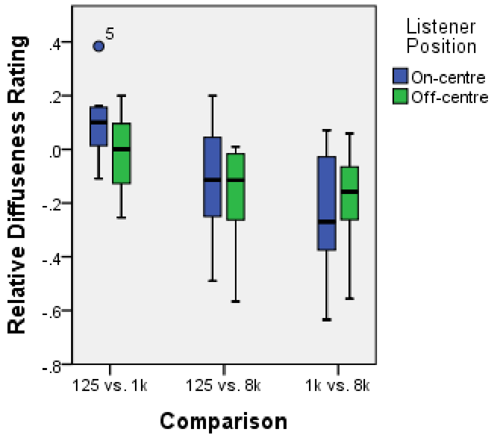


Figure 6.7: Box plots of mean ratings from each listener. In each case the value is the diffuseness of the higher frequency relative to the lower frequency. i.e. positive values mean the higher frequency is more diffuse and negative values mean the higher frequency is less diffuse.

Table 6.2 shows ICCC to be a significant factor for both listening positions and at all centre frequencies.

Table 6.3 shows that the centre-frequency is only a significant factor in some on-centre cases. It is expected that there is no difference at ICCC=0 where the reference and test are identical and therefore the rating should be zero for all centre frequencies.

Wilcoxon Signed Ranks tests (table 6.4) were used to investigate significant differences between the two listening positions. These show several of the ICCC, centre frequency

Centre Frequency	ICCC	Z	p
125 Hz	0	-0.135	0.893
	0.2	-2.240	**0.025**
	0.4	-2.380	**0.017**
	0.6	-1.540	0.123
	0.8	-2.521	**0.012**
	1	-1.680	0.093
1 kHz	0	-0.524	0.600
	0.2	-1.820	0.069
	0.4	-2.100	**0.036**
	0.6	-2.521	**0.012**
	0.8	-2.371	**0.018**
	1	-2.521	**0.012**
8 kHz	0	-0.405	0.686
	0.2	-0.524	0.600
	0.4	-2.375	**0.018**
	0.6	-1.122	0.262
	0.8	-2.380	**0.017**
	1	-1.400	0.161

Table 6.4: Wilcoxon Signed Ranks Test comparing the results from the different listening positions.

Position	X^2	df	p	Kendall's W
On-Centre	12.250	2	**0.002**	0.766
Off-Centre	2.889	2	0.236	0.160

Table 6.5: Friedman Tests showing the effect of the comparison pair (which frequencies were being compared) for the different listening positions.

combinations to be statistically significantly different between the two listening positions. No correction has been applied due to the large number of comparisons due to the conservative nature of family-wise error rate correction methods such as Bonferroni Adjustment. Although it should be highlighted that the p-values are relatively high and there are many comparisons. Applying any correction for multiple comparisons would likely show these differences to not be significant. The differences are far less significant than, for example, the differences due to the ICCC.

Friedman tests and Wilcoxon Signed Rank Tests were used to assess the statistical differences between the inter-frequency comparisons. Tables 6.5 and 6.6 show the effects of the comparison pair (which 2 frequencies are being compared) and the listener position respectively. On-centre there was a significant difference in the rating depending on which frequencies were being compared. Off-centre the differences were not significant. The listening position was not significant regardless of frequency.

Comparison	Z	p
125 Hz vs. 1 kHz	-1.014	0.310
125 Hz vs. 8 kHz	-0.338	0.735
1 kHz vs. 8 kHz	-1.120	0.263

Table 6.6: Wilcoxon Signed Ranks Test comparing the effect of listener position for different frequency comparisons.

6.8.3 Discussion

The intra- and inter- frequency comparisons are discussed separately.

Intra-frequency Comparisons

To view the trends in the intra-frequency comparisons, the mean ratings are taken and shown in figure 6.8.

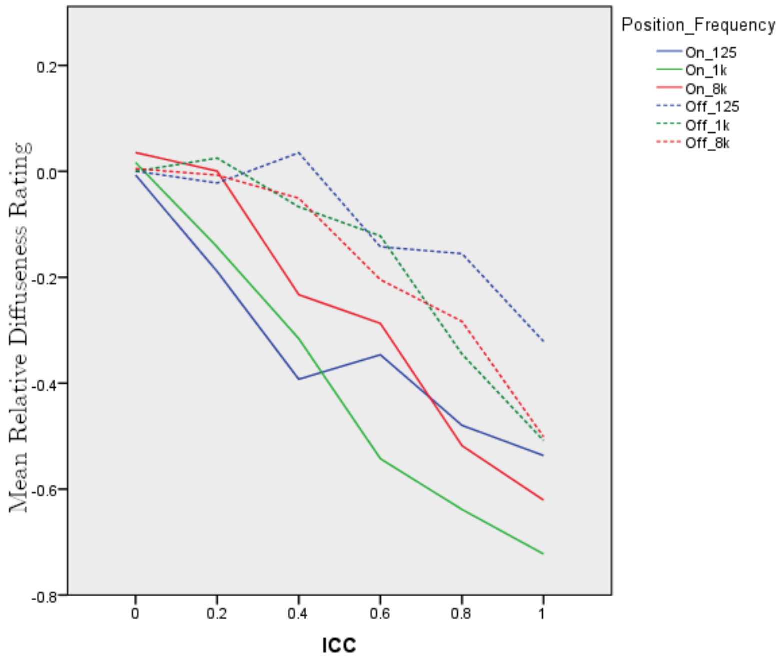


Figure 6.8: Mean relative diffuseness ratings for all intra-frequency stimuli comparisons.

As in experiment 3 (chapter 5), the ICCC has a large effect with high ICCC less diffuse than the low ICCC reference.

As in experiment 3, when off-centre, low ICCCs have less effect on the perceived diffuseness than when on-centre. This can be seen off-centre at ICCC=0.2 which is rated the same as ICCC=0. Interestingly, on-centre, 8 kHz appears to demonstrate a similar trend with ICCC=0.2 seemingly indistinguishable from ICCC=0. It is possible that the short wavelength of 8 kHz makes even the on-centre position exhibit the same trends seen off-centre at lower frequencies.

The stimuli at 125 Hz also show a strong significant dependence on the ICCC despite the high correlation between ear signals at this low frequency.

In this experiment the listener position is also highly significant. In the previous experiments the lines converged at high ICCC. It is possible that the MUSHRA type listening test of experiment 3 exhibited some range equalisation bias between the two different listening positions (Zielinski et al., 2007) encouraging the use of the bottom of the scale when off-centre and thereby normalising both ranges and hiding any differences.

The variation on-centre is greater than off-centre with larger interquartile range and standard deviation. This could be due to the fact that small variations in listener position have a larger effect on-centre than off-centre. Alternatively, it may be harder for listeners to be consistent when the difference in perceived diffuseness between the reference stimulus and the test stimulus is large and whilst using an A/B type experiment (as opposed to the MUSHRA method from experiment 3).

Inter-frequency Comparisons

The data from the inter-frequency comparisons are less intuitive to understand. Therefore, to visualise the data from the inter-frequency comparisons, the mean ratings were combined to give a perceived diffuseness for each frequency arbitrarily normalised to the diffuseness at 1 kHz. The inter-frequency comparisons are labelled as A (125 vs. 1k), B (125 vs. 8k) and C (1k vs. 8k) in table 6.1. Therefore the perceived diffuseness of the different frequencies, with reference to the diffuseness of 1 kHz, can be calculated for 125 Hz by either $-A$ or by $C-B$, and the for 8 kHz by either C or $-A+B$. The two alternatives were averaged so that $Diffuseness_{125} = (-A+C-B)/2$ and $Diffuseness_{8k} = (C-A+B)/2$ where A, B and C are the mean diffuseness ratings for the inter-frequency comparisons. These relative frequency diffusenesses, that have been calculated from the inter-frequency comparisons, are plotted in figure 6.9 and discussed in the following section.

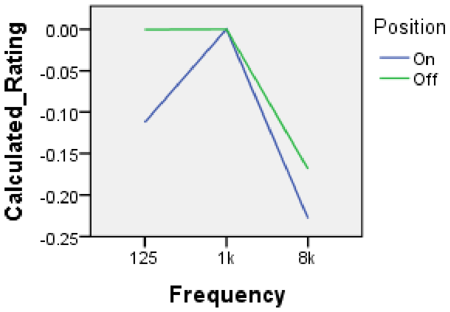


Figure 6.9: Diffuseness of each frequency band calculated with reference to the perceived diffuseness at 1 kHz.

Looking at the calculated diffusenesses of the different frequencies (figure 6.9) we see that on-centre 1 kHz was more diffuse than the 125 Hz or 8 kHz bands. Off-centre, 1 kHz was equally diffuse as 125 Hz but more diffuse than 8 kHz. Whilst the differences between frequencies are significant, the differences are small especially when compared to the effect of the ICCC.

Adjusted Comparisons

The mean inter-frequency values for each frequency and listening position can be added to the intra-frequency comparisons to give values that should be comparable across frequencies and ICCCs. These adjusted mean values are plotted in figure 6.10.

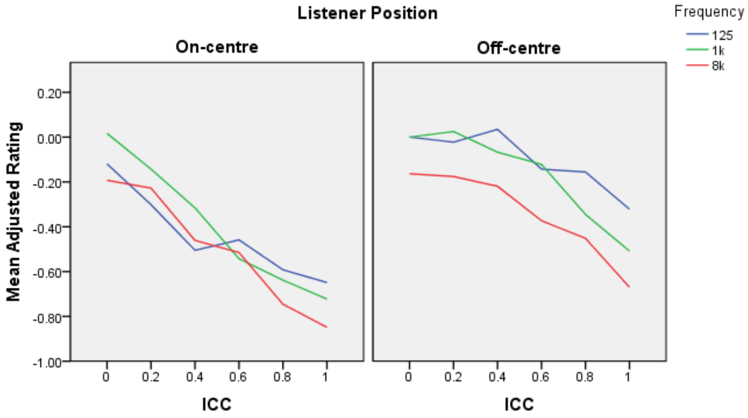


Figure 6.10: Mean intra-frequency diffuseness of each stimulus adjusted by the differences between references (from inter-frequency comparisons)

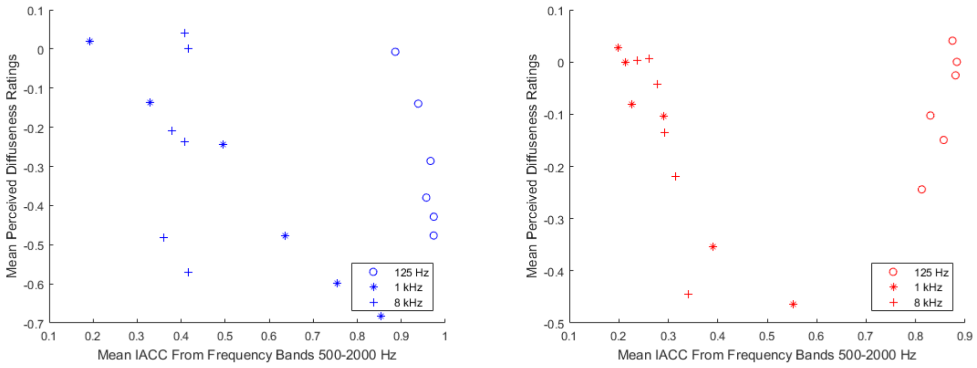
Even with the differences between the references applied to all the stimuli, there are no clear trends that show any differences between the different frequencies. Off-centre 8 kHz is less perceptually diffuse than the other two frequencies although this is not the case on-centre.

6.9 Comparison with Objective Metrics

In this section the results are compared to the values obtained from the objective metrics of IACC–which is the basis for most estimates of diffuseness or envelopment–and the spatial coherence which was found to correlate well with the data in experiments 2 and 3.

IACC

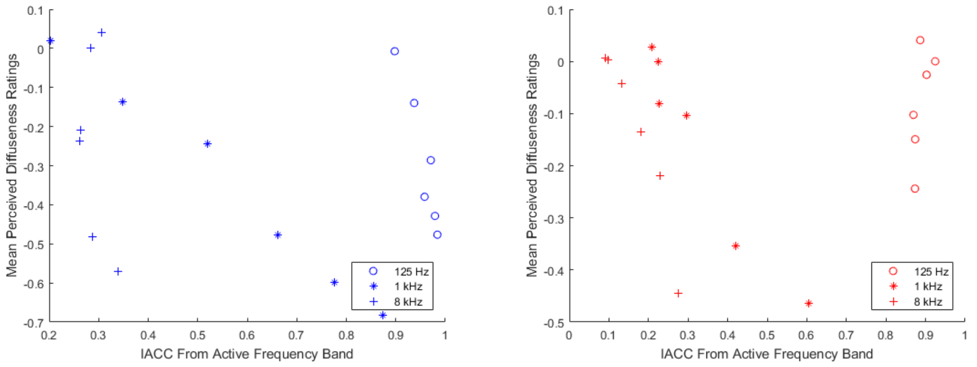
As for experiments 2 and 3, the IACC was calculated from binaural recordings of the stimuli made using a Neumann dummy head (KU 100). Figure 6.11 shows the IACC averaged



(a) IACC$_3$ On-centre

(b) IACC$_3$ Off-centre

Figure 6.11: IACC averaged across octave frequency bands 500 Hz, 1 kHz and 2 kHz.



(a) IACC$_{Active}$ On-centre

(b) IACC$_{Active}$ Off-centre

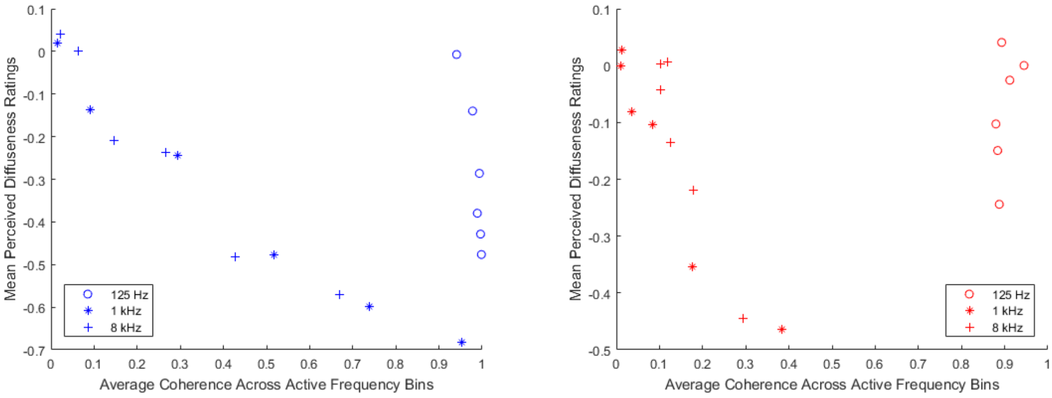
Figure 6.12: IACC of the active frequency bands.

across the central frequencies of 500 Hz, 1 kHz and 2 kHz. Figure 6.12 shows the IACC value relating to the active frequency band. This active frequency band is the octave band in which the stimulus falls considering this experiment used narrow band stimuli.

At 1 kHz there is good correlation between the IACC and the perceived diffuseness. However at both high and low frequencies the prediction is poor. At low frequencies this is to be expected. The correlation between points is known to be high at low frequencies where the wavelength is long relative to the head width. However the fit to the data is also poor at high frequencies with the IACC seemingly independent of the ICCC. In the off-centre listening position there is a small amount of correlation between the subjective data and the IACC although there is no consistency between the predictions at different frequencies and the perceived diffuseness.

Spatial Coherence

All stimuli were recorded using microphones separated by 17 cm at the listening positions. The spatial coherence was calculated between these points and the function averaged across



(a) On-centre

(b) Off-Centre

Figure 6.13: Spatial coherence averaged across the relevant frequency bins.

all frequency bins that relate to the bandwidth of the stimulus. For example, for the 125 Hz stimuli, the single frequency bin at 93 Hz is the only relevant frequency bin. For 8 kHz, the frequency bins are narrower and therefore more coefficients are relevant to the 1/3 octave bandwidth. For example, a third octave centred at 8 kHz would include 19 frequency bins from 71255 Hz to 8906 Hz for a 48 kHz sample and an FFT length of 512 samples. The coherence values for each of these ranges of frequency bins were averaged to give a single coherence-based diffuseness estimate for each stimulus. These values are plotted against the perceived diffuseness in figure 6.13.

As with the IACC, the coherence at low frequencies is always high independently of the difference in perceived diffuseness. However, unlike the IACC, the coherence accounts for both listening position and frequency–for the two high frequencies at least– and gives a prediction for the diffuseness that more generally valid. It is hypothesised that the narrower frequency resolution at high frequencies allows the metric to distinguish the differences in diffuseness at high frequency as well as for the mid frequencies.

However, at low frequencies there is no obvious relation between any objective metric, and the perceived diffuseness.

6.10 Summary

Even when applying the differences between the references (the calculated inter-frequency diffusenesses) to the intra-frequency comparisons, the difference in diffuseness between frequencies is not obvious. It is therefore suggested that to ensure high perceived diffuseness, low correlation should be used at all frequencies (at least 125 Hz to 8 kHz) and this is especially important on-centre.

This independence of the perception on frequency highlights the limitations of metrics such as Interaural Cross-Correlation Coefficient (IACC) and spatial coherence for measuring low frequency envelopment where the perception of variable diffuseness is strong to a much lower frequency than these metrics are effective. However, the spatial coherence does show a good fit to the data for both 1 kHz and 8 kHz as well as both listening positions where the IACC is only effective at 1 kHz.

Chapter 7

Development of a Robust Metric for Diffuseness

7.1 Overview

In this chapter, the results of the previous experiments are used to create a metric that can reasonably be expected to predict the perceived diffuseness under a range of conditions. The following section reports a range of challenges. These challenges refer to features of the coherence and test material that should be considered to ensure the metric is robust to a range of test materials. The final metric is designed to be robust to these challenges and the algorithm is described in detail in section 7.3. This metric is refined using the data from the subjective experiment in chapter 8 and tested in chapter 9.

7.2 Challenges: Spatial Coherence as a Metric for Perceived Diffuseness

The spatial coherence has shown to have a high correlation with the perceived diffuseness in a range of situations from the results of the previous experiments.

Although the coherence has shown promise it inherently has issues that potentially limit it's usefulness as a perceptual metric. In this section these issues are highlighted and possible solutions are proposed. These solutions are combined into a single algorithm for predicting diffuseness. The resulting metric will therefore provide very similar predictions for the stimuli tested so far, but also be reasonably robust to stimuli that exploit known limitations of the coherence. The issues are divided into the following subsections and finally the metric is presented.

7.2.1 Linearity with frequency

The first issue is that the coherence is a function of frequency that is linear with frequency. meaning each frequency bin contains the same range of frequencies. Human hearing in contrast is logarithmic with greater frequency resolution at low frequencies where the critical bandwidths are narrower. Averaging the coherence across all frequency bins gives excessive weighting to the more numerous high frequency coefficients.

This bias can be avoided by dividing the linear FFT bins between logarithmically selected groups using third octave band weighting or barc scale weighting. This gives a logarithmic set of values that approximate an individual coherence value for each of the critical bands in human hearing. Averaging across these third octave bands provides an average coherence that is not biased towards high frequencies. The barc scale has the added advantage that at very low frequencies the width of the bands is larger than for third octave bands. This is useful if the FFT window is short and therefore there are not many low frequency bins in the FFT to average between.

7.2.2 Frequency Content of Input Signal

The second issue is that the coherence is independent of the spectrum of the input signal. Even if there is no signal at a given frequency, the coherence will be a value between 0 and 1 and will be calculated on the background noise on each channel. In experiments 2 and 3 pink noise was used and so averaging across the full range of FFT frequency bins was appropriate. However, when narrow frequency bands are used, only the coherence values of the active frequency bands should be used. This was done manually in experiment 4 but should be automatic in a coherence-based metric. The solution is a weighting of the coherence based on the spectrum of the frequency response of the input.

A first simple suggestion is to use a noise gate methodology to ignore the coherence in frequency bands that are much quieter than others. A threshold of -15 dB relative to the loudest bin could be used as a criterion for selecting/ ignoring the coherence in a frequency band. In this research so far there have been two stimulus spectra used: one is full bandwidth pink noise (flat in the third octave domain) and the other is 1/3 octave band noise. This binary on/off methodology for selecting frequency band works well in these cases but maybe a more elegant solution for the future might be one that weights the coherence values based on the perceptual loudness of bands.

7.2.3 High Coherence at Low Frequency

When using loudspeakers and microphone spacings akin to the width of the head, the coherence is always high at low frequencies as the wavelength is shorter than the microphone spacing.

When combining coherence coefficients into a single value, if low frequencies are included then signals with more low frequency content are predicted to be more coherent overall. If low frequencies are ignored entirely then the the variable diffuseness at low frequencies are ignored, and low frequency only signals are not assigned a diffuseness value. This is known from experiment 4 to not be suitable as low frequencies do play a role in the perception of diffuseness.

The solution to this issue is not trivial. This is an inherent limitation of the coherence as a metric of diffuseness. The simplest solution–and the one generally employed in IACC–is to ignore any frequency bands for which the metric is known not to work. The optimal answer would be to find a metric that also correlates with perceived low frequency diffuseness and use a crossover to swap between the estimation methods. Although at this time it is not clear what form this might take.

A holistic and non-rigorous approach could be to weight the lower frequencies as less important. Therefore, for signals with only low frequency content, the metric tries its best to estimate the diffuseness but is likely underestimated. But any signal that includes higher frequency content or wide bandwidth content would be estimated based on the coherence at higher frequencies only avoiding the bias of the highly coherent lower frequencies.

Another alternative is to take the coherence as related to that of a theoretical diffuse sound field given by $(\sin(kR)/kR)^2$. However this is also not ideal as there are sound fields that have a lower coherence than the theoretical diffuse sound field and would therefore be given negative values.

The method for handling low frequencies in the metric presented at the end of this chapter, is to apply a weighting based on the theoretical minimum spatial coherence for a sound field composed of plane waves given by $\cos(2\pi Rf/c)^2$ below 500 Hz. This is derived in detail in chapter 10. The weighting means that the minimum coherence becomes zero and the maximum stays as one. There are cases where this expansion might lead to negative coherence values where the assumption of incoming signals being plane waves is not correct and therefore the weighted coherence should be constrained to between 0 and 1.

7.2.4 Perceived Diffuseness of Silence?

The coherence is independent of level. Therefore what should the predicted diffuseness be in the case of silence? A good example is listening to music in a room with very quiet background noise. Based on the coherence, when the music stops, the diffuseness is then estimated based on the background noise of the room and the measurement noise. This is typically estimated to be highly diffuse especially considering that measurement noise is generally uncorrelated between channels. On one hand, background noise such as air conditioning may well be perceived as diffuse, however, in the context of spatial audio and usable tools for production, the interesting diffuseness estimation would be that of the music and not the background noise. For this scenario, a gate should be applied to the perceived diffuseness estimation based on the overall level of the programme material. However, this leads to missing data when there is little or no audio. If the goal is to visualise the diffuseness then the diffuseness should only be displayed when the loudness is assumed to be sufficiently high that the diffuseness estimate is associated to the main programme material and not background noise.

An arbitrary cut-off of level of -50dB relative to the maximum peak level of the audio could be considered quiet enough to be not part of the main content and therefore the coherence of any windows that quiet would be irrelevant for any diffuseness meter.

This is similar to the frequency dependent weighting based on the frequency content of the material and could be combined into a perceptual loudness based weighting. However, this issue is related to the macro dynamics of an audio scene rather than the instantaneous frequency response of the scene and so is worth separating.

7.2.5 Window Length

Coherence is calculated in the STFT domain. Therefore the FFT window length needs to be chosen carefully. Consider the following example: noise is played through one loudspeaker of a stereo system and then delayed and played through the other loudspeaker. When this delay is very short a single small source (phantom image) signal is perceived. If this delay is very long (several seconds), then the human hearing system is unable to distinguish this coherent pair of noise signals from uncorrelated noise signals as the auto-correlation of the signals is very short and the delays very long.

The calculation of the coherence does this naturally. A delayed noise signal will be coherent if the delay falls within the length of the FFT window and will be measured as incoherent if the delayed version falls outside the FFT analysis window. Therefore the window length should be chosen to reflect the perceptual time windowing of the human hearing system. The experiments carried out so far are not heavily affected by the coherence window length. The reverberation in the listening room will be affected by the choice of coherence analysis window length as will moving off-centre. However in both these cases the choice of window length has only a small effect on the predicted diffuseness values and even less effect on the quality of the fit to the data as it forms a general bias rather than an error. Therefore the coherence should be considered a valid prediction method for diffuseness even considering this previously unaccounted for variable. However, to make the coherence generic, the analysis window length should match that of the perceptual window length. The next experiment described in chapter 8 was used to tune this parameter subjectively.

7.2.6 Microphone Spacing

The spatial coherence has been shown to fit the data well for the experiments conducted so far. The coherence between points is a feature of the sound field. However, as human beings, the sound field is only perceived through the signals at the ears of the listener.

Looking at the coherence function equation for a sound field composed of uncorrelated plane waves from section 4.7,

$$\gamma_{xy}(f) = \left(\frac{1}{\sum A} \sum_{n=1}^{N} A_n \cos(kR\sin(\theta_n)) \right)^2 + \left(\frac{1}{\sum A} \sum_{n=1}^{N} A_n \sin(kR\sin(\theta_n)) \right)^2 \qquad (7.1)$$

This periodic function is highly dependent on the value R which is the spacing between microphones. Because this function is periodic, this implies that any loudspeaker layout is not diffuse because at given frequencies, even uncorrelated signals will lead to a high coherence between points. Equally, a pair of uncorrelated signals will have frequencies at which the coherence is low and frequencies at which it is high. This is the case when the microphones are replaced by the ears of a listener. However, the spacing R, although approximately equivalent to the interaural distance, does not necessarily give the same time delay between the signals as the ITD. For the spaced microphone case, the factor that determines the coherence function is the delay between the two microphones for an incoming plane wave given by $\tau = R\cos(\theta)/c$. This determines whether the uncorrelated plane waves will add to give a coherent microphone signals or incoherent microphone signals. When the microphones are replaced by the head of a listener the time delay that determines whether the uncorrelated signals will add to be coherent between the ears or incoherent is instead the ITD.

If the stimulus material is wide bandwidth noise, the entire periodic coherence function is averaged to give a single value. Because the function is periodic, the average is more related to the “duty cycle" of the function and not heavily biased by the specific microphone spacing. However, if the frequency content of the material is considered, then the particular frequencies of the peaks and notches in the coherence function becomes important and should be accurately predicted.

This leads to two possibilities. Firstly, the coherence is calculated using binaural recordings which would include perfectly the ITD as the delay between microphone signal, or use a microphone spacing that approximates the ITD.

Whilst the first option seems the most practical, it has been shown that head shadowing effects seem to heavily bias the prediction of diffuseness, especially at high frequencies, and lead to a low diffuseness prediction regardless of the stimulus. Conversely, the coherence between microphone signals was shown to work well for at both 1 kHz and 8 kHz in the previous experiment. Therefore in the metric presented at the end of this chapter, a spaced microphone technique is recommended but using a microphone spacing that relates to the ITD.

7.3 The Coherence-Based Metric

In this section the working of the metric is described and the factors from the previous section implemented. The metric has been implemented in python using the VISR framework developed as part of the S3A project on object-based audio (Franck and Fazi, 2018).

The flow diagram for the visr implementation is shown in figure 7.1. The python code is shown in Appendix A and is available from `https://git.soton.ac.uk/mpc1r13/Diffusness_Meter.git`.

The sound field is sampled using a pair of omnidirectional microphones separated by 0.17 cm. The signal is normalised giving an approximate reference maximum loudness to allow the main material to be differentiated from the background noise at a later stage. The audio is then split into windows. The FFT length is 512 samples at 48 kHz sampling frequency was chosen based on the results of experiment 5 in chapter 8. Multiple FFTs are required to obtain a good estimate for the coherence and so the calculated FFTs are placed in a circular buffer. The use of more windows allow for a more accurate estimate of

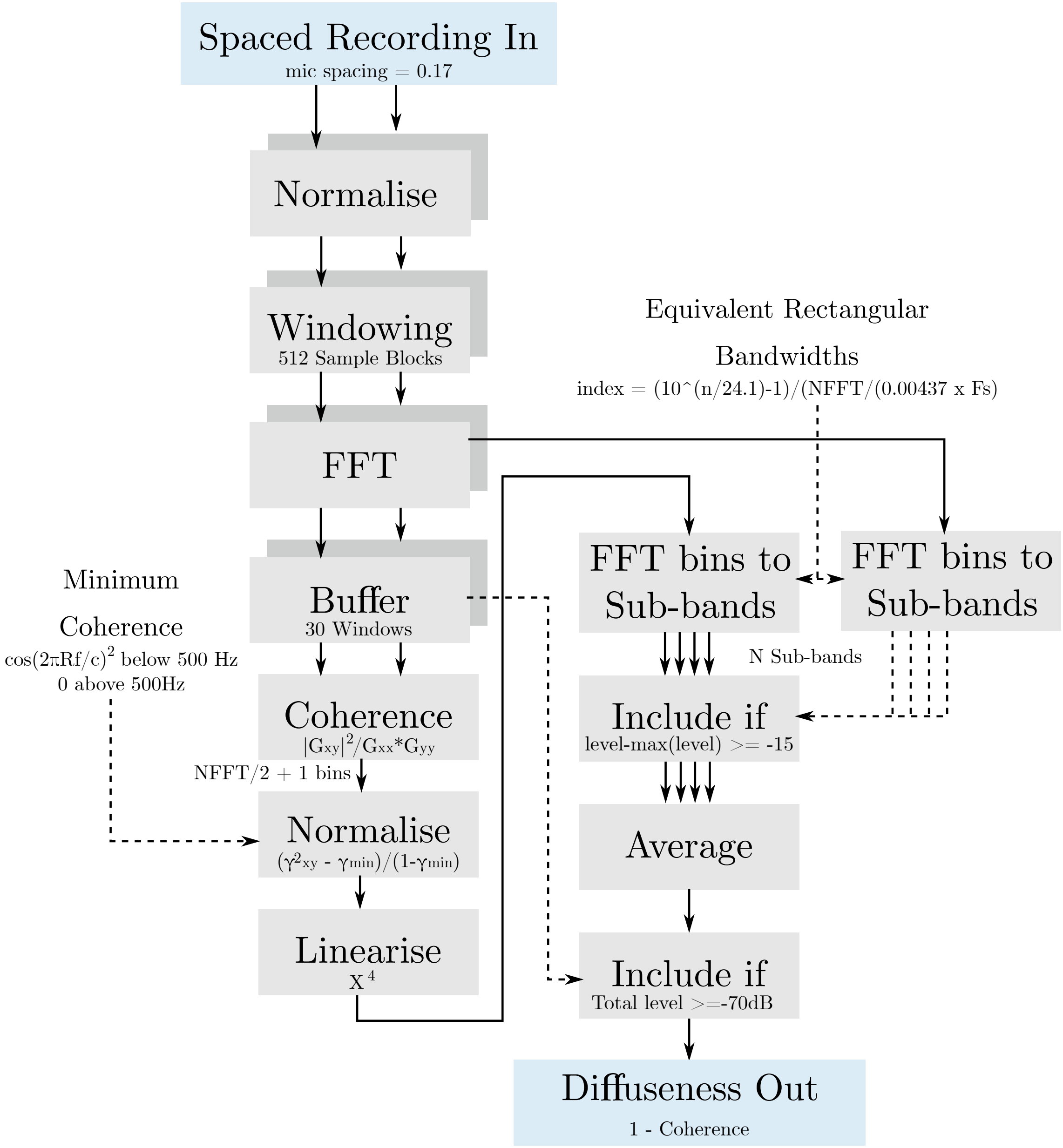


Figure 7.1: Flow Diagram describing the perceived diffuseness prediction metric.

the coherence for a stationary signal but is less suitable for dynamic material with around 30 FFTs appearing to be a good middle ground.

The coherence between the two microphone signals is calculated for each input window by normalising the squared cross-spectrum by the two auto-spectra. At this point the coherence is known to be high at low frequencies. Therefore the coherence is normalised such that the coherence ranges from zero to one at all frequencies. The minimum coherence is calculated based on a pair of plane waves arriving from ±90° below 500 Hz and zero above 500 Hz. The coherence has been found to not be linear with perception. Based on the results of the following experiment in chapter 8, there is a non-linear relationship between the coherence and the perceived diffuseness. In the data for that experiment, a power curve was used to fit the data and had the form $ax^b + c$. However the listeners did not always use the full range of the scale. Therefore, the exponent of the function b can be used to linearise the relation between the coherence and the perceived diffuseness but also normalise the range of possible coherence values to between zero and one. Zero for a sound source that is composed of plane waves with the minimum possible coherence between the microphone signals, and one for a single plane wave. Therefore in the metric, an exponent of $b = 4$ is used to perceptually linearise the coherence.

Each of the 257 frequency bins has a coherence value that ranges from zero to one. These FFT bins are linearly spaced in frequency and are then placed into sub-bands that are equally spaced on the equivalent rectangular bandwidth scale that approximates the critical bandwidth scale by adding the levels and averaging the coherences.

Each of the FFT bins and therefore each of the sub-bands has a coherence value. This is independent of the level of the signal in that frequency bin or sub-band. Therefore, for signals with a sparse frequency response, the coherence of the inactive frequency bands would also be be included in the diffuseness estimate. To avoid the coherence of inactive frequencies affecting the diffuseness estimate, a threshold of -15 dB relative to the most active (highest level) sub-band is used to determine which bands should be included when averaging across frequency. These coherence values are then averaged to give a single value across all frequencies.

It was found that recorded background noise would be rated as highly diffuse whereas digital silence would be rated as extremely not diffuse. In most cases it would be better if these cases were displayed identically. Therefore, a second threshold of -70 dB relative to maximum digital level–remembering that the input audio was normalised– is used to distinguish between the main programme material and the background noise. It is this threshold that allows both of these cases to generate the same diffuseness predictions. This does lead to missing diffuseness estimate values, but should mean that all the diffuseness values relate to the relevant programme material.

7.4 Summary

Experiments one to four all used very similar material. Material that is both stationary and of a known bandwidth. From these first experiments, the spatial coherence between spatially separated points was found to have a good fit to the majority of the data. However, the spatial coherence is limited as it only describes the similarity between the signals but is agnostic of factors such as the frequency content of the signal. In order to predict the diffuseness for a wider range of potential materials, a coherence-based metric was developed that will provide the same good fit to the existing data, whilst also being more generally appropriate for a other test materials.

The remaining issue is that for sound fields such as reverberation, the resulting spatial coherence value is highly dependent on the choice of analysis window length. Therefore, experiment 5 was conducted to subjectively assess the optimal analysis window length to be used in the metric.

This coherence-based metric is then tested in chapter 9 with both a subjective experiment and data taken from the literature.

Chapter 8

Experiment 5: Choice of Window Length for Metric

8.1 Overview

Coherence based metrics look at the phase between subsequent windows. If the phase difference between channels is consistent then the coherence is said to be high. However, depending on the length of the windows under test the coherence value can vary greatly. Figure 8.1 shows how the mean coherence value changes as the delay between the two signals is increased. When the delay is zero, both signals are identical and the coherence is high. As the delay increases, then the coherence decreases. When the delay is longer than the analysis window length, then the coherence is zero even though the signals are identical.

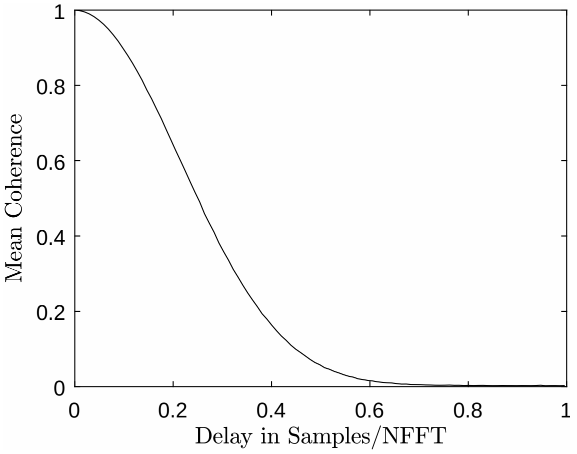


Figure 8.1: Mean coherence as a function of the delay between identical noise signals with reference to the analysis window length (NFFT).

In the context of perceptual audio, this is a good thing, as a delayed noise signal will sound uncorrelated with the non-delayed version, and, if the delay is sufficiently large, would not be differentiable from truly uncorrelated signals. Likewise two signals with very short delay will be similar to having highly correlated sources. Therefore the time window of the metric should be chosen to best match the perceptual time windowing.

For the stimuli from the previous experiments, the analysis window length has not had a large effect on the coherence. However, to ensure a robust metric, the window length should be chosen based on subjective results for stimuli for which the analysis window length is critical. This experiment uses stimuli for which the coherence varies with analysis window length to find the optimal window length to fit the coherence to the subjective perceived diffuseness. This should additionally allow the metric to be of use in architectural acoustics where the length of the reverberation has a large effect on the perception of diffuseness but has no effect on the coherence if the measurement window is always too long (Jacobsen and Roisin, 2000).

8.2 Stimuli

When measuring the coherence from previous experiments, the result was not dependent on the analysis window length. This is because there is no time dependency to the correlation. The method used to vary the ICCC was not time dependent and was just a linear sum of uncorrelated signals.

In reverberation, all the reflections are identical to the original signal but with a different level and phase. The coherence is therefore dependent on the analysis window. If the analysis window is short, then the coherence will appear to be low because some of the “delays” from the difference in path length for reflections are large with respect to the analysis window length. If the analysis window is long then all the “delays" are short relative to the analysis window. Figure 8.1 shows how short delays with respect to the NFFT length leads to a high coherence.

In order to subjectively optimise the analysis window length, the stimuli in this experiment need to be chosen such that the measured coherence is dependent on the analysis window. This is shown in more detail in section 8.7.

The stimuli were generated using noise bust decorrelation filters

As with previous experiments, pink noise was chosen as the material as it is stationary, the energy is logarithmic with frequency–like human hearing–and can be completely uncorrelated.

A single pink noise source was decorrelated for each loudspeaker by convolving it with different, uncorrelated white noise bursts of variable length. This type of stimulus has coherence values that are highly dependent on the analysis window length and It was decided to use both exponentially-windowed white noise and rectangularly-windowed white noise as the decorrelation filters. The exponential decay approximates reverberation whereas the rectangular noise burst were used in way similar to Kendall (1995) where they are used specifically as decorrelation filters.

The total energy of the stimuli was maintained by scaling the amplitude of the decay curve to match the energy of the rectangular window. However, there remains a build up of the energy over the length of the decorrelation window at the beginning of the sample and equally a decay at the end. These build ups and decays were removed leaving a stationary noise signal with constant amplitude. Although the length of a reverberant tail in response to an impulse may well be a salient point of the perception of diffuseness in a concert hall, for the purpose of this experiment, stationary signals are more appropriate and more easily compared.

In previous experiments the spatial coherence was found to vary with loudspeaker layout when using uncorrelated signals. Therefore the spatial coherence is dependent not only on the length of the decorrelation filters, but also on the loudspeaker layout used. Therefore by including multiple loudspeaker layouts, then both of these effects can be compared directly. The loudspeaker layouts of 5.0 and 12/12/13 were the lower and upper diffuseness layouts used in experiment 3 and so were also used in this experiment.

The white noise burst lengths were chosen following the results of informal listening shown in table 8.1. A range of decorrelation filter lengths were listened to and compared. A marking of X denotes stimuli that were not diffuse and equally diffuse as having no decorrelation applied. The stimuli marked with a Y were highly diffuse and near indistinguishable from being truly uncorrelated (by comparing the decorrelated stimulus with signals generated with independent noise generators.) The filter lengths in bold were chosen for the experiment a they cover a good range of perceived diffusenesses for both layouts and window shapes.

8.3 Listener Response Methodology

Once again a MUSHRA style listening test methodology was chosen allowing for quick comparison of multiple stimuli especially where accurate relative ratings are more relevant than unbiased absolute values for the diffuseness.

Decorrelation filter length (seconds)	Layout			
	12/12/13		5.0	
	Exponential	Rectangular	Exponential	Rectangular
0.001	X	X	X	X
0.002	X	X	X	X
0.005	**X**	**X**	**X**	**X**
0.01	X	X	X	X
0.02	**X**		**X**	**X**
0.05	X		X	
0.1			**X**	
0.2				
0.5		**Y**		**Y**
1	Y	Y		Y
2	**Y**	**Y**	**Y**	**Y**
5	Y	Y	Y	Y
10	Y	Y	Y	Y

Table 8.1: Informal pretest to choose decorrelation filter lengths.

As in previous experiments, uncorrelated stereo was included as a low anchor. In this case the layout 12/12/13 with uncorrelated pink noise was used as a high reference as this is the has the lowest possible inter-channel cross-correlation.

Uncorrelated pink noise was generated using the dsp.ColoredNoise object in Matlab. These signals are generated using independent random sequence generators to ensure a correlation coefficient of 0 (for a long sequence).

The user interface implemented in Matlab is shown in figure 8.2. The buttons across the top of the page would select and play a 2 s burst of the stimulus. The order of the stimuli was randomised for every listener and repeat.

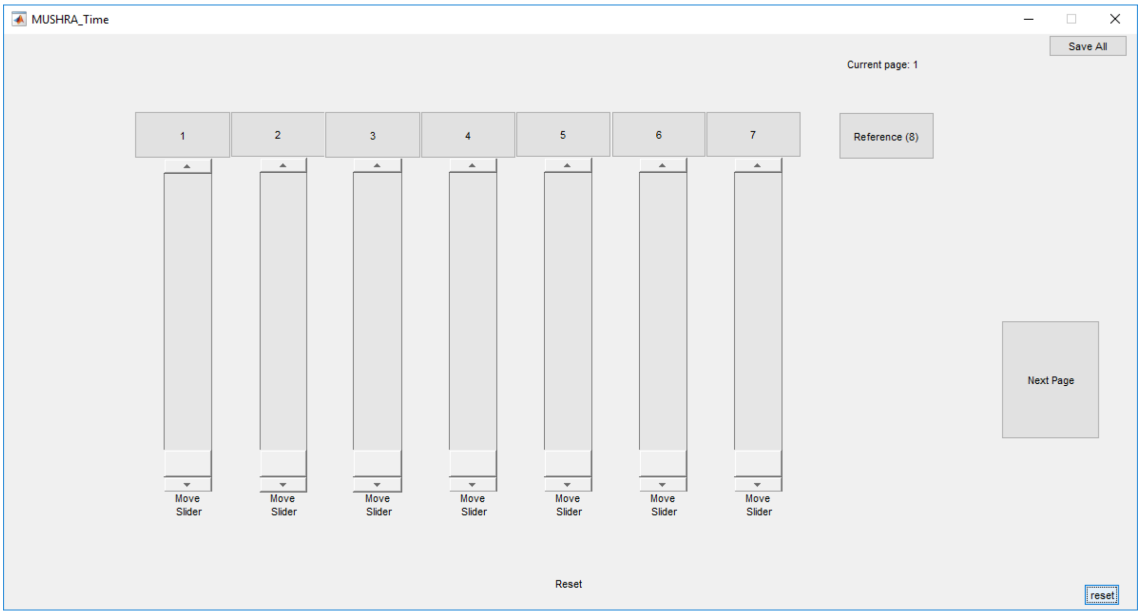


Figure 8.2: MUSHRA-style user interface used in experiment 5 implemented in Matlab.

8.4 Reproduction System

The listening test was performed using the same reproduction system as the previous experiments. The Audio Lab at the University of Southampton has an T_{60} of 0.12 s ±0.02 s in 1/3 octave bands between 125 Hz and 8 kHz and 39 Kef HS3001SE loudspeakers mounted to the walls, floor and ceiling equalised in 1/3 octave bands from 95 Hz to 20 kHz. Below 95 Hz a -24 dB per octave roll off followed the approximate frequency response of the loudspeakers. Digital delay was applied to each loudspeaker to compensate for the differing propagation delays between the loudspeakers and the central listening position to the nearest sample (the sampling frequency was 48 kHz throughout).

8.5 Subjects

The listening test was approved by the ethics and research governance committee (ID: 18709). This experiment was sat by a mixture of 8 experienced and 8 inexperienced listeners. Experience was self reported and generally related to sitting previous listening tests and/working in audio related subjects. All were either undergraduate or postgraduate students at the University of Southampton with self reported normal hearing.

The most common responses following the experiment were that not many of the stimuli fell into the middle of the scales and most listeners opted to use the extremes of the scale. Some listeners mentioned in head localisation. All listener appeared to understand the task and felt they were able to perform the task reliably.

8.6 Results

8.6.1 Post-screening

The average standard deviation between repeats was again used as measure of repeatability. These are plotted in figure 8.3. The standard deviation between repeats shows all listeners to be sufficiently consistent when rating the same stimulus. The three repeats were then averaged for each stimulus and listener and these results are shown in the box plot in figure 8.4.

Figure 8.4 shows that listener 11 has several outliers and made ratings very different to other listeners and did not manage to detect the reference regularly. The listener may have just misunderstood the task and therefore it is reasonable to exclude their data. In all further analysis, listener 11 was excluded leaving 15 listeners.

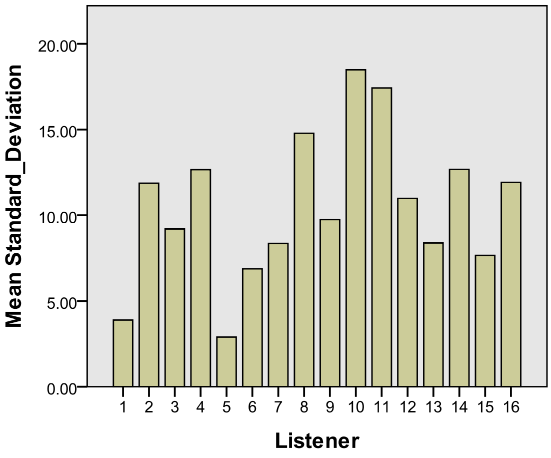


Figure 8.3: The average standard deviation between the three repeats for each stimulus.

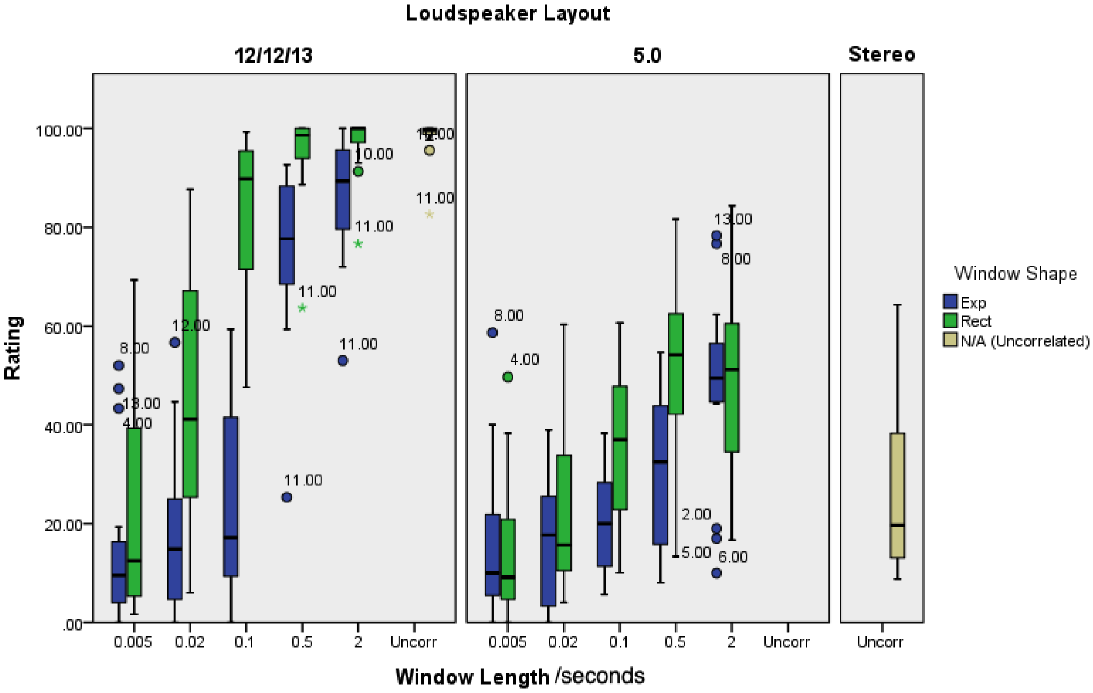


Figure 8.4: Box plots for all listeners results (averaged across the three repeats). Outliers are labelled with listener IDs.

8.6.2 Analysis

With listener 11 excluded the remainder of the data are displayed in the box plot in figure 8.5.

Listener 6 seemed to consistently rate 5.0 as not diffuse. Looking at their individual ratings it was clear they tended towards using the extremes of the scales with very little in the middle where the majority of listeners placed 5.0 as moderately diffuse.

The fixed scale and use of references means that the full range of the scale is used. This tends to mean the distribution for stimuli ranked at the top and the bottom of the

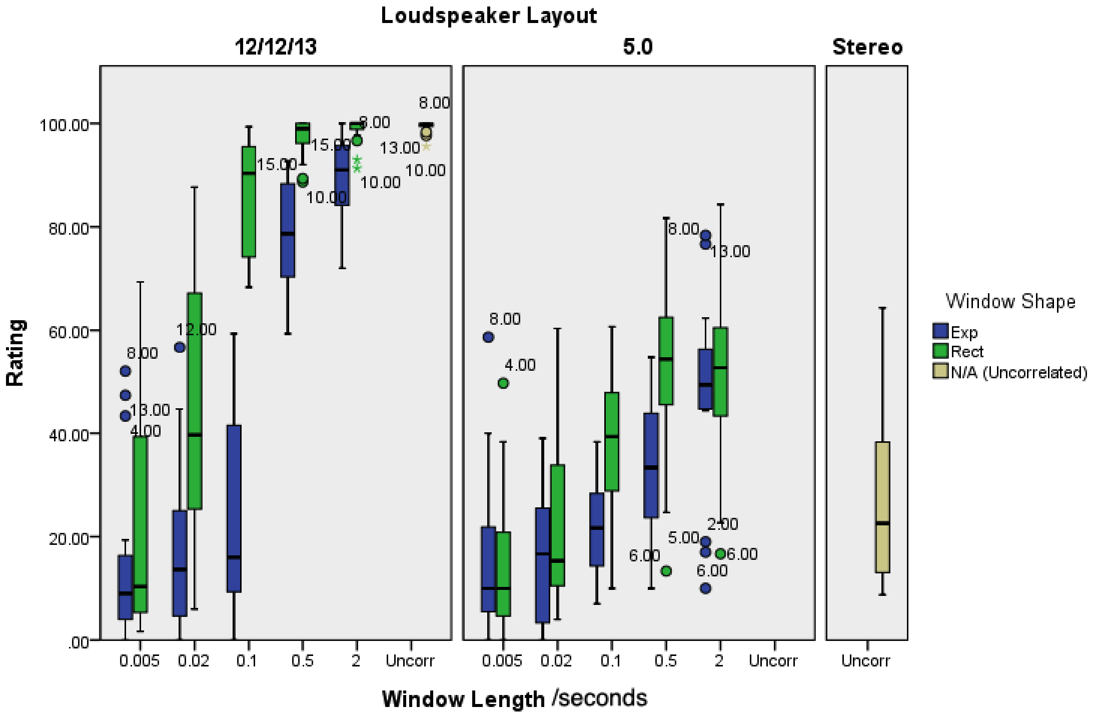


Figure 8.5: Box plots of all data for consistent and congruent listeners averaged across the three repeats. Outliers are labelled with listener IDs.

Layout	Envelope	X^2	df	p	Kendall's W
5.0	Exponential	51.646	4	$<$**0.0005**	0.861
	Rectangular	53.932	4	$<$**0.0005**	0.899
12/12/13	Exponential	38.896	4	$<$**0.0005**	0.648
	Rectangular	47.279	4	$<$**0.0005**	0.788

Table 8.2: Friedman Tests showing the effect of filter length for different layouts and filter windows.

scale is skewed and not normally distributed. Therefore Friedman Tests and Wilcoxon Signed Rank Tests are used to assess the statistical significance of any factors. These tests are shown in tables 8.2 and 8.3 for the effects of the decorrelation filter length and the decorrelation filter envelope respectively.

The filter length is significant for both loudspeaker layouts and both filter envelopes. The different filter envelopes are significantly different for all but the shortest and longest filters. In other words, for the shortest filters, there is little to no perceivable decorrelation regardless of the filter envelope. For mid-length filters, there is a statistical difference. But for very long filters, the signals are indistinguishable from uncorrelated signals, again, regardless of the filter envelope.

Loudspeaker Layout	Filter Length /Seconds	Z	p
5.0	0.005	-1.306	0.191
	0.02	-3.233	**0.001**
	0.1	-3.408	**0.001**
	0.5	-3.409	**0.001**
	2	-3.045	**0.002**
12/12/13	0.005	-0.140	0.889
	0.02	-2.104	**0.035**
	0.1	-3.171	**0.002**
	0.5	-3.408	**0.001**
	2	-0.682	0.496

Table 8.3: Wilcoxon Signed Ranks Test comparing the effect of decorrelation filter envelope for different loudspeaker layouts and filter lengths.

8.6.3 Discussion

The mean ratings for all stimuli are shown in figure 8.6.

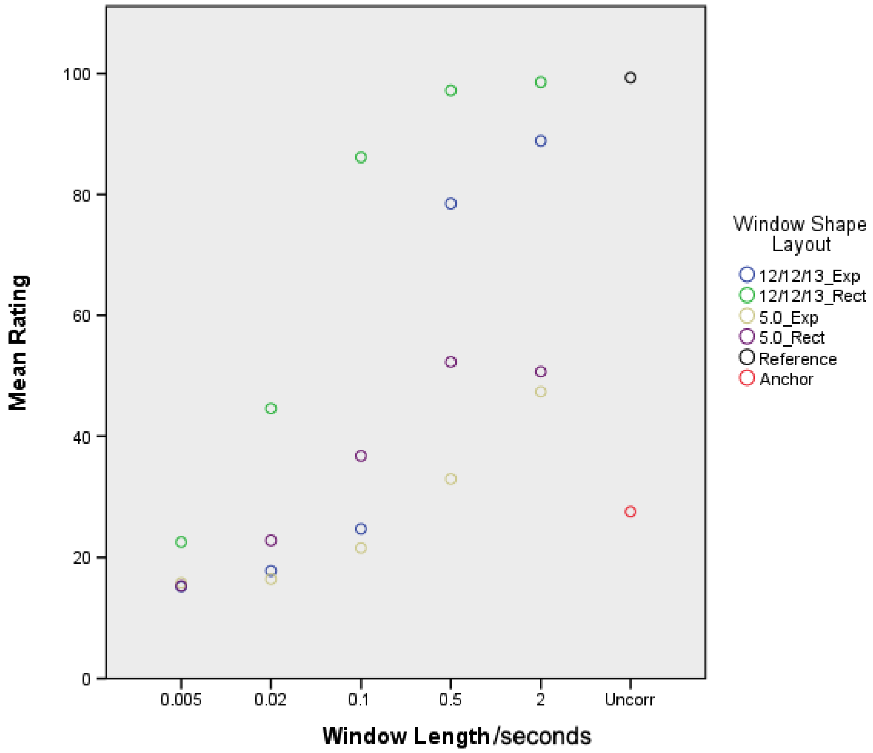


Figure 8.6: Mean diffuseness ratings for all stimuli.

On the right is the most decorrelated, on the left is the most correlated with the shortest decorrelation window. As expected we see 5.0 is limited in the maximum perceived diffuseness due to the fewer loudspeakers reaching a maximum of around 50, although this is more diffuse than uncorrelated stereo.

Layout	Experiment 2	Experiment 3	Experiment 5
Stereo	12.2061	27.8403	27.5444
5.0	52.4242	72.1944	50.6889 ✓
12/12/13	79.8485	96.9722	99.3000

Table 8.4: Differences in ratings of the same stimuli between experiments. ✓Technically this stimulus was highly decorrelated and not truly uncorrelated although it is believed this difference should be negligible based on the similarity between the ratings for the reference 12/12/13 and the 2 second rectangular decorrelated 12/12/13 in this experiment.

Also as expected the rectangular window can be shorter for the same perception of decorrelation. The long rectangular decorrelation is nearly indistinguishable from truly uncorrelated. It is possible that by providing an explicit reference, very slight changes in the colouration allow listeners to differentiate the test stimuli even if the perceived diffuseness is nearly identical. As in experiment 3, when the correlation is high (in this case due to short decorrelation filters) the number of loudspeakers has no effect on the perceived diffuseness and both layouts are always rated not diffuse.

8.6.4 Comparison of Coherence Measurements Between Experiments

Previous experiments have had similar stimuli, namely, uncorrelated stereo, 5.0 and 12/12/13. Although uncorrelated 5.0 was not included in this test, the 2 second rectangular decorrelation window should be very close to uncorrelated 5.0 considering that 2 second exponential decay decorrelation converges to the same value. This allows some comparison between experiments and reveal some interesting results.

The table 8.4 shows the mean values for uncorrelated stereo, 5.0 and 12/12/13 from experiments 2, 3 and 5.

The main difference between experiment 2 and experiment 3 can be explained by the lack of explicit reference in experiment 2 and the generally more diffuse stimuli leading to a shift in the ratings.

However, the difference between experiments 3 and 5 shows a different issue. Here there is a non-linear scaling between the different experiments. With the reference available in both experiments stereo and 12/12/13 are rated similarly in both experiments. However 5.0 is rated notably differently. This is likely due to stimulus spacing bias. For example in experiment 3 there were many stimuli rated between the diffusenesses of stereo and 5.0. In this experiment there are fewer stimuli in the same range. Therefore in experiment 3, 5.0 appears more diffuse as there are more stimuli that it is more diffuse than. This highlights how these results should be taken as relative and may not be directly comparable between experiments and they are not placed on a globally consistent scale. This is why it is important that the metric is able to fit a given curve within a single experiment but it is less important to accurately match the values of the experiments between experiments where the scaling may be different.

8.7 Tuning the Metric: Choosing the Analysis Window Length

Using coherence as a metric for perceived diffuseness means that–for the stimuli tested–the prediction is dependent on the choice of analysis window length. The optimal choice of window length is therefore the window that best fits the data. In this section different window lengths are applied during the coherence calculation, the predictions are plotted against the subjective data for each stimulus and a curve is fitted to the data.

The value of the prediction will vary with the choice of window length. However, the important factor is the relative prediction between stimuli with the absolute value easily scaled afterwards. i.e. it is less important that the average coherence of 0.5 responds to a perceived diffuseness of 50% because the perceptual scale is arbitrary. It is more important that if one stimulus is perceived more diffuse than another, then the metric reflects this regardless of whether the diffuseness is due to the loudspeaker layout, the ICCC or the decorrelation filter length. The result can then be scaled and stretched afterwards to give a prediction that is linear with perception.

Figure 8.7 shows the effect of running 2 different stimuli through the diffuseness metric using different analysis window lengths. The first stimulus is pink noise decorrelated using convolution with exponentially-windowed noise. The second shows the same layout but using truly uncorrelated noise. The stimuli decorrelated using noise tails is predicted differently depending on the chosen window. In contrast the truly uncorrelated signal does not change with the analysis window.

Figure 8.8 shows the coherence estimates made using different powers of two plotted against the perceived diffuseness. For each analysis window length, a line of fit based on a second order power curve was calculated and the fit to the data is plotted in table 8.5.

Window	Adjusted R-sq
64	0.6142
128	0.8683
256	0.8835
512	0.8967
1024	0.8305
2048	0.7627

Table 8.5: Adjusted R-sq for a second order power curve fitting to the data of the form $ax^b + c$ for different analysis window lengths.

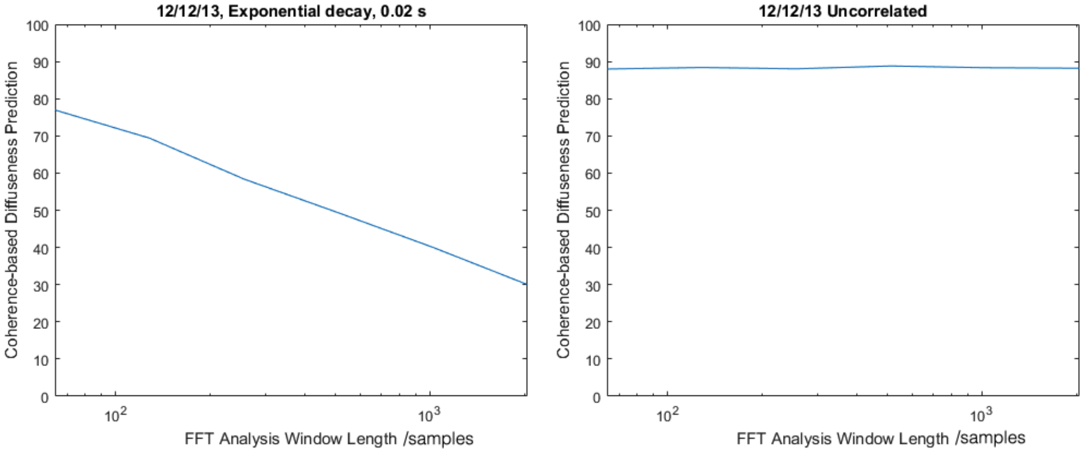


Figure 8.7: Coherence based diffuseness predictions using different analysis window lengths. On the left 12/12/13 but with a single pink noise signals decorrelated using 0.02 s exponential decays of white noise to decorrelate between loudspeakers. On the right is 12/12/13 with uncorrelated signals.

Interestingly, although the window shape is statistically significant, the coherence metric combines the effects of window length and window shape with both decorrelation windows falling on the same line regardless of the analysis window length.

From the data we see that 5.0 is approximately half as diffuse as 12/12/13. Equally we see that 12/12/13 with a rectangular decorrelation window of approximately 0.02 s or an exponential decay of approximately 0.2 s are both equally as diffuse as the highly uncorrelated 5.0. The predicted diffuseness is dependent on the analysis window if the signal is artificially decorrelated. If the signals are truly decorrelated or the analysis window is much shorter than the decorrelation window then the analysis window length is less significant. Therefore the technique to choose the correct window length is to find a window length that scales the effect of the decorrelation length to the uncorrelated effect (the layout change).

Using fit to the data as a method to demonstrate suitability of one window length over another shows 256 or 512 to be the best window lengths to use. A length of 512 samples equates to 10.6 ms considering the sample rate of 48 kHz. Figure 8.9 shows how this window length relates to the coherence between two delayed identical white noise signals.

It is interesting to compare this window length with time constants reported in the literature. For example, delays of less that 2 ms are considered highly coherent and this is similar to the ±1 ms used in the IACC measurement, although the coherence features a slower roll off with delays greater than 6 ms also considered in the coherence-based diffuseness estimate. This window length of 512 samples is shorter than the 80 ms that relates to the fusing of reverberation with the source when describing ASW and envelopment.

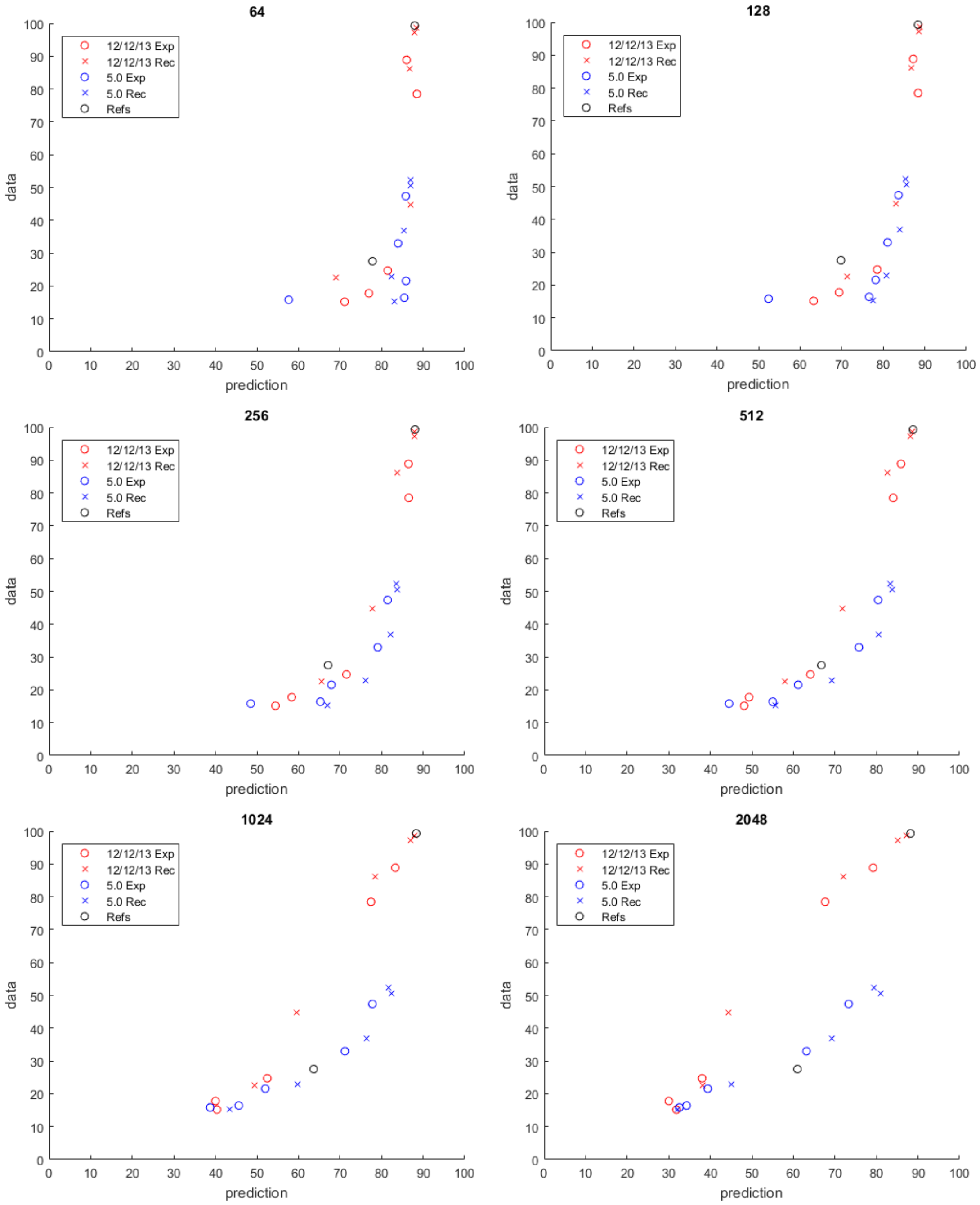


Figure 8.8: Predicted diffuseness using different analysis window lengths.

The curvature of the line represents a non-linear relationship between the coherence and perceived diffuseness. This curvature, given by the exponenet of the line of fit, can therefore be used to linearise the relationship between the coherence and the perceived diffuseness.

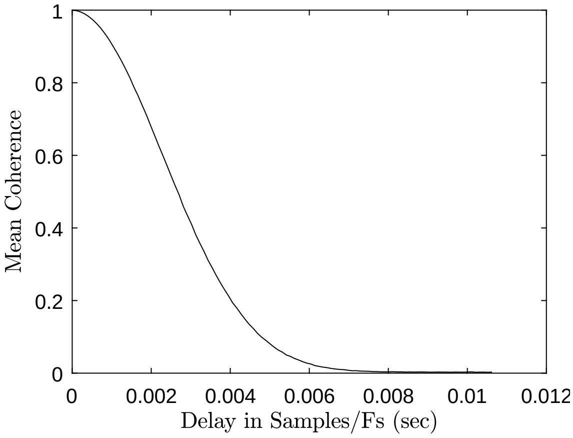


Figure 8.9: Mean coherence as a function of delay between two identical white noise signals when using the optimal analysis window length of 512 samples at a sampling frequency of 48 kHz.

8.8 Summary

This experiment was used to elicit subjective diffuseness responded for a range of stimuli for which their spatial coherence is dependent on the choice of analysis window length. This allowed the coherence-based metric to be subjectively tuned. A window length of 512 samples for a sample rate of 48 kHz was the power of 2 analysis window length that gave the best fit to the data and equates to 10.6 ms. This window length should allow the metric to have some more usability for stimuli such as reverberation. It is at this window length that the metric fits the subjective data for changes in both the length of decorrelation filter and the changes in loudspeaker layout.

Chapter 9

Testing The Metric

9.1 Overview

In this section the metric developed in chapter 7 is critically assessed to discover the limitations and shortcomings of the metric.

This was achieved through an experiment that uses material that is more similar to the type of material used in spatial audio, rather than the pink noise and band limited noise stimuli used thus far.

9.2 Experiment 6: Testing “Real” Material

9.2.1 Overview

Up until this point all the stimuli have used pink noise or band limited noise and the metrics have been tested on this type of stimulus. However in order to assess the usability of the metric in more situations, it is important that it is investigated using more different and realistic material. In this experiment a range of stimuli materials are used. Some materials are similar to that tested in previous experiments and some are chosen deliberately as critical tests of the metric and thereby highlight the limitations of the metric.

It is worth reiterating at this point the decision to use the term envelopment in this experiment rather than perceived diffuseness (Section 2.5.1). The use of complex sound scenes with multiple objects makes the use of the term perceived diffuseness ambiguous. Therefore, the term envelopment is used in this experiment in place of perceived diffuseness. On one hand, this change in nomenclature might reduce the effectiveness of any metric based on experiments of perceived diffuseness to predict the results of experimental data collected about envelopment. But on the other hand, the term envelopment should be easier

Material	Notes
Radio Drama Ambience	Distant background music, Sound effects ✓
Pop Music Background	Drums, percussion, guitar, piano ✓
Reverberation Jazz	Reverberation for piano and upright bass ✓
Sports Crowd	Recording of sports crowd ✓
Pink Noise	Uncorrelated pink noise as used in the other experiments
Rain	Static rain recording decorrelated by delaying each loudspeaker by a long delay (>10 s)

Table 9.1: Material for Experiment 6. ✓Were provided as separate mixes for the separate layouts by Francombe et al. (2018). The pink noise and rain stimuli were created using uncorrelated loudspeaker feeds depending on the loudspeaker arrangement.

for listeners to rate for complex scenes as it can be considered a higher level attribute. The differences arising from naming conventions can also be considered minimal due to the similarity with which the two attributes are described.

9.2.2 Stimuli

In this experiment loudspeaker signals that are more realistic than the noise stimuli from experiments 1 to 4 are tested. There are two categories to investigate. Firstly, the sorts of real sound fields that are highly physically diffuse with examples being rain, audience noise and reverberation. Secondly combinations of non-diffuse sound sources, specifically, multiple non-diffuse sound sources that together form a potentially surrounding sound scene.

For each material a range of loudspeaker layouts allows variable perceived diffuseness within each material to be compared to the inter-material differences.

The material for this experiment is described in table 9.1. The four materials marked with a ✓were provided by Francombe et al. (2018). In each case there were separate mixes for three standardised loudspeaker layouts (ITU-R, 2014b), stereo, 5.0, and 22.0. Each mix was engineered to be highly enveloping given the available loudspeaker layout. Therefore the stereo and 5.0 stimuli were not simply downmixes of a 22.0 recording, but instead standalone mixes for that native layout aimed at maximising envelopment. In these materials the foreground elements such as narrators and soloists were very clearly localisable and the main part of the scene. They also would appear and disappear meaning the overall envelopment would likely change over time. For this experiment, to maintain an approximately constant and static envelopment throughout the length of the clips, only the background elements were used–where appropriate–and foreground elements were removed. These foreground elements were the narrator in the radio drama; the voice and

Material	Gain in dB
Forest	-4
Pop	-6
Jazz	2
Sport	2
Pink Noise	0
Rain	-6

Table 9.2: Gains used to match the loudness between materials. The same gains were used for each loudspeaker layout.

flute in the pop music, the direct sound in the jazz recording and the commentator in the sports crowd example leaving the sources mentioned in the notes of table 9.1.

The additional two materials were instead systematically generated. Uncorrelated pink noise was generated using the dsp.ColoredNoise object in Matlab. These signals are generated using independent random sequence generators to ensure a correlation coefficient of 0 (for a long sequence). The rain was generated by dividing a single mono recording of rain into multiple uncorrelated signals by means of a long delay (>10 s) between loudspeaker signals.

Loudness between the layouts appeared fairly well matched. The loudness differences between the materials was measured objectively by comparing the 22.0 mixes with the 22.0 pink noise stimulus. Because the loudness is not static it is not possible to match all the loudnesses exactly but the following gains–in table 9.2–left no obvious differences in overall loudness between the 22.0 mixes. Because the loudness differences between layouts seem minimal the adjustment gains found for the 22.0 layout were also used for the other 2 layouts.

The combination of all six materials and three loudspeaker layouts in all combinations led to 18 stimuli.

9.2.3 Listener Response Method

In this experiment the material is different from that of previous experiments. Up to this point the focus of the research has specifically been on diffuse sound fields. These sound fields are composed of a sufficiently large number of essentially uncorrelated signals so that the sound field is perceived as a single, spatially large entity. In this experiment complex sound scenes are included that are composed of both multiple concurrent and localisable instruments/components as well as diffuse components such as reverberation. This difference necessitates the use of a different description of the feature that will be assessed subjectively by the listeners. The previous experiments used the term “perceived diffuseness" primarily when asking for responses from listeners. The description of perceived diffuseness mentioned that something that is localisable is not diffuse. However in a complex sound scene, the individual elements may be localisable but may form together an enveloping sound scene. Therefore, in this experiment the standard description of envelopment was used to avoid ambiguity when the sound field has both diffuse and non-diffuse components. The more interesting element is the overall surroundingness of the sound field rather than the listener's ability to pick out the diffuse sound components in a complex sound scene. The definition given to the listeners was,

> "Envelopment may be heard when standing and listening to the rain hitting the pavement; applause in a concert hall; atmosphere or air conditioning (room tone). Being able to localise the source of the sound will decrease envelopment. Holes (an absence of sound from a certain directions) would normally reduce envelopment. Feeling the sound inside your head or as moving/unstable would also usually be less enveloping."

It was predicted that differences between the same materials might be easier to compare than between different materials. For this reason, on each page, stimuli were grouped by material type. The order of the materials was still randomised and the order in which the different layouts were presented was also randomised for every material. However the listeners would be able to compare the three different versions of the same material within the same page. Listeners were not informed of the differences between the three loudspeaker layouts in an attempt to avoid bias.

Once again a MUSHRA style listening test was used in this experiment (figure 9.1). References were not used as it was not clear what the most enveloping stimulus might be. In previous experiments, uncorrelated stereo was included as an anchor. In this experiment, all stimuli fit across only two pages of MUSHRA and uncorrelated stereo was included as a stimulus anyway for comparison of the data with the previous experiments. Equally, including a hidden anchor, for example uncorrelated pink noise, on both pages would be obviously different to the other stimuli because the material is different and therefore give no additional information

The 18 stimuli were divided between two pages of MUSHRA each with 9 stimuli. The 9 stimuli on each page would be of 3 materials with each of the different loudspeaker layouts grouped together by colour. The order of the materials and the order of the layouts were both randomised.

Three repeats for each stimulus led to a total of 54 ratings per person.

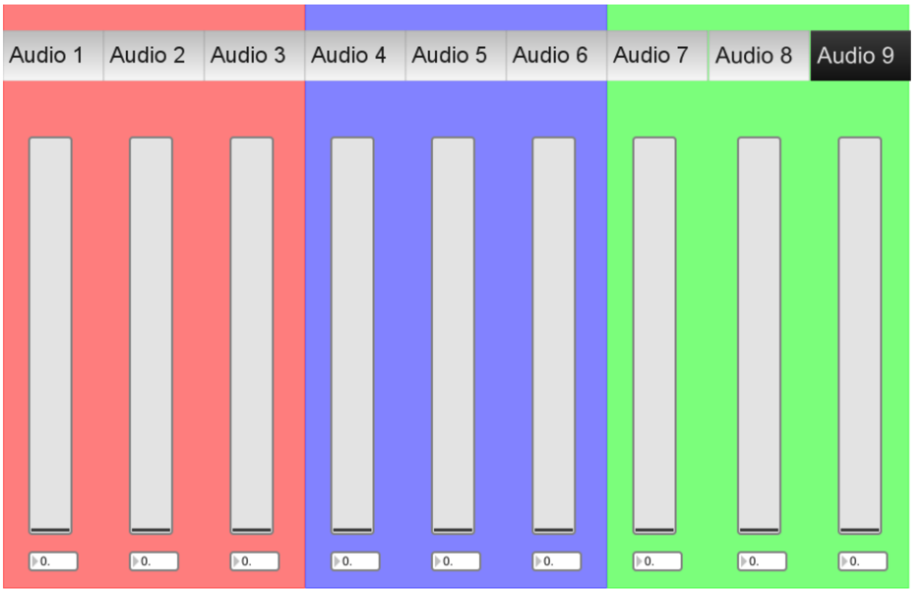


Figure 9.1: MUSHRA-style user interface used in experiment 6 implemented in Max 6.1.

9.2.4 Reproduction System

For the final time the listening test was performed in the Audio Lab at the University of Southampton with T_{60} of 0.12 s ±0.02 s in 1/3 octave bands between 125 Hz and 8 kHz. As in the previous experiments, the 39 Kef HS3001SE loudspeakers were mounted to the walls, floor and ceiling and were equalised in 1/3 octave bands from 95 Hz to 20 kHz. Below 95 Hz a -24 dB per octave roll off followed the approximate frequency response of the loudspeakers. Digital delay compensated for the differing propagation delays between the loudspeakers and the central listening position to the nearest sample (the sampling frequency was 48 kHz throughout).

9.2.5 Subjects

The listening test was approved by the ethics and research governance committee (ID: 18709). The listening test was sat by 7 postgraduate listeners from the University of Southampton. All could be be considered to be expert listeners with previous experience in listening tests with self reported normal hearing.

9.2.6 Results, Analysis and Discussion

All listeners were consistent between repeats of the same stimulus. The three repeats were then averaged together to give a single rating for each stimulus from each listener. The results are plotted in the box plot in figure 9.2

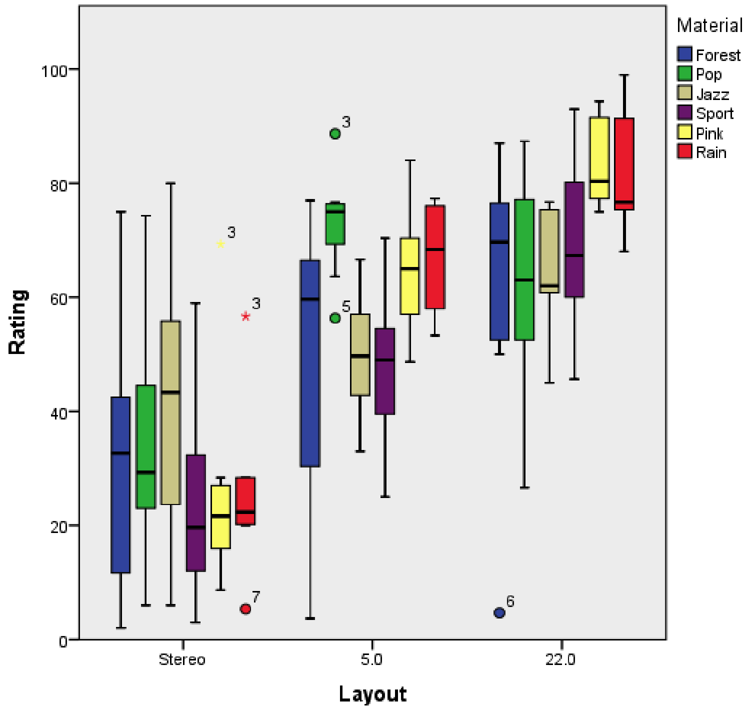


Figure 9.2: Mean envelopment rating for each stimulus and listener. Outliers are labelled with listener IDs.

Layout	X^2	df	p	Kendall's W
Stereo	9.025	5	0.108	0.258
5.0	16.5	5	**0.006**	0.471
22.0	20.517	5	**0.001**	0.586

Table 9.3: Friedman Tests showing the effect of the stimulus material for the different loudspeaker layouts.

The box plots show listener 3 to have a few outliers with Pink Noise and Rain in stereo consistently rated s not enveloping by the other listeners. Listener 6 consistently rated the forest recording as not enveloping, even at 22.0. Listener 6 commented that there were sources that were easily localised and led to a low rating. However the pop music sample was very similar and yet they rated that as more enveloping. No listeners were excluded from the following analysis.

The boxplots show the data to be skewed for some stimuli. Therefore, Friedman Tests are used to show significant differences between stimuli.

Table 9.4 shows that the stimulus material is only a significant factor for 5.0 and 22.0.

Table 9.3 shows the loudspeaker layout to be a significant factor for all materials except the Jazz.

Material	X^2	df	p	Kendall's W
Forest	10.286	2	**0.006**	0.735
Pop	9.852	2	**0.007**	0.704
Jazz	5.429	2	0.066	0.388
Sport	12.286	2	**0.002**	0.878
Pink	13.556	2	**0.001**	0.968
Rain	12.074	2	**0.002**	0.862

Table 9.4: Friedman Tests showing the effect of the loudspeaker layout for the different stimulus materials.

Discussion

Although this listening test was designed primarily to test the metric with a variety of material, the perceptual data are also interesting in and of itself. Therefore, the discussion is split into two parts, the discussion on the perceptual data, and a discussion on the suitability of the coherence for the prediction of envelopment for this range of material.

Overview of the Results

The means for all stimuli are plotted against the loudspeaker layout in figure 9.3. As expected, 22.0 was more enveloping than 5.0 which was in turn more enveloping than stereo. Interestingly, we see this was the case for all stimuli except the pop music recording, where 5.0 was rated as more enveloping. Of all the stimuli presented, 5.0 pop music was the only stimulus that was notably louder than the other layouts for any of the materials. At the time of planning the experiment, this was assumed to be negligible as the stimuli were mixed to be of the same loudness originally. However, following the results, the residual loudness differences that were overlooked would serve as explanation for this inconsistency.

For the 22.0 layout, the rain and pink noise are perceived as most enveloping. These stimuli scale to larger layouts well with truly uncorrelated signals from all loudspeakers. In some of the other materials, not all the loudspeakers are used depending on the mix. Conversely, in stereo, the pink noise and rain materials are amongst the least enveloping. Although the loudspeakers have a lower degree of correlation in the pink noise and rain materials, they are perceived as less enveloping. A possible explanation is that the static nature of these stimuli make them easier to localise to the front. Conversely material such as a bird tweeting does not give enough information to distinguish the correct position on the cone of confusion. Therefore, with no knowledge of the loudspeaker layout used, the ambiguous position of fleeting sounds could create a greater sense of envelopment. A second possibility is that the rain is expected to be all around and surrounding and therefore when it appears from only the front it is considered less diffuse than expected. The pop music

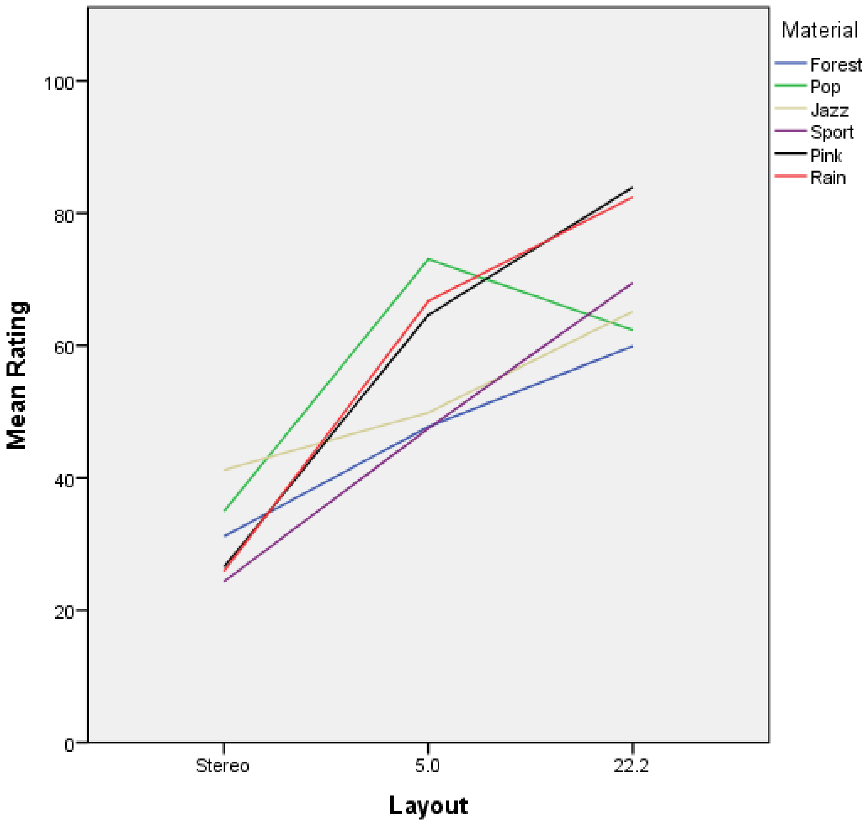


Figure 9.3: Mean envelopment scores for all stimuli.

in comparison is expected to be in the front of the listener and therefore it is possible the stereo sounds as enveloping as expected and by extension, more enveloping than the rain. A third possibility is that the wide band static nature of the rain and pink noise stimuli highlights the room effects more strongly than the more dynamic programme materials. Factors such as comb filtering are far more obvious with noise like stimuli than with the dynamic materials. With more loudspeakers this becomes less obvious and therefore the more obvious comb filtering may lead to a assumption of fewer sources and therefore a lower amount of envelopment.

The only material that was not rated significantly differently for the three loudspeaker layouts was the jazz. This was interestingly rated the most enveloping in stereo but the least enveloping for 22.0. It is possible this is a combination of 2 factors. Firstly, the reverberation in the stereo recording gives a sense of space (even with only 2 loudspeakers) leading to a high rating. However, with loudspeaker layouts with rear channels, these channels are only used for reverberation with the main instruments remaining in the frontal area. In this way the distribution of sources varies less for the Jazz than for the other materials. This leads to a lower rating at 22.0

Comparison of the Results with the Coherence Metric

The stimuli were recorded using a pair of A B&K free-field microphones type 4190 with B&K preamplifiers type 2669 separated by 17 cm at the listening position. These signals were run through the diffuseness metric described in chapter 7 and the results plotted against the subjective envelopment scores in figure 9.4.

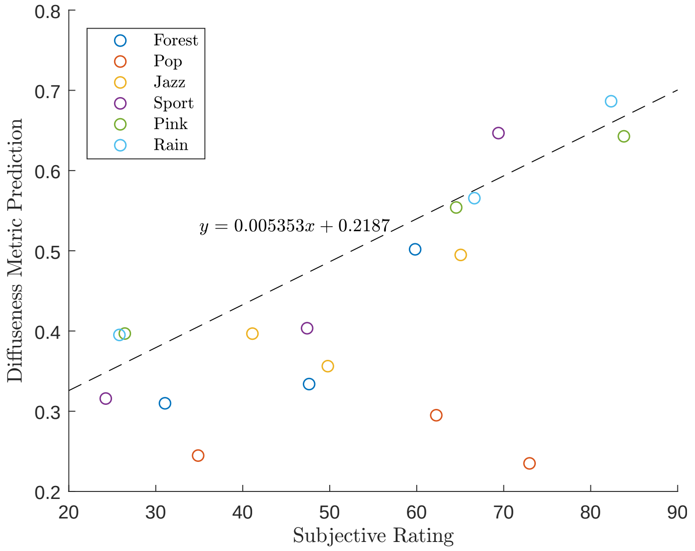


Figure 9.4: Mean envelopment scores for all stimuli plotted against coherence-based diffuseness prediction. Line of fit to the noise like stimuli (Sport, Pink Noise and Rain) with adjusted r-squared fit of 0.8967.

At first glance the correlation is poor. Looking at specifically the pink noise and rain samples we see the positive correlation expected as the number of loudspeakers increases and the perceived envelopment increases. For the pink noise this is the trend observed in experiment 2. The rain is a very similar material. It is noise like and highly uncorrelated between loudspeakers. The sports crowd noise follows a similar trend and is estimated surprisingly diffuse for 22.0 considering that the pink noise and rain are truly uncorrelated between loudspeakers. Considering only these three materials, the fit to the data is good with an adjusted r-squared of 0.8967.

The pop music recording which has multiple point sources is predicted not diffuse for all loudspeaker layouts.

The forest recording is also predicted low except for the 22.0 layout. Interestingly, the majority of the mixes were very consistent with all instruments equally loud between the different loudspeaker layouts, however the 22.0 mix of the forest has a notably quieter flute. The flute is a particularly tonal and appears to have a large effect on the diffuseness metric although perceptually the 22.0 mix is only slightly more enveloping.

The jazz is also predicted low with 5.0 strangely lower than both stereo and 22.0. This appears to correspond to a strange effect in the recording where the 5.0 mix appears to have the piano off to the side. It is not immediately clear why this is in the binaural recordings and why this did not affect the rating of diffuseness if it was clearly off-centre. Equally why this off-centre piano would affect the of diffuseness any differently to having the piano on-centre.

It is clear that the metric is unable to predict the envelopment in cases where the sound field is comprised of multiple spatially separated signals. These are sound fields with multiple localisable components that, on their own may not be considered enveloping, but together form an enveloping scene.

Whilst this does appear to be a poor fit to the data, the stimuli tested here are designed to be highly enveloping. The mixes were created to the remit of being maximally enveloping. Including less enveloping mixes or mono for example would interrogate more the range of possible envelopment levels as well as the range of the metric. This would likely show a higher correlation with the metric. Equally, if the question had been rephrased and localisable scenes rated as not diffuse, there would again be likely better correlation.

9.2.7 Summary

In this experiment the limitations of the metric were found. Whilst the metric appeared somewhat successful when the material was similar to the material of the previous experiments, complex sound scenes composed of non-diffuse components were underestimated. The metric likely has value for monitoring the diffuseness of the sound field components but may not be suited to the overall envelopment of a recording. A sound field that has a high coherence-based diffuseness estimation is likely highly enveloping, but a sound field that has a low coherence-based diffuseness estimation is not necessarily not enveloping although it does indicate it is at least composed of non-diffuse elements.

The difference between the static stimuli and the dynamic stimuli shows an interesting further limitation of the metric at the current stage. Although there is less correlation between loudspeakers in the stereo static stimuli, they are rated as less enveloping. This material dependence could be related to a higher cognitive level of perception and may therefore not be possible to predict from simple measurements of the sound field. Any time varying dependences are not easily captured from the frequency domain based coherence metric.

Whilst these specific cases highlight some of the limitations of the metric, the metric still has many advantages over other methods of predicting the perceived diffuseness. The spatial coherence was found throughout experiments 2 to 5 to have a high correlation with the subjective data and the metric that was developed based on the coherence, maintains this good agreement with the subjective diffuseness.

Chapter 10

Diffuse Sound in Object-Based Audio

10.1 Overview

In this chapter the results of the experiments and the coherence-based metric are described in the context of spatial audio practice and with reference to object-based audio.

The spatial coherence is a parameter of the sound field evaluated at the listening position. This has been found to, in general, correlate well with perception. Because the spatial coherence has an analytical solution, it is possible to determine the factors that affect the shape of the spatial coherence curve and equally, this allow optimisation of the spatial audio reproduction chain. This minimisation process allows some recommendations to be made regarding loudspeaker layouts, loudspeaker signal processing and microphone techniques for minimising the spatial coherence in an attempt to maximise the perceived diffuseness.

Whilst it is outside the scope of this thesis to evaluate these techniques subjectively, initial informal listening suggests that there is some advantage to these techniques to increase the sense of perceived diffuseness.

Firstly, the analytical function for the coherence is reintroduced and decomposed to demonstrate how it may be possible to minimise the spatial coherence. This thinking can then be applied to different parts of the spatial audio signal chain to show how the spatial coherence might be best minimised.

10.2 Minimising the Coherence

Firstly, back in in equation 4.4 in section 4.7, an analytical function of the coherence $\gamma_{xy}(f)$ was described for uncorrelated loudspeaker signals.

$$\gamma_{xy}(f) = \left(\frac{1}{\sum A}\sum_{n=1}^{N} A_n \cos(kR\sin(\theta_n))\right)^2 + \left(\frac{1}{\sum A}\sum_{n=1}^{N} A_n \sin(kR\sin(\theta_n))\right)^2 \tag{10.1}$$

Where k is the wavenumber, R the distance between measurement points,N the number of loudspeakers and A_n and θ_n are the linear gain and angle to the median plane respectively for the n-th loudspeaker.

For the sake simplification, if the signals are assumed to be of the same amplitude (i.e. $A_n = 1/N$ to maintain overall level), then the equation becomes,

$$\gamma_{xy}(f) = \left(\frac{1}{N}\sum_{n=1}^{N} \cos(kR\sin(\theta_n))\right)^2 + \left(\frac{1}{N}\sum_{n=1}^{N} \sin(kR\sin(\theta_n))\right)^2 \tag{10.2}$$

This coherence function is a sum of cosines squared plus a sum of sines squared. To minimise the coherence, firstly the sine component is removed by using any pairs of loudspeakers positioned symmetrically about the median plane. For example using a pair of loudspeakers at $\pm\theta^\circ$ allows,

$$\left(\frac{1}{N}\sum_{n=1}^{N} \sin(kR\sin(\theta_n))\right)^2 \tag{10.3}$$

to become,

$$\left(\frac{1}{2}(\sin(kR\sin(\theta)) + \sin(kR\sin(-\theta)))\right)^2 = 0 \tag{10.4}$$

However the cosines part of the coherence function cannot fully cancel out. In the theoretical diffuse sound field with an infinite number of uncorrelated plane waves, these cosines will cancel at higher frequencies leading to the coherence of $(\sin(kR)/kR)^2$. However, for signals reproduced by loudspeakers, there are a finite number of “plane" waves and therefore the coherence function will always be periodic. The cosine part of the coherence function, given by,

$$\left(\frac{1}{N}\sum_{n=1}^{N} \cos(kR\sin(\theta_n))\right)^2 \tag{10.5}$$

can be minimised for a given frequency, but not for all frequencies.

10.3 Loudspeaker Positions

Minimising the coherence function across all frequencies requires many loudspeakers. The periodic coherence function will, by definition be high at low frequencies and also equal one periodically.

However, for a given narrow bandwidth frequency, the coherence can be minimised or even zero using only a pair of loudspeakers. By using a pair of loudspeakers symetrically about the median plane, the coherence can be minimised by minimising,

$$\left(\frac{1}{N}\sum_{n=1}^{N}\cos(kR\sin(\theta_n))\right)^2 \tag{10.6}$$

This can be achieved if,

$$kR\sin(\theta_n) = n\pi \tag{10.7}$$

where n is an odd integer. Therefore, at high frequencies, a pair of loudspeakers positioned at $\pm\theta$ where θ is given by,

$$\begin{aligned}\theta &= \sin^{-1}\left(\frac{n\pi}{2kR}\right) && (10.8)\\ &= \sin^{-1}\left(\frac{nc}{4fR}\right) && (10.9)\end{aligned}$$

will theoretically produce microphone signals with a coherence of zero.

However, there is only a valid value for θ when $4fR \geq c$ where c is the speed of sound. This means that below a frequency of $c/4R$, the coherence must be greater than zero. For a head diameter of $R = 17$ cm and a speed of sound of 340 ms^{-1}, the coherence below 500 Hz will always be greater than zero–when using uncorrelated plane waves.

Figure 10.1 shows the angle from the median plane for a pair of uncorrelated narrow band signals to give a coherence of 0.

The span angle of the loudspeakers to minimise coherence is interestingly equivalent to the distribution of an Optimal Source Distribution (OSD) transaural array from Takeuchi and Nelson (2002). Although in this case the distribution leads to low inter-microphone coherence rather than the purpose of OSD, which is one of efficiency when using transaural filters.

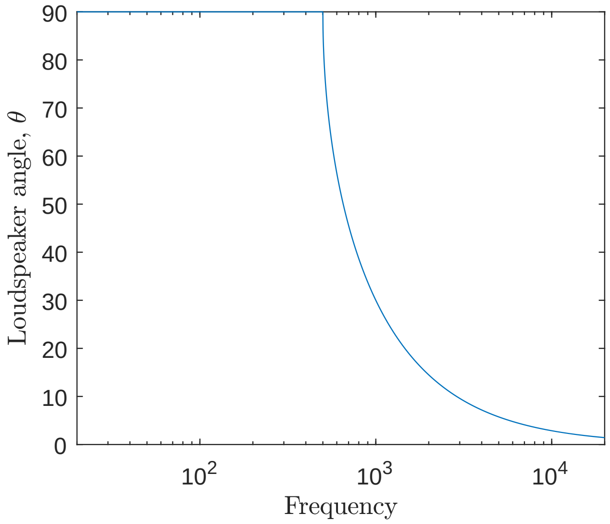


Figure 10.1: Optimal source distribution. Equivalent to the angle from the median plane to create a minimal coherence between spatially separated microphone signals separated by 17 cm.

Whilst these findings are of interest and are useful, they simultaneously highlight some limitations or peculiarities of the coherence as a metric of diffuseness. The coherence function only depends on the angle to the median plane. Therefore any coherence function relating to a given 3D loudspeaker layout could also be generated by an equivalent layout constrained to an arc of loudspeakers from 0 to 180 from the inter-microphone axis. This seems to suggest that 3D loudspeaker layouts have no benefit over loudspeaker layouts that cover a single arc. If head rotation is considered, then the advantage of a 2D layout of loudspeakers over a single arc is apparent. However, because head rotation is usually only considered in a single axis, there is still no obvious advantage for 3D layouts that were found to be significantly more diffuse than the most diffuse 2D layout in experiment 2 but appear to have no benefit to reducing coherence. Because listeners can move their head, and 3D layouts are more perceptually diffuse than a pair of loudspeakers, the coherence at one head orientation is not sufficient to explain the whole perception of diffuseness.

Another point is that, the minimum coherence at low frequencies is given by a pair of loudspeakers at ±90°. This coherence value is less coherent than the coherence of a theoretical diffuse sound field. This implies that either the diffuseness estimation should be taken with reference to the coherence of the theoretical diffuse sound field, or the using a pair of loudspeakers at ±90° is more perceptually diffuse than even a theoretical diffuse sound field. Whilst this would only be the case for a single head orientation, if head rotation is included but limited to the horizontal plane, then a 2D diffuse sound field would render a lower coherence than a 3D diffuse sound field for all head angles. A 2D diffuse field leads to a coherence between spatially separated microphones of $J_0(kR)$ where J_0 is the zero-th order Bessel function; which is lower than the coherence of a 3D diffuse field given by $(\sin(kR)/kR)^2$. This seems to be in contrast to the results that showed increased perceived diffuseness for 3D loudspeaker layouts.

10.4 Loudspeaker Panning

The previous section derived the optimal layout of loudspeakers to minimise the coherence at a given frequency. In reality, a more likely scenario is that the loudspeaker layout is fixed and there are a number of diffuse sound field components that should be mapped to loudspeakers.

For a given loudspeaker layout and a given number of diffuse sound field components the signals can be panned to minimise the coherence and thereby to maximise the perceived diffuseness. Additionally, the input signals can be split into narrow frequency bands and these panned differently. This method allows for optimal panning of sources to minimise the coherence whilst using fewer input signals then the number of loudspeakers and no decorrelation (which would add artefacts).

The minimisation of the coherence should be done in combination with some additional metric that seeks to maximise the diversity of loudspeakers used in order to avoid the optimisation only working for a single listener orientation. This could be achieved by using an algorithm that attempts to reduce the coherence but for a wide variety of head rotation angles.

10.5 Microphone Techniques

To minimise the coherence the correlation between the loudspeaker signals needs to be minimal. This can usually be achieved using spaced microphone techniques where the wide spacing allows low correlation between parts of the sound field and therefore low correlation between loudspeaker signals. Low correlation is also possible using ambisonic recording techniques. If the sound field being recorded is diffuse, then the different components of the the ambisonic recording should also be uncorrelated. Typically, virtual microphones are used to render ambisonic material to loudspeakers. However this will tend to add correlation between loudspeakers. Using the raw ambisonics components should retain the low correlation, and if the source was diffuse to begin with, should not degrade the spatial components of the recording.

10.6 Summary

In summary, low spatial coherence has been found to correlate well the perception of diffuseness. The spatial coherence has an analytical solution that can be used to minimise the spatial coherence depending on the constraints of the system. Low correlation between loudspeakers is the most important factor and an assumption of the analytical model. This leads to recommendations for microphone techniques and capturing methods. Then based on the model, either the positions of the loudspeakers or the panning of frequency components between loudspeakers can be exploited to minimise the spatial coherence with the goal of maximising the perceived diffuseness.

Chapter 11

Conclusions

In this chapter, the results and findings of this research are summarised. The topic of diffuseness in relation to object-based audio is explained, the findings are highlighted and the remaining issues that could be used as the basis for future research are covered in the final section

11.1 Thoughts on Diffuseness

The goal of this research was to investigate how diffuse sound fields are perceived and how that relates to object-based spatial audio. A diffuse sound field proposes a particular challenge in spatial audio and one with no obvious solution. This section covers some of the main outcomes of this research. Subjective experiments generated new data–subjective and objective–which has led to a deeper understanding of the perception of diffuseness and informed the design of a new metric, based on the spatial coherence for predicting the perceived diffuseness.

The first issue comes with defining a diffuse sound field. A genuinely diffuse sound field does not exist in the real world. It would have to be generated by an infinite number of uncorrelated plane waves from all directions simultaneously. It is one that is completely homogeneous and isotropic. Whilst impossible in reality, it has mathematical properties that make it very useful for making measurements; is a good approximation of several realistic sound fields and also represents the upper limit of the type fo sound field that is so difficult to reproduce accurately in spatial audio.

The second issue comes with defining a partially diffuse sound field. A sound field is also only diffuse or not diffuse. To be diffuse it must be both homogeneous and isotropic and any deviation from either of these conditions means the sound field is not diffuse.

Sound fields such as late reverberation, crowd noise or rain are often described as diffuse and might have properties similar to the theoretical diffuse sound field, but they are not truly diffuse. In the case of reverberation, the reflections are all correlated with each other and so the sound field is not truly diffuse. The noise of a large crowd or rain is uncorrelated, but there are not an infinite number of uncorrelated signals and not from all directions simultaneously. For these sound fields, there is a degree of diffuseness, but diffuseness is extremely ambiguous. One sound field may be homogeneous but not isotropic and another might be isotropic but not homogeneous. It is impossible to determine which of these two sound fields is more or less diffuse because the factor of diffuseness is multidimensional. In this research though, the factor of sound field diffuseness is secondary to the perception of the sound field. Listeners were asked how similar they thought the sound to a diffuse sound field in terms of its spatial impression. The average response from a set of listeners is once again one dimensional as listeners combine the differences they hear into a single rating of perceptual diffuseness. Therefore, in this research, the multidimensionality of diffuseness was investigated over a series of listening tests with each experiment eliciting both the subjective perceived diffuseness but also objective measurements of the sound field.

The subjective data was analysed and revealed several factors relating to the perception of diffuseness. The first two experiments investigated how the isotropy of the sound field affected the perceived diffuseness. As expected, a greater number of loudspeakers produces a more subjectively diffuse sound field. Three-dimensional loudspeaker layouts have the potential to be more perceptually diffuse than two-dimensional layouts. Whilst it is important to have loudspeakers from all directions, standardised loudspeaker layouts that tend to have more loudspeakers in the frontal direction were not notably less perceptually diffuse thank more evenly distributed layouts.

Although the isotropy of the sound field was found to be important to the perception of diffuseness, it was not the only factor. In the third experiment, by varying the Inter-Channel Correlation Coefficient (ICCC), the similarity between loudspeaker signals could be varied and the homogeneity of the sound field varied whilst maintaining the isotropy. This was found to be highly important to the perception of diffuseness. Stimuli with high ICCC were rated as not diffuse independently of the number of loudspeakers. A loudspeaker system with 37 loudspeakers was rated as less diffuse than uncorrelated stereo highlighting why more loudspeakers does not necessarily guarantee high diffuseness

Whilst the ICCC experiment highlighted the importance of high isotropy towards a high perception of diffuseness, the human hearing system is limited to its two sensors, the ears. The limited separation of the ears means that the isotropy and homogeneity of the sound field have a different effect on the signals at the ears of a listener at different frequencies. The wavelength of the sound can be much longer than the ear separation or much shorter. This was predicted to have a large effect on the perception of diffuseness. However, it was found in experiment 4 that at all frequencies the ICCC had the same effect on the perceived diffuseness.

During all of these listening tests objective measurements of the sound field were taken. Metrics that only consider either the homogeneity or the isotropy were found to be insufficient to predict all the subjective data. This led to the realisation that a metric of diffuseness must include both the homogeneity and also the isotropy of the sound field at the listening position. The coherence between spatially separated points was therefore employed as a metric of diffuseness as it factors both the homogeneity of the sound field and the isotropy. By using a microphone spacing akin to the separation of the ears, the coherence is based on pseudo-binaural signals that reflect the way in which the individual loudspeaker signals combine at the ears of a listener. Interestingly, the coherence is a function that determines how similar the signals at the ears of a listener are. As a sound wave approaches a listener it is heard by both ears. In a stereo system, the sound from the left speaker goes to both ears as does the signal from the right loudspeaker. This loudspeaker system does not deliver the two signals to the ears separately, instead it is the summation of the left and right loudspeaker signals at both ears of the listener that contributed to the perceived sound field and the given spatial impression. In a diffuse sound field, at mid and high frequencies, there are enough incoming plane waves from all directions that the signals at the ears are highly uncorrelated, even though individually all the incoming plane waves are heard by both left and right ears, the culmination of all the infinite plane waves from all directions, the ear signals are uncorrelated. Whilst the coherence metric could have been developed to use binaural recordings, it was found that the head-shadowing at higher frequencies led to a low coherence and therefore high diffuseness estimate at high frequencies despite the results of experiment 4 indicating that the perception of diffuseness at high frequencies was similar to that at low and mid frequencies.

Using the coherence as a metric of diffuseness was found to fit the data well, but also there were limitations of the coherence. Therefore a diffuseness metric based on the coherence was developed to avoid these likely pitfalls and the algorithm implemented in python using the VISR framework allowing the model can be used in real-time or offline.

The algorithm takes spaced microphone recordings of a sound field and calculates the coherence. The minimum coherence is limited by the separation of the microphones but the separation should be chosen to represent the ITD of a listener (17 cm). Therefore at low frequencies the coherence function is normalised using the range of possible coherence values based on the summation of plane-waves. This leads to a value between zero and one at all frequencies with zero representing the lowest possible coherence and one representing a single plane-wave. The coherence is calculated in the FFT domain and is therefore linear with frequency meaning averaging across frequency gives unfair weighting to the higher frequency coefficients. Therefore the coherence is divided into frequency sub-bands that lie on the equivalent rectangular bandwidth scale. The coherence is independent of the loudness of the frequency band or the time window and therefore the coherence is weighted based on the frequency content of the input signals but also the loudness of the input signal. This averaged coherence gives an estimated diffuseness between zero and one for each window. The algorithm was designed to make the coherence metric robust to a range of input signals.

One of the components of the coherence-based metric that required particular attention and could not be approximated was the choice of analysis window length. For particular stimuli that have a strong dependence on time delays or long filter lengths–for example reverberation–the coherence depends on the length of the analysis window. Therefore a subjective experiment was conducted using analysis window sensitive material to determine the optimal window length for the coherence-based diffuseness metric. This was found to be 512 samples at a sampling frequency of 48 kHz.

The metric was then tested using some auditory scenes that are more realistic to the types of content in a broadcast environment. In these experiment the coherence-based metric was found to predict well the diffuseness for the type of stimulus that it was trained with, however material that had multiple discrete non-diffuse components was perceived as enveloping by the listener although was predicted as not diffuse by the metric. This appears to a be the main limitation of the metric. Whilst the metric is likely of use for material that is surrounding and perceived as a single entity, it is a poor predictor of envelopment for more complex scenes. However, it is likely that the diffuseness metric would serve as a component of a more complex envelopment meter.

The coherence-based metric could well serve as a standalone tool for use in production environment for monitoring diffuseness. However the nature of the coherence as the basis for the metric also allows some recommendations to be made for object-based audio. Firstly, methods of recording diffuse sound fields, reproducing diffuse sound fields using loudspeakers or generating perceptually diffuse sound fields based on decorrelation or panning have been discussed. It also highlighted some of the inherent difficulties when trying to take a non-diffuse source and make it diffuse. For example, tonal sources cannot be diffuse according to the metric and therefore decorrelation algorithms should be developed that minimise the audible artefacts when the diffuseness of the source is limited by its type.

For the factors tested over the course of this research, the coherence based-metric has provided a good degree of correlation with the subjective data. However, there are still a range of factors that were not investigated and the coherence has features that could not have been investigated in the available time. The following section provides some basis for the work that has not been investigated yet and could provide the basis for future research.

11.2 Possible Future Subjective Experiments

The ability of a loudspeaker system to reproduce diffuse sound fields is important when trying to create surrounding and enveloping sound fields. The metric that was developed based on the research described in this thesis provides a starting point for the optimisation of object-based audio. Ideally this metric would provide the perceived diffuseness for all stimuli with 100% accuracy. There would be no need to run subjective diffuseness experiments. From this research it was possible to make some initial recommendations for how to reproduce diffuse sound fields. However, the metric is not perfect in all conditions and there remain unanswered questions.

Firstly there are known limitations of the metric. Low frequencies are always coherent and yet have variable perceived diffuseness. There is scope to investigate this phenomenon in more detail. It is likely that at low frequencies there may be a better metric to predict perceived diffuseness, and adding this to a hybrid model would improve the generality of the metric. Similarly there are the stimuli that were poorly predicted in chapter 9. Complex scenes composed of multiple localisable sources or high frequency components that do not come from all directions were poorly estimated and there is further scope to refine the metric. Additionally, components of the metric such as the frequency based weighting and the silence detection could be more refined and based on loudness weighting rather than arbitrary cut-off thresholds.

Secondly, there are several factors that have not been tested at all as yet. For example, the early experiments used wide bandwidth noise to investigate the perceived diffuseness in a controlled manner. The later experiments used realistic material, however one variable not investigated in the earlier experiments was the effect of bandwidth. This was fond to be relevant by Santala and Pulkki (2011) but the coherence is not bandwidth dependent. Therefore an experiment that combined the factors of centre frequency and bandwidth would provide a fuller picture for a wider range of stimuli.

Similarly, in experiment 3, variable decorrelation was achieved by combining a global correlated signal with uncorrelated signals to generate variable correlation. However, referring back to section 5.2.2, the method of decorrelation used was just one of four ways proposed in that section. There are several other methods for getting variable correlation that could be closer to real world signals. Especially where adjacent loudspeakers are highly correlated but spatially separated loudspeakers have lower correlation. This leads to a complex inter-loudspeaker correlation coefficient matrix that has more than the single variable of inter-channel correlation coefficient. It is not clear how a different method of achieving variable inter-channel correlation might affect both the perception of diffuseness and the coherence-based metric. An experiment that compared different methods of decorrelation could provide some additional insight.

Additionally, the findings of this research could be assessed critically. For example, the coherence was found to be the best metric of perceived diffuseness. however, the spatial coherence assumes that sources in the same cone-of-confusion are equivalent. This implies that any loudspeaker layout that is rotationally symmetrical about the interaural axis is equivalent in terms of perceived diffuseness. An interesting further experiment would be to create loudspeaker layouts that are rotationally symmetric and then restrict head rotation for some stimuli and allow head rotation for others. This would allow this counter-intuitive feature of the metric to be tested and give further insight into the role of head rotation in the perception of diffuseness.

Finally, the coherence-based metric can be used to design optimal diffuse sound field capture, processing and reproduction. Subjective testing of any capture or reproduction based on this metric would allow for highly critical assessment of the metric whilst potentially providing usable tools based on subjective research for use by spatial audio engineers.

The metric provides a better estimate than existing tools and it's simplicity allows for development of further spatial audio tools. However, the metric is not perfect. There are areas in which it can be improved, there are areas in which it has not yet been tested, and there are areas in the spatial audio signal chain where this metric can be implemented to maximise the perception of diffuseness.

Bibliography

Ahonen, J., Kallinger, M., Küch, F., Pulkki, V., and Schultz-Amling, R. (2008). Directional analysis of sound field with linear microphone array and applications in sound reproduction. In *Audio Engineering Society Convention 124*. Audio Engineering Society.

Ando, Y. and Kurihara, Y. (1986). Nonlinear response in evaluating the subjective diffuseness of sound fields. *The Journal of the Acoustical Society of America*, 80(3):833–836.

Barron, M. (2001). Late lateral energy fractions and the envelopment question in concert halls. *Applied Acoustics*, 62(2):185–202.

Barron, M. and Marshall, A. (1981). Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure. *Journal of Sound and Vibration*, 77(2):211–232.

Berg, J. (2009). The contrasting and conflicting definitions of envelopment. In *Audio Engineering Society Convention 126*. Audio Engineering Society.

Berg, J. and Rumsey, F. (1999). Spatial attribute identification and scaling by repertory grid technique and other methods. In *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*. Audio Engineering Society.

Berg, J. and Rumsey, F. (2003). Systematic evaluation of perceived spatial quality. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society.

Bradley, J. S. and Soulodre, G. A. (1995). The influence of late arriving energy on spatial impression. *The Journal of the Acoustical Society of America*, 97(4):2263–2271.

Clark, D. (1981). High Resolution Subjective Testing Using a Double Blind Comparator. In *Audio Engineering Society Convention 69*.

Cook, R. K., Waterhouse, R. V., Berendt, R. D., Edelman, S., and Thompson, M. C. (1955). Measurement of Correlation Coefficients in Reverberant Sound Fields. *The Journal of the Acoustical Society of America*, 27(6):1072–1077.

Daele, B. V. and Baelen, W. V. (2012). Productions in Auro-3d.

De Man, B. and Reiss, J. D. (2013). A pairwise and multiple stimuli approach to perceptual evaluation of microphone types. In *Audio Engineering Society Convention 134*. Audio Engineering Society.

Franck, A. and Fazi, F. M. (2018). Visr – a versatile open software framework for audio signal processing. *Audio Engineering Society.*

Francombe, J., Brookes, T., and Mason, R. (2018). Determination and validation of mix parameters for modifying envelopment in object-based audio. *Journal of the Audio Engineering Society*, 66(3):127–145.

Gade, S. (1982). Technical Review No. 3: Sound Intensity (Theory). Technical report, Brüel & Kjær.

Gover, B. N., Ryan, J. G., and Stinson, M. R. (2002). Microphone array measurement system for analysis of directional and spatial variations of sound fields. *The Journal of the Acoustical Society of America*, 112(5):1980.

Gover, B. N., Ryan, J. G., and Stinson, M. R. (2004). Measurements of directional properties of reverberant sound fields in rooms using a spherical microphone array. *The Journal of the Acoustical Society of America*, 116(4):2138.

Gribben, C. and Lee, H. (2017). A Comparison between Horizontal and Vertical Interchannel Decorrelation. *Applied Sciences*, 7(11):1202.

Griesinger, D. (1999). Objective Measures of Spaciousness and Envelopment. In *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction.*

Hidaka, T., Beranek, L. L., and Okano, T. (1995). Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls. *J. Acoust. Soc. Am.*, 98 (2)(Pt. 1):988:1007.

Hiyama, K., Komiyama, S., and Hamasaki, K. (2002). The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field. In *Audio Engineering Society Convention 113*. Audio Engineering Society.

Hosseini, M. and Georganas, N. D. (2002). MPEG-4 BIFS streaming of large virtual environments and their animation on the web. In *Proceedings of the seventh international conference on 3D Web technology*, pages 19–25. ACM.

ISO (1999). ISO 3741:1999 Acoustics. Determination of sound power levels of noise sources using sound pressure. Precision methods for reverberation rooms. *International Organization for Standardization.*

ISO (2009). BS EN ISO 3382-1:2009 Acoustics, Measurement of room acoustic parameters. *International Organization for Standardization.*

ITU-R, B. S. (1997). Rec. ITU-R BS.1116-1:1997 Method for the subjective assessment of small impairments in audio systems including multi-channel sound systems. *Int. Telecommun. Union, Geneva, Switzerland.*

ITU-R, B. S. (2007). Rec. ITU-R BS.775-2: 2007 Multichannel stereophonic sound system with and without accompanying picture. *Int. Telecommun. Union, Geneva, Switzerland.*

ITU-R, B. S. (2011). Rep. ITU-R BS.2159-1: 2011 Multichannel sound technology in home and broadcasting applications. *Int. Telecommun. Union, Geneva, Switzerland.*

ITU-R, B. S. (2014a). Rec. ITU-R BS.1534-2:2014 Method for the subjective assessment of intermediate quality level of audio systems. *Int. Telecommun. Union, Geneva, Switzerland.*

ITU-R, B. S. (2014b). Rec. ITU-R BS.2051-0: 2014 Advanced sound system for programme production. *Int. Telecommun. Union, Geneva, Switzerland.*

Jacobsen, F. and Roisin, T. (2000). The coherence of reverberant sound fields. *The Journal of the Acoustical Society of America*, 108:204.

Jang, I., Kang, K., Lee, T., and Park, G. Y. (2005). An Object-based 3d Audio Broadcasting System for Interactive Services. In *Audio Engineering Society Convention 118*. Audio Engineering Society.

Kendall, G. S. (1995). The Decorrelation of Audio Signals and Its Impact on Spatial Imagery. *Computer Music Journal*, 19(4):71.

Koenen, R. (2002). R. Koenen, Overview of the MPEG-4 Standard, ISO/IECJTC1/SC29/WG11, Document N4668, (2002).

Laitinen, M.-V., Küch, F., and Pulkki, V. (2011). Using spaced microphones with directional audio coding. In *Audio Engineering Society Convention 130*. Audio Engineering Society.

Loftis, B. (2014). Object Panning for Film: Challenges and Solutions. In *Audio Engineering Society Convention 137*. Audio Engineering Society.

Loutridis, S. J. (2009). Quantifying sound-field diffuseness in small rooms using multifractals. *The Journal of the Acoustical Society of America*, 125(3):1498.

Merimaa, J. and Pulkki, V. (2005). Spatial impulse response rendering I: Analysis and synthesis. *Journal of the Audio Engineering Society*, 53(12):1115–1127.

Meyer, E. (1954). Definition and Diffusion in Rooms. *The Journal of the Acoustical Society of America*, 26(5):630–636.

Nélisse, H. and Nicolas, J. (1997). Characterization of a diffuse field in a reverberant room. *The Journal of the Acoustical Society of America*, 101(6):3517:3524.

Nolan, M., Fernandez-Grande, E., Brunskog, J., Richard, A., and Jeong, C.-H. (2016). A wavenumber approach to characterizing the diffuse field conditions in reverberation rooms. In *Proceedings of the 22nd International Congress on Acoustics*, Buenos Aires. International Congress on Acoustics.

Paine, G., Sazdov, R., and Stevens, K. (2007). Perceptual Investigation into Envelopment, Spatial Clarity, and Engulfment in Reproduced Multi-Channel Audio. In *Audio Engineering Society Conference: 31st International Conference: New Directions in High Resolution Audio*. Audio Engineering Society.

Plogsties, J., Baum, O., and Grill, B. (2003). Conveying spatial sound using MPEG-4. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society.

Power, P., Davies, B., and Hirst, J. (2014). Investigation into the Impact of 3d Surround Systems on Envelopment. In *Audio Engineering Society Convention 137*. Audio Engineering Society.

Pulkki, V. and Merimaa, J. (2006). Spatial impulse response rendering II: Reproduction of diffuse sound and listening tests. *Journal of the Audio Engineering Society*, 54(1/2):3–20.

Romblom, D., Guastavino, C., and Depalle, P. (2016). Perceptual thresholds for non-ideal diffuse field reverberation. *The Journal of the Acoustical Society of America*, 140(5):3908–3916.

Santala, O. and Pulkki, V. (2011). Directional perception of distributed sound sources. *The Journal of the Acoustical Society of America*, 129(3):1522–1530.

Schmidt, J. and Schroeder, E. F. (2004). New and advanced features for audio presentation in the MPEG-4 standard. In *Audio Engineering Society Convention 116*. Audio Engineering Society.

Spring, N. and Randall, K. E. (1969). The measurement of sound diffusion index in small rooms. Technical Report 16.

Takeuchi, T. and Nelson, P. A. (2002). Optimal source distribution for binaural synthesis over loudspeakers. *The Journal of the Acoustical Society of America*, 112(6):2786–2797.

Veit, I. and Sander, H. (1985). The Production of a Spatially Limited 'Diffuse' Sound Field in an Anechoic Room. In *Audio Engineering Society Convention 77*. Audio Engineering Society.

Zacharov, N., Pedersen, T., and Pike, C. (2016). A common lexicon for spatial sound quality assessment-latest developments. In *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*, pages 1–6. IEEE.

Zielinski, S., Hardisty, P., Hummersone, C., and Rumsey, F. (2007). Potential biases in MUSHRA listening tests. In *Audio Engineering Society Convention 123*. Audio Engineering Society.

Glossary

IACC - InterAural Cross-correlation Coefficient

IACF - InterAural Cross-correlation Coefficient Function

ICCC - Inter-Channel Correlation Coefficient

HF - High Frequency

LF - Low Frequency

ASW - Apparent/Auditory Source Width

LEV - Listener EnVelopment

HRTF - Head-Related-Transfer-Function

ISLD - Inter-Subset Level Difference

ISCLD - Inter-Subset Channel Level Difference

LEDT - Lateral Early Decay Time

IAD - InterAural Difference

FFT - Fast Fourier Transform

STFT - Short-Time Fourier Transform

d.c - Direct Current

BIFS - Binary Information For Scenes/ BInary Format for Scenes

AABIFS - Advanced Audio BInary Format for Scenes.

VRML - Virtual Reality Modelling Language

VBAP - Vector Based Amplitude Panning

SPL	- Sound Pressure Level
ANOVA	- ANalysis Of VAriance
p	- Pressure
$\overrightarrow{u}$	- Particle Velocity
$\overrightarrow{I}$	- Intensity Vector
k	- Wave Number
E	- Energy
A	- Area
θ	- Azimuth
ϕ	- Elevation
$\langle\rangle$	- Time averaging
R	- Microphone Spacing
$\langle\rangle$	- Time averaging
c	- Speed of Sound
ρ_{xy}	- Cross-correlation Coefficient
R_{xy}	- Cross-correlation
S_{xy}	- 2-sided Cross-spectrum
S_{xx}	- 2-sided Auto-spectrum
G_{xy}	- 1-sided Cross-spectrum
G_{xx}	- 1-sided Auto-spectrum
p	- p-value

Appendices

Appendix A

Code

This code, written in python, estimates the perceived diffuseness of a spaced microphone input signal. The microphones should be placed at a separation of 17 cm at the listening position. The code is implemented as a VISR atomic component and outputs both the diffuseness estimated for each incoming buffer, but also the level which is used as a threshold between the main audio material and the background noise. This therefore requires the appropriate VISR python libraries that will, in the future, be made available to the public as part of the S3A project.

```
# -*- coding: utf-8 -*-
"""
git repository: https://git.soton.ac.uk/mpc1r13/Diffusness_Meter.git
@author: Michael Cousins
"""


# Template for an atomic component that takes in an audio signal and outputs a stream of diffuseness values.

# %% Module imports

import visr   # Core VISR module, defines components and ports
import pml    # Parameter message library , defines standard parameter types and communication protocols.

# Python standard library calls.
```

```
import numpy as np

# Definition of a component.
# The class is derived from the VISR base class AtomicComponent,
    that means that
# its function is implemented in code (basically in the process()
     method)
class DiffusenessMeter( visr. AtomicComponent ):
# Defines the constructor that creates an DiffusenessMeter.
def __init__( self, context, name, parent,
numberOfChannels,
audioOut = True ):

# Call the base class constructor
super(DiffusenessMeter, self).__init__(context,name,parent)
# Define an audio input port with name "audioIn" and width (
    number of signal waveforms) numberOfChannels
self.audioInput = visr.AudioInputFloat( "audioIn", self,
    numberOfChannels )

# If the option is set, add an audio output to put out the input
    signals
# Some audio interfaces don't like configs with no outputs.
if audioOut:
self.audioOutput = visr.AudioOutputFloat( "audioOut", self,
    numberOfChannels )
else:
self.audioOutput = None

# Define a parameter output port with type "Float" and
    communication protocol "MessageQueue"
# MessageQueue means that all computed data are hold in a first-
    in-first-out queue,
# which decouples the parameter update rate from the buffer size.
```

```
self.diffusenessOut = visr.ParameterOutput( "diffusenessOut",
   self, pml.Float.staticType, pml.MessageQueueProtocol.
   staticType, pml.EmptyParameterConfig() )
self.levelOut = visr.ParameterOutput( "levelOut", self, pml.Float
   .staticType, pml.MessageQueueProtocol.staticType, pml.
   EmptyParameterConfig() )

# %% Setup data used in the process() function.

# Round the measurement period to the next multiple of the buffer
    period
self.numWin = 30    # multiplr FFTs are required to calculate the
    coherence
self.NFFT = 512     #FFT length
self.winIndex = 0   #counter for the windows
indexes = ((10**(np.arange(0.5,43.5,1)/21.4)-1)/0.00437)
   *512/48000
self.indexes = indexes.astype(int)  #These are the indexes of FFT
    bins that determine the cutooff between bands placed on the
   ERB scale. Chossing a coarser grid in the arrange function
   woul mean evenly spaced bands but wider than a critical
   bandwidth.
self.circBuff = np.zeros( (self.numWin,2,257) , dtype = np.
   complex128 ) #buffer that holds all the FFTs adds new ones
   and deletes old ones.
self.cohBands = np.zeros( self.indexes.size , dtype = np.float64
   ) #array to hold the cohernece values
self.levelsBands = np.zeros( self.indexes.size-1 , dtype = np.
   float64 ) #array to hold diffuseness
self.target = np.zeros(257)

self.target[0:6]= np.square(np.cos(0.17*2*np.pi*48000*np.arange
   (6)/(340*512)))#array with weighting for LF. Based on the
   minimum coherence possible using loudspeakers at +-90. Signals
```

```
    with lower coherence values (e.g binaural/non planewave etc)
    are limited in maximum diffuseness.)


# The process() method implements the runtime operation of the
    component.
# It is called regularly (every context.period samples) by the
    runtime system.
def process( self ):
# Retrieve the new input samples as a Numpy ndarray (dimension
    numChannels )
x = self.audioInput.data()

#calculate FFT
xf = np.fft.fft(x, self.NFFT, 1, None)
#make 1 sided
xf1 = xf[:,0:257:] / self.NFFT
xf1[:,1:] = xf1[:,1:] * 2
#place in buffer
self.circBuff[self.winIndex] = xf1;

#calculate auto and cross spectra
Gxx = (2 / self.NFFT) * np.mean(np.square(abs(self.circBuff
    [:,0,:])), 0)
Gyy = (2 / self.NFFT) * np.mean( np.square(abs(self.circBuff
    [:,1,:])), 0)
Gxy= (2 / self.NFFT) * np.mean( np.multiply(np.conj(self.circBuff
    [:,0,:]),self.circBuff[:,1,:]), 0)

#calculate coherence
coh = np.divide(np.square(abs(Gxy)),np.multiply(np.clip(Gxx, 1.0e
    -12, None ),np.clip(Gyy, 1.0e-12, None )))
#weight the coherence by the minimum possible coherence (Above 0
    Hz where the minimum is 1 anyway to avoid divide by zero)
```

```
coh = np.divide(np.clip((coh[1:]-self.target[1:]), 0, 1,),(1-self
    .target[1:]))
#perceptually linearise the coherence before averaging.
coh = np.power(1 - coh,4)

#put into frequency bands on the ERB scale..

for cf in range(0,self.indexes.size-1):
#

self.cohBands[cf]=np.mean(coh[self.indexes[cf]:self.indexes[cf
    +1]+1])
self.levelsBands[cf]=10 * np.log10(np.clip(np.sum(abs(xf1[:,self.
    indexes[cf]:self.indexes[cf+1]+1])), 1.0e-12, None ))



# Only select bands that are within 15dB of the loudest band.
Diff = np.mean(self.cohBands[self.levelsBands >= max(self.
    levelsBands) - 15])



# The diffuseness is only relevant if the audible sound is part
    of the main programme material based on an arbitrary cut off
    of -70dB. i.e. digital silence should not be displayed as
    having a coherence of 1 and therefore diffuseness of 0.
Level = 10* np.log10(np.clip(np.mean(np.sum(abs(self.circBuff
    [:,0,:]), 1)), 1.0e-12, None))
Include = Level >= -70;

#update window number for circular buffer
self.winIndex = self.winIndex + 1

if self.winIndex >= self.numWin:
self.winIndex = 0
```

```
# Create a parameter message holding a single float.
outParam = pml.Float( Diff )
outParam2 = pml.Float( Include )

# Send the computed diffuseness value to the output.

self.diffusenessOut.protocolOutput().enqueue( outParam )
self.levelOut.protocolOutput().enqueue( outParam2 )


# Send the data to the audio output if this option has been
    chosen.
if self.audioOutput is not None:
self.audioOutput.set( x )
```