2018

Reproduction of Personal Sound in Shared Environments

Jacob Donley
University of Wollongong

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Reproduction of Personal Sound in Shared Environments

Jacob Donley

Supervisors:
Professor Christian Ritz & Professor W. Bastiaan Kleijn

This thesis is presented as required for the conferral of the degree:

Doctor of Philosophy

The University of Wollongong
School of Electrical, Computer and Telecommunications Engineering

January, 2018

Abstract

The experience and utility of personal sound is a highly sought after characteristic of shared spaces. Personal sound allows individuals, or small groups of individuals, to listen to separate streams of audio content without external interruption from a third-party. The desired effects of personal acoustic environments can also be areas of minimal sound, where quiet spaces facilitate an effortless mode of communication. These characteristics have become exceedingly difficult to produce in busy environments such as cafes, restaurants, open plan offices and entertainment venues. The concept of, and the ability to provide, spaces of such nature has been of significant interest to researchers in the past two decades.

This thesis answers open questions in the area of personal sound reproduction using loudspeaker arrays, which is the active reproduction of soundfields over extended spatial regions of interest. We first provide a review of the mathematical foundations of acoustics theory, single zone and multiple zone soundfield reproduction, as well as background on the human perception of sound. We then introduce novel approaches for the integration of psychoacoustic models in multizone soundfield reproductions and describe implementations that facilitate the efficient computation of complex soundfield synthesis. The psychoacoustic based zone weighting is shown to considerably improve soundfield accuracy, as measured by the soundfield error, and the proposed computational methods are shown capable of providing several orders of magnitude better performance with insignificant effects on synthesis quality. Consideration is then given to the enhancement of privacy and quality in personal sound zones and in particular on the effects of unwanted sound leaking between zones. Optimisation algorithms, along with *a priori* estimations of cascaded zone leakage filters, are then established so as to provide privacy between the sound zones without diminishing quality. Simulations and real-world experiments are performed, using linear and part-circle loudspeaker arrays, to confirm the practical feasibility of the proposed privacy and quality control techniques. The experiments show that good quality and confidential privacy are achievable simultaneously. The concept of personal sound is then extended to the active suppression of speech across loudspeaker boundaries. Novel suppression techniques are derived for linear and planar loudspeaker boundaries, which are then used to simulate the reduction of speech levels over open spaces and suppression of acoustic reflections from walls. The suppression is shown to be as effective as passive fibre panel absorbers. Finally, we propose a novel ultrasonic parametric and electrodynamic loudspeaker hybrid design for acoustic contrast enhancement in multizone reproduction scenarios and show that significant acoustic contrast can be achieved above the fundamental spatial aliasing frequency.

Acknowledgements

First, I would like to express my utmost gratitude to my two supervisors, Prof. Christian Ritz and Prof. W. Bastiaan Kleijn, who have provided me with truly invaluable knowledge and guidance over the last several years. Their insightful and thought provoking discussions always made me question the next path to take; without, this thesis would never have come to be.

Thank you to everyone in my school, those in my office, those who I have taught with and those who I have taught. You have all made this a pleasurable, bearable and unforgettable journey through the ups and downs of everyday research.

To my closest of friends, thank you for being there, always, and for the fun times had when I was not stuck in the books.

I wish to express my warm thanks to my sister, cousin and family, who have given me the love and support that has allowed me to focus on completing this research project and degree.

To my mother and father, you are both the reason I am where I am and I thank you both whole heartedly for everything you have provided me with over the years.

Thank you all! The few words here are simply not enough.

Statements

Format

Much of this work has either been published or submitted for publications as journal articles and conference proceedings. The following is a list of the manuscripts that formed the basis for the thesis.

Peer-reviewed Publications

- J. Donley and C. Ritz, “An efficient approach to dynamically weighted multizone wideband reproduction of speech soundfields,” in *China Summit Int. Conf. Signal Inform. Process. (ChinaSIP)*, IEEE, Jul. 2015, pp. 60–64.

- J. Donley and C. Ritz, “Multizone reproduction of speech soundfields: A perceptually weighted approach,” in *Asia-Pacific Signal Inform. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, IEEE, 2015, pp. 342–345.

- J. Donley, C. Ritz, and W. B. Kleijn, “Improving speech privacy in personal sound zones,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2016, pp. 311–315.

- J. Donley, C. Ritz, and W. B. Kleijn, “Reproducing personal sound zones using a hybrid synthesis of dynamic and parametric loudspeakers,” in *Asia-Pacific Signal & Inform. Process. Assoc. Annu. Summit and Conf. (APSIPA ASC)*, IEEE, Dec. 2016, pp. 1–5.

- J. Donley, C. Ritz, and W. B. Kleijn, “Active speech control using wave-domain processing with a linear wall of dipole secondary sources,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2017, pp. 1–5.

- J. Donley, C. Ritz, and W. B. Kleijn, “Multizone soundfield reproduction with privacy- and quality-based speech masking filters,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1041–1055, 2018.

- J. Donley, C. Ritz, and W. B. Kleijn, “On the comparison of two room compensation / dereverberation methods employing active acoustic boundary absorption,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, Apr. 2018, pp. 221–225.

The research presented in this thesis has been performed jointly with Professor Christian Ritz and Professor W. Bastiaan Kleijn. Approximately 80% of this work is my own.

Declaration

I, Jacob Donley, declare that this thesis submitted in fulfilment of the requirements for the conferral of the degree Doctor of Philosophy, from the University of Wollongong, is, the majority, my own work. This document has not been submitted for qualifications at any other academic institution.

Jacob Donley
Saturday 10th November, 2018

Professor W. Bastiaan Kleijn
Saturday 10th November, 2018

Professor Christian Ritz
Saturday 10th November, 2018

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Overview: *This chapter provides an introduction to the thesis with background knowledge of sound and spatial audio reproduction. Current difficulties with existing technologies of personal sound reproductions are outlined along with motivations to solve existing problems. The contributions of this thesis are summarised at the end of this chapter.*

1.1 Background

In nature, the fundamental physical process of sound is that of pressure oscillations through a medium. Humans perceive the pressure oscillations through air as audible sounds such as speech or wind noise. A fundamental mechanism of communication for humans is the sound of speech [1]. Speech carries information to be conveyed from one location in space to another across open areas.

The pressure amplitudes as a function of space are referred to as soundfields and can theoretically contain any desired content [2]. Soundfields are produced by physically exciting a medium such that it oscillates a wave over a space. The physical nature of sound leads to sound waves that either subtract from each other, add to each other or result in a field of pressures that is somewhere in-between. A soundfield can be interpreted by measuring the strength of oscillations in pressure at locations within a field, for instance, with human hearing or pressures sensors, such as microphones. The human ears channel sound pressures to the cochlea which the brain then perceives as sound [3]. The superposition of sound waves can lead to mixtures of desired and undesired soundfields resulting in media content that is of perceivably poor quality and difficult to distinguish [4].

Spatial audio is sound that gives listeners a sense of immersion and presence in an acoustic environment. Listeners perceive sounds as if they arrive from various directions or locations in space. The concept of artificially reproducing a high quality soundfield has existed for many decades. Early implementations contained only a small number of loudspeakers as monophonic, stereophonic and quadraphonic systems [5]. In the 1930's stereophony was developed and created spatial impression for listeners using a combination of delay and level differences [6], [7]. The most common approach to stereophony and surround sound is level (amplitude) panning due to the accurate results it delivers for localisation [8]–[10]. Vector-based Amplitude Panning (VBAP) has generalised the concept of pairwise (two loudspeaker) [11], [12] amplitude panning by adding a further dimension for three-way systems [13]. The perception of amplitude panning-based reproduction is plausible compared to a real sound source but there are differences, such as: reduced accuracy in localisation [14]–[16], greater sense of width [17], [18], and acoustic colouration of the media content [15], [19].

Later commercialised surround sound products, which are popular in home entertainment and cinema for their spatial impression, use proprietary reproduction technologies such as Dolby Digital [20], DTS [21], Dolby Atmos [22] and DTS:X [23]. Until the late 90's, soundfield reproduction technologies reproduced the same content over the desired area. The concept of personalised sound was introduced in the late 90's [24] which saw the research field of soundfield reproduction move towards the reproduction of individual zones of sound [25]. Over the last two decades, spatial audio and personalised sound have made advances in numerous fields, such as virtual reality (VR) [26]–[29], mobile devices [30], [31], medicine [32], [33], teleconferencing

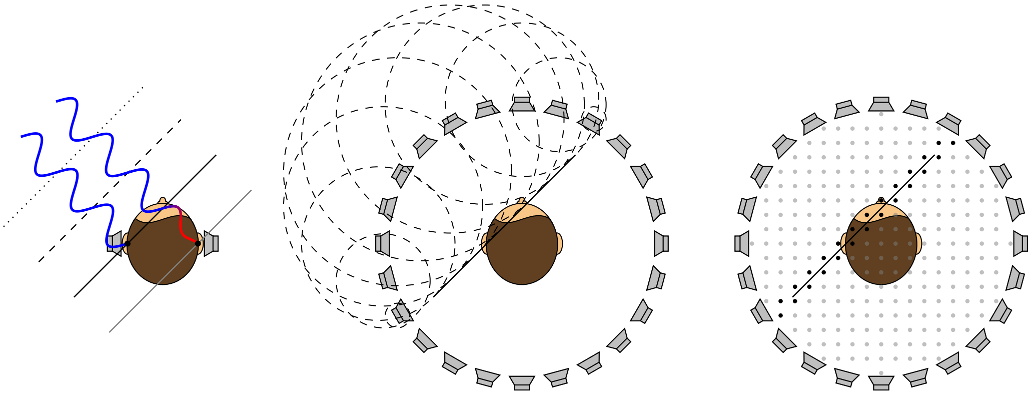

Figure 1.1: Comparison of three spatial audio reproduction techniques. Binaural reproduction (left), Kirchhoff-Hemlholtz integral-based (WFS, SDM and HOA) soundfield reproduction (middle) and point-based (SFR and LSO) soundfield reproduction (right).

[34], vehicle cabin sound [35]–[37] and active noise control [37]–[41].

1.1.1 Spatial Audio

Spatial audio can be produced by systems either at the listener's ear directly, for instance with headphones or earphones, or from larger distances using multiple loudspeakers. The former is commonly known as *binaural reproduction* and the latter as *soundfield reproduction*. A stereo loudspeaker system is one of the simplest approaches to soundfield reproduction.

Binaural audio reproductions rely on several mechanisms to provide realistic acoustic scenes for listeners [1]. Sound arriving on the azimuthal plane towards a listener's head is modelled using Interaural Time Differences (ITDs) and Interaural Level Differences (ILDs). ITDs model delays in the time of arrival of wave fronts to a point in a listener's ear. ITDs are mainly used for frequencies below 1 kHz where the listener's head provides little attenuation to the sound level. For frequencies at which the listener's head attenuates the level of the audio, generally above 1.5 kHz, ILDs are used to provide spatial impression. The ILDs are changes in the level of the sound at the listener's ears. The sound levels produced at the listener's ears are such that they match those that would have been heard by the listener for a given virtual sound at a particular location. The symmetrical nature of two receivers, such as two ears, makes it difficult to distinguish between sounds arriving from different elevations. The human pinnae has adapted in such a way that different frequencies are attenuated or amplified depending on the elevation of the source. Humans perceive sound elevation with pinnae-based spectral cues and rely on individually shaped pinnae to determine the specific elevation. The ITDs, ILDs and spectral cues can all be derived from system models known as Head-Related Transfer Functions (HRTFs). Binaural recordings are performed to capture individually tailored HRTFs which are then used for spatial audio reproduction.

Soundfield reproductions require the synchronous use of numerous loudspeakers in order to produce constructing and deconstructing sound waves. Several advanced techniques have been investigated over the last century, such as the Least Squares Optimisation (LSO) method (sometimes referred to as Sound Field Reconstruction (SFR)) [42]–[51], Wave Field Synthesis (WFS) approach [52]–[62], Spectral Division Method (SDM) [61], [63], [64] and Higher-Order Ambisonics (HOA) [65]–[75].

WFS was first theorised in the late 1980's and early 1990's as a method of acoustical holography with the underlying theory based on the Kirchhoff-Helmholtz integral [52]–[57]. The Kirchhoff-Helmholtz integral states that any soundfield can be described with the complete knowledge of the sound pressure and velocity on the enclosing boundary. WFS uses this relationship to reproduce virtual wave fields on the interior and exterior of a, theoretically, continuous distribution of secondary monopole sources [58]–[60]. Non-smooth secondary source distributions and multiple parallel linear arrays have also been integrated into the WFS framework [60], [62].

Ambisonics was first introduced in 1973 initially looking at zeroth and first order harmonics [65], [66] and was further extended to HOA in the 1990's and early 2000's [67]–[70], [72] with harmonic orders of two or greater. The fundamental idea of Ambisonics and HOA is that any soundfield can be described by a combination of soundfields resulting from the harmonic expansions of secondary source signals. Ambisonics is an alternative solution to the wave equation using the Kirchhoff-Helmholtz integral equation [2]. A desired soundfield is matched to the expansion of either cylindrical (2D) or spherical (3D) harmonics up to a finite harmonic mode limit [68], [71], [74]. The finiteness of the mode results in predictable truncation errors [73]. The concept of matching the desired soundfield using the harmonic modes is known as the *mode-matching* approach [71].

LSO based methods were initially introduced in 1964 [42] and further investigated in soundfield reproduction for their simplicity [43], [44]. Various other benefits of LSO methods have been shown and conclusions have been made which state that, in some circumstances, the discretised sampling, often involved with LSO, is better suited to discrete loudspeaker arrays [47], [49], [50]. The active control of soundfields has also benefited from the LSO based reproduction methods [45], [46], [51]. Pseudo-inversion based on singular-value decomposition (SVD), as often used in multiple-input multiple-output (MIMO) inversion solutions, has been investigated to help reduce the effect of ill-conditioned matrix inversions in LSO-based reproductions [48], [50].

1.1.2 Personal Sound Zones

The techniques described in section 1.1.1 are all focused on the reproduction of a soundfield which contains the same content across the entire region of reproduction. A single region of reproduced content is commonly called a *zone* in soundfield reproduction literature. The reproduction of more than one zone simultaneously from a single loudspeaker array system is called multizone soundfield reproduction (MSR). The process of MSR produces personal sound zones.

The reproduction of personal sound zones was first conceptualised in 1997 [24] as the reproduction of sound programs to different individuals with minimum annoyance from the other programs. The original work looked at single loudspeaker directivity, array directivity and MIMO active control. This idea was further researched and, over the following two decades, multiple new reproduction techniques were published.

Beamforming can be considered a type of multizone soundfield reproduction technique even though it is not explicitly defined as one [76]. The behaviour exhibited by beamforming techniques, where energy is focused along a single direction, results in spatially separated acoustic energy densities [77], [78]. Beamforming is traditionally not constrained to reduce sound energy in areas that are not the target beamformer direction. In 2007, Microsoft's I. Tashev, J. Droppo and M. Seltzer demonstrated the use of a uniform linear array of loudspeakers providing personal audio spaces [79], [80]. The system was based on a steerable beamformer that was designed to amplify the sounds in one area and cancel in another. The WFS and SDM based reproduction methods described in section 1.1.1 have also been extended to the multizone case by deriving wavenumber domain spatio-temporal filters to restrict sound pressure in spatial rectangular windows [81], [82]. The combination of linear and circular arrays has also been investigated for sound zoning applications [83], [84].

One of the earlier multizone reproduction techniques involved a maximisation of the energy ratio between zones and was termed acoustic contrast control (ACC) [85], [86]. The aim of the energy control method is to find loudspeaker weights that reproduce a soundfield with a maximum separation in energy between zones. The efficiency of the method was later improved by ensuring that source strengths were evenly distributed [87]. Further robustness has been considered by formulating the trade-off between performance and array effort as a regularisation problem [88]. The ACC method has been realised for use with personal computers and televisions [86] and has been adopted to reduce annoyance from mobile device audio [30], [31].

Pressure Matching (PM) approaches to MSR using the LSO discussed in section 1.1.1 and the Least Absolute Shrinkage and Selection Operator (LASSO) have been investigated for various loudspeaker array geometries [89]–[94]. The use of the LASSO is motivated by the prohibitive number of loudspeakers currently required for MSR. Using the LASSO reduces the number of simultaneously used loudspeakers by selecting only those which provide significant influence on the resulting soundfield for a given virtual source. The underlying assumption with the use of the LASSO to select loudspeakers is that the virtual source locations are fixed [92], which always results in virtual sounds coming from fixed locations for any given loudspeaker setup.

The wideband two-stage LASSO-LS algorithm selects the loudspeakers and then performs a regularised LSO with the selected loudspeakers to reproduce the desired wideband soundfield [92], [93]. Further, efficient harmonic nested dictionaries have been incorporated into the LASSO-LS algorithm to perform loudspeaker selections based on the frequency bands of interest [94].

Planarity Control (PC) was motivated by the need for a combination of soundfield synthesis methods and energy control methods [95]–[100]. The energy constraint applied in energy control methods (such as ACC) can create unpredictable distributions in pressure, whereas soundfield synthesis methods, whilst providing smoother pressure distributions, produce lower acoustic contrast between zones [96]. A cost function was formulated for the PC method which optimises both the attenuation into the quiet zone and the reproduction of the plane wave into the bright zone with limitations on the direction of wave propagation [96], [98], [101].

Harmonic Expansion (HE) can be useful for soundfield reproduction (as discussed in section 1.1.1) and has been successfully employed for MSR by spatially filtering and translating zone-based soundfield coefficients for circular and linear loudspeaker arrays [102]–[108]. The coefficient translation theorem spatially relocates the soundfield of a particular zone, defined by its coefficients, to an alternative global position [103]. To avoid unwanted leakage from one zone to another additional angular windowing can be applied to the coefficients after translation [103]. Spatial band stop filters have also been derived, as an alternative to angular windowing, which use the higher order spatial harmonics of a zone to cancel undesired effects of its lower order harmonics on other zones [104]. In [107], the prioritised control of regions was introduced to the harmonic expansion method.

Orthogonal Basis Expansion (OBE) as a method of MSR was first introduced in 2013 [109]. In the OBE method, the desired multizone soundfield is described as an orthogonal expansion of basis functions over the reproduction region [109]–[114]. Orthogonalisation on a set of plane waves, using a method of QR factorisation such as the modified (weighted) Gram-Schmidt process or Householder transform, is performed to find a set of suitable basis functions [113]. The OBE method considers both the pressure and velocity in the optimisation of soundfield coefficients. Using the modified Gram-Schmidt process, zones can be weighted with relative reproduction importance to other zones which is useful for prioritising the control of individual regions [109], [110] and for controlling leakage between zones as we will see in a later chapter.

1.2 Motivation

It is often desired that high quality and private media can be presented to individuals in shared spaces without affecting others in that shared area. Sound is difficult to control over space as it generally radiates in all directions; from both the original source and any reflections. The difficulty in controlling sound over large, and separate, areas has drawn the attention of researchers in recent years. MSR has provided realisable solutions to the spatial separation of audible media content but there is little work on the perceptual effects of MSR and the information that is carried by the soundwaves of the reproduction process. Current MSR techniques assume clean signal reproduction and do not consider an information theoretic approach to the distribution of media content, which can be heavily influenced by signal noise. The control of perceptual quality in MSRs also lacks investigation in the current literature.

In addition to the information carried by soundwaves discussed above, which may be desired to be private, the lack of acoustic absorption in an open space can allow unintended listeners to eavesdrop. The addition of reflections, commonly induced by room walls, increases the energy of the freely propagating sound waves, which may carry information and compounds the issue of reduced privacy. Control of the direct path sound and any additional reflections is required if information is to be kept private in open spaces, such as open-plan offices, restaurants/cafés, libraries and conferencing rooms.

The overall goal of this thesis is to provide feasible, practical and robust solutions to yield both high perceptual quality and high privacy acoustic environments for individuals, or small groups, in complex public acoustic settings (e.g. reverberant environments containing human talkers). This overall aim is divided into several objectives that we investigate throughout the chapters in this thesis. These objectives are:

- to efficiently process complex soundfields in a way that also considers human perception of the reproduced content in the bright and quiet zones;
- to increase the privacy between zones whilst at the same time maintaining quality for other listeners;
- to reduce the energy of soundwaves travelling between human talkers and unintended listeners, including those that are reflected off rigid walls;
- to increase the acoustic energy contrast between spaces in reproductions above that of current methods and, in particular, above the spatial aliasing frequency; and
- to facilitate the feasibility of practical implementations, i.e. reducing loudspeaker counts, reducing computational effort, etc.

We begin addressing the objectives of this thesis with novel extensions to weighted multizone reproductions. We present a method to efficiently compute multizone soundfield weights based on the pressure distributions they reproduce by using interpolated pre-computed look-up tables. We show that by using the reverse look-up method, to efficiently compute relative zone weights, it is possible to perceptually weight the frequency responses in each zone whilst reducing reproduction error in other zones. We further investigate the perceptual aspects of MSR and propose novel control methods for improving speech quality and speech privacy. New field metrics for speech quality and speech zone privacy are also proposed. An analytical solution is derived for the influence of spatial aliasing on speech privacy in shared acoustic environments. External effects on speech privacy, such as that from people talking in a room, are suppressed using novel active speech control techniques by predicting speech traversing across active loudspeaker barriers. Room reflections are also considered and new proposed methods are compared for three dimensional enclosed rigid wall rooms. The proposed active dereverberation approach does not assume any known room geometry or soundfield to be suppressed. Finally, the fundamental physical limitation that small loudspeaker numbers (and therefore spatial aliasing) impose on the acoustic separation of zones, is addressed using a novel hybrid loudspeaker method consisting of ultrasonic parametric loudspeaker arrays.

Real-world multizone soundfield reproductions were implemented and confirmed the feasibility of providing good quality and private personal sound zones. The methods and implementation of multizone soundfield reproduction systems facilitate personal sound reproduction in shared public environments without requiring physical barriers and/or wearable playback systems.

1.3 Outline

The work in this thesis is organised into 7 chapters.

In **chapter 2**, we provide an overview of fundamental acoustics theory and lead into discussion of techniques for soundfield reproduction. Green's functions are introduced along with the Kirchhoff-Helmholtz integral equation and solutions to the wave equation. The WFS, SDM and HOA approaches to soundfield reproduction are explored with respect to their solutions to the wave equation. Point-based methods, such as LSO, are also discussed. The state-of-the-art algorithms that extend single zone methods to multizone techniques are formulated and reviewed. Further, background on human perception and acoustic privacy are covered with an emphasis on speech-based acoustics. The chapter is summarised with links to work presented throughout the remainder of the thesis.

In **chapter 3**, we propose efficient methods to facilitate the practical reproductions of multizone soundfields for speech sources. An interpolation method is proposed for predicting weighting parameters of the multizone soundfield model. The pre-determined soundfields help facilitate real-time reproduction of dynamically weighted multizone reproductions. The dynamic weighting aspect of the method is further extended to dynamic perceptual weighting. It is shown that the perceptual weighting can be implemented in particular ways so as to reduce the spatial error in soundfield reproductions by considering thresholds of human hearing and auditory masking. The reduction in spatial error also corresponds to a reduction in loudspeaker signal power.

Chapter 4 covers the proposed methods that allow for high quality and private multizone reproductions. New field-based metrics are proposed for evaluating speech quality and speech privacy over soundfields. Novel optimisation algorithms, which make use of the quality and privacy field metrics, are derived for improving the intelligibility contrast between zones whilst maintaining speech quality in target reproduction zones. *A priori* estimates of acoustic contrast between zones are shown to be useful in optimising the shape of masking spectra for improving speech privacy and maintaining quality. Analytically derived descriptions of spatial aliasing artefacts, in the form of grating lobes, are used to further enhance the robustness of the optimisation algorithms. Physical implementations are realised, for the approaches in the chapter, to verify the effectiveness and feasibility of the proposed techniques.

Chapter 5 investigates new methods for the active control of undesired soundfield interference. Several techniques are proposed to help mitigate sound propagating from uncontrollable sources, such as human talkers or reflections from rigid walls. An autoregressive method is proposed to compensate for filter delay in an active soundfield cancellation system that suppresses speech across a barrier. A trade-off is shown to exist between soundfield reproduction accuracy and prediction accuracy required for non-stationary speech sources. The concept of barrier cancellation is further extended to active boundary cancellation. Two approaches to the active cancellation of sound traversing a boundary are proposed. The methods are derived using the Kirchhoff-Helmholtz integral equation as a solution to the wave equation and both consider the pressure and velocity on the boundary. The first method uses a WFS pre-filter to compensate for the approximation of dipole sources in the WFS method and the second method proposes to directly determine the velocity and pressure at the boundary using differential sources. The methods are compared and show significant suppression across active acoustic boundaries.

In **chapter 6**, a method to further improve the acoustic contrast between zones is proposed. It is based on the use of a hybrid dynamic and parametric loudspeaker system. The hybrid approach makes use of parametric loudspeakers to reproduce high frequency wave components in zones where spatial aliasing, caused by systems with few dynamic loudspeakers, would otherwise hinder the performance. Linkwitz-Riley filters are used for the cross-over between reproduction methods in the frequency domain. The filters are designed using the geometrically determined aliasing frequency of the multizone system. Results show that acoustic contrast is significantly improved above the spatial Nyquist frequency when using the proposed hybrid parametric loudspeaker approach.

Chapter 7 concludes the thesis. A discussion is given on possible future directions for the research which would further improve the practical feasibility and performance of personal sound zoning systems.

1.4 Contributions

The main contributions made in this thesis are summarised in the following list. The references provided here correspond to the list of papers published from this work (see also section 1.4.1).

- An efficient interpolation scheme is proposed for dynamically weighting zone importance in personal sound zone reproductions [115].
- The interpolation scheme is extended to a novel dynamic perceptual weighting approach based on spreading functions and human hearing thresholds [116].

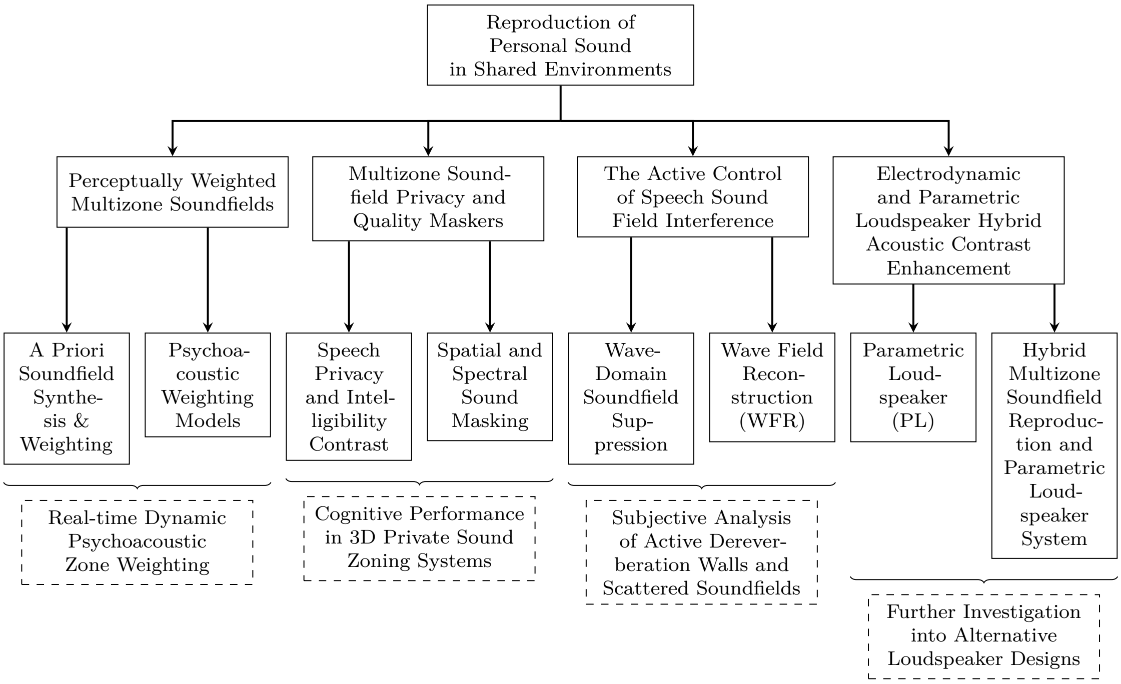
Reproduction of Personal Sound in Shared Environments
Perceptually Weighted Multizone Soundfields
Multizone Soundfield Privacy and Quality Maskers
The Active Control of Speech Sound Field Interference
Electrodynamic and Parametric Loudspeaker Hybrid Acoustic Contrast Enhancement
A Priori Soundfield Synthesis & Weighting
Psychoacoustic Weighting Models
Speech Privacy and Intelligibility Contrast
Spatial and Spectral Sound Masking
Wave-Domain Soundfield Suppression
Wave Field Reconstruction (WFR)
Parametric Loudspeaker (PL)
Hybrid Multizone Soundfield Reproduction and Parametric Loudspeaker System
Real-time Dynamic Psychoacoustic Zone Weighting
Cognitive Performance in 3D Private Sound Zoning Systems
Subjective Analysis of Active Dereverberation Walls and Scattered Soundfields
Further Investigation into Alternative Loudspeaker Designs

Figure 1.2: Thesis framework and research topics with potential future research directions in dashed blocks.

- New field metrics are proposed for speech quality and privacy over soundfields [117], [118]

- An optimisation approach to improve speech privacy using the newly defined field metrics is presented. The optimisation is further extended to include speech quality [117], [118].

- Descriptions of spatial aliasing grating lobe boundaries for multizone soundfield reproductions are analytically derived for accurate estimation of zone leakage [118].

- Multizone soundfield spectral sound maskers are analytically derived from estimates of acoustic contrast and spatial aliasing artefacts. The sound maskers are derived with a trade-off parameter for speech privacy enhancement and speech quality preservation [118].

- An active speech control method for the cancellation of speech across loudspeaker barriers is proposed [119].

- Soundfield reproduction loudspeaker weights are extended to dipole weights for speech suppression across active acoustic barriers [119].

- A novel autoregressive model is proposed for predicting non-stationary speech and is used to compensate for real-time filter delay in active soundfield control systems [119].

- A minimum-phase weighted least square (WLS) compensation pre-filter for WFS/SDM is proposed. It allows for delayless real-time suppression of acoustic reflections [120]. The WLS-based method does not assume knowledge of the room.

- A first-order differential (FOD) source/receiver model is proposed for active dereverberation [120]. The FOD-based method does not assume knowledge of the room.

- A comparison of the proposed WLS-based and FOD-based dereverberation methods is given for acoustic suppression in the time domain and frequency domain [120].

- Parametric loudspeakers are incorporated into a multizone soundfield reproduction scenario using a novel hybrid crossover approach. The hybrid approach is designed to improve acoustic contrast above the spatial aliasing frequency in multizone soundfield reproduction scenarios [121].

- A real-world multizone soundfield reproduction system is implemented to verify the feasibility of the proposed speech privacy and quality control approaches [118]. The system is realised for reproductions of frequencies up to 8 kHz, for wideband speech.

1.4.1 Publications

The following peer-reviewed publications resulted from this thesis:

- J. Donley and C. Ritz, “An efficient approach to dynamically weighted multizone wideband reproduction of speech soundfields,” in *China Summit Int. Conf. Signal Inform. Process. (ChinaSIP)*, IEEE, Jul. 2015, pp. 60–64

- J. Donley and C. Ritz, “Multizone reproduction of speech soundfields: A perceptually weighted approach,” in *Asia-Pacific Signal Inform. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, IEEE, 2015, pp. 342–345

- J. Donley, C. Ritz, and W. B. Kleijn, “Improving speech privacy in personal sound zones,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2016, pp. 311–315

- J. Donley, C. Ritz, and W. B. Kleijn, “Reproducing personal sound zones using a hybrid synthesis of dynamic and parametric loudspeakers,” in *Asia-Pacific Signal & Inform. Process. Assoc. Annu. Summit and Conf. (APSIPA ASC)*, IEEE, Dec. 2016, pp. 1–5

- J. Donley, C. Ritz, and W. B. Kleijn, “Active speech control using wave-domain processing with a linear wall of dipole secondary sources,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2017, pp. 1–5

- J. Donley, C. Ritz, and W. B. Kleijn, “Multizone soundfield reproduction with privacy- and quality-based speech masking filters,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1041–1055, 2018

- J. Donley, C. Ritz, and W. B. Kleijn, “On the comparison of two room compensation / dereverberation methods employing active acoustic boundary absorption,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, Apr. 2018, pp. 221–225

Chapter 2

Literature Review and Acoustics Theory

Overview: *This chapter provides an overview of fundamental acoustics theory for wave propagation, soundfield generation and various aspects of personalised sound. The overview provides a general discussion of advanced soundfield reproduction techniques and gives theoretical grounds for derivations and discussions presented throughout the remainder of the thesis. The chapter is focused on describing the relationship between acoustics theory of soundfield reproduction systems and personalised sound, with emphasis on perception and communication. We start by building mathematical foundations with the wave equation and Euler's equation which we lead to a derivation of Green's function and then the definition of the Kirchhoff-Helmholtz integral equation (KHIE). Mathematical definitions for soundfield reflections and reverberation are given as equivalent acoustic scattering problems. Following the acoustic fundamentals are descriptions of several approaches to soundfield reproduction that have been published in the literature over the last century. The concept of personal sound is then discussed, which forms much of the motivation for this thesis. State-of-the-art techniques for the reproduction of personal sound zones are formulated. We then consider the human perception of personal sound and its relationship to the field of psychoacoustics. The link between information theory and personal* *sound in shared environments is discussed with a focus on speech intelligibility and speech privacy. In conclusion, a summary of the chapter is provided with connections to later chapters in the thesis.*

2.1 Acoustics Theory

In this section, we cover the fundamental theory of acoustic wave propagation for arbitrary geometries, which serves to provide a firm foundation for the rest of this thesis. The mathematics discussed are based on boundary integral solutions to the wave equation.

2.1.1 The Wave Equation and Euler's Equation

The propagation of acoustic sound waves in a homogeneous medium, which contains no other sources, is defined by the time domain homogeneous acoustic wave equation. We let the acoustic pressure, $p(\mathbf{x};t)$, at any given point in space, $\mathbf{x}$, be an infinitesimal change in acoustic pressure, which satisfies the acoustic wave equation [2], [122],

$$\nabla^2 p(\mathbf{x};t) - \frac{1}{c^2}\frac{\partial^2 p(\mathbf{x};t)}{\partial t^2} = 0, \tag{2.1}$$

where c is a constant for the speed of sound in the homogeneous fluid medium. The left hand side (LHS) of (2.1) describes the source and the zero value of the right hand side (RHS) indicates that there are no other sources in the volume. One can express the Laplace operator, ∇^2, in Cartesian coordinate space as,

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}, \tag{2.2}$$

where $\mathbf{x} \equiv (x, y, z)$. The homogeneous Helmholtz equation is the acoustic wave equation (2.1) in the frequency domain, $P(\mathbf{x};\omega)$, and is given by

$$\left(\nabla^2 + k^2\right) P(\mathbf{x};\omega) = 0, \tag{2.3}$$

where $k = \omega/c$ is the acoustic wave number, $\omega = 2\pi f$ is the angular frequency and f is the temporal frequency.

The direction of the velocity of particles in the fluid medium is represented as a vector quantity, $\boldsymbol{v}$. In the time domain, the velocity vector and the sound pressure are related to one another by Euler's equation as [2], [122]

$$\rho_0 \frac{\partial \boldsymbol{v}}{\partial t} = -\nabla p(\mathbf{x}; t), \tag{2.4}$$

where ρ_0 is the fluid density when there is zero change in the medium, i.e. in equilibrium. The velocity vector, with components $\dot{u}$, $\dot{v}$ and $\dot{w}$, is given by

$$\boldsymbol{v} = \dot{u}\hat{\imath} + \dot{v}\hat{\jmath} + \dot{w}\hat{k}, \tag{2.5}$$

where the unit vectors in the x, y and z directions are $\hat{\imath}$, $\hat{\jmath}$ and $\hat{k}$, respectively. The spatial gradient is denoted with the nabla and is defined in Cartesian coordinates in terms of the unit vectors as

$$\nabla \equiv \frac{\partial}{\partial x}\hat{\imath} + \frac{\partial}{\partial y}\hat{\jmath} + \frac{\partial}{\partial z}\hat{k}. \tag{2.6}$$

Performing a Fourier transform on Euler's equation from (2.4) yields, in the frequency domain,

$$i\omega\rho_0\boldsymbol{v} = \nabla P(\mathbf{x}; \omega). \tag{2.7}$$

2.1.2 Green's Function

Consider an infinitesimally small point that is a source of acoustic wave energy, we call this a *point source.* In an unbounded volume, the solution to the inhomogeneous Helmholtz equation is a point source [2], [122]

$$\left(\nabla^2 + k^2\right)\boldsymbol{\Phi} = \Box G(\mathbf{x}, \mathbf{x}'; k) = \delta(\mathbf{x} - \mathbf{x}'), \tag{2.8}$$

where $\mathbf{\Phi}$ is a solution to the Helmholtz equation, $\Box$ is the d'Alembert operator, $G(\mathbf{x}, \mathbf{x}'; k) : \Omega \times \mathbb{R} \to \mathbb{C}$ is the free space Green's function for the non-homogeneous wave equation, $\mathbf{x}'$ is the location of the point source and the impulse is denoted by the multidimensional Dirac delta function, $\delta(\cdot)$. The general solution to (2.8) is

$$\mathbf{\Phi} = \mathbf{\Phi}_p + \mathbf{\Phi}_h, \tag{2.9}$$

where the particular solution is $\mathbf{\Phi}_p$ and the homogeneous solution is $\mathbf{\Phi}_h$. We set the homogeneous solution, $\mathbf{\Phi}_h$, to zero and obtain the free space Green's function with positive time dependency,

$$\mathbf{\Phi} = \mathbf{\Phi}_p = G(\mathbf{x}, \mathbf{x}'; k) = \frac{\exp(ik\|\mathbf{x} - \mathbf{x}'\|)}{4\pi\|\mathbf{x} - \mathbf{x}'\|}, \tag{2.10}$$

where $\exp(\cdot)$ is the exponentiation of Euler's number and $\|\cdot\|$ is the Euclidean, or ℓ^2, norm. The homogeneous solution, $\mathbf{\Phi}_h$, from (2.9), is any solution to the homogeneous Helmholtz equation (2.3).

2.1.3 Green's Theorem

A volume in three dimensional space, $\Omega \subset \mathbb{R}^3$, with a bounding surface, $\mathcal{C} \equiv \partial\Omega$, is assumed. We let $\mathbf{x}$ be any point inside Ω, i.e. $\mathbf{x} \in \Omega$, and $\mathbf{x}_0 \in \mathcal{C}$ a point on the surface. Within the volume, Ω, we have two finite and continuous functions whose first and second partial derivatives are also finite and continuous. We name the two functions $\mathbf{\Phi}(\mathbf{x}; k)$ and $\mathbf{\Psi}(\mathbf{x}; k)$. Green's theorem, which is Green's second identity, then applies as [2],

$$\iiint_{\Omega} \Big(\mathbf{\Phi}(\mathbf{x}; k) \nabla^2 \mathbf{\Psi}(\mathbf{x}; k) - \mathbf{\Psi}(\mathbf{x}; k) \nabla^2 \mathbf{\Phi}(\mathbf{x}; k) \Big)\, d\Omega$$

$$= \iint_{\mathcal{C}} \left(\mathbf{\Phi}(\mathbf{x}; k) \frac{\partial \mathbf{\Psi}(\mathbf{x}; k)}{\partial \mathbf{n}} - \mathbf{\Psi}(\mathbf{x}; k) \frac{\partial \mathbf{\Phi}(\mathbf{x}; k)}{\partial \mathbf{n}} \right) d\mathcal{C}, \tag{2.11}$$

where the derivative (gradient) with respect to the outward facing normal, $\mathbf{n}$, is $\partial/\partial\mathbf{n}$. In the direction perpendicular to the surface that is $\mathbf{n}$, this gradient is the rate of change of $\mathbf{\Phi}(\mathbf{x};k)$ or $\mathbf{\Psi}(\mathbf{x};k)$.

If we assume that the homogeneous Helmholtz equation is satisfied on the surface and in the volume by the two functions, $\mathbf{\Phi}(\mathbf{x};k)$ and $\mathbf{\Psi}(\mathbf{x};k)$, and the two functions have no singularities within the bounding surface or on it, then the LHS of (2.11) becomes

$$\begin{aligned}\mathbf{\Phi}(\mathbf{x};k)\nabla^2\mathbf{\Psi}(\mathbf{x};k)-\mathbf{\Psi}(\mathbf{x};k)\nabla^2\mathbf{\Phi}(\mathbf{x};k)&\\ &= \mathbf{\Phi}(\mathbf{x};k)\Big(-k^2\mathbf{\Psi}(\mathbf{x};k)\Big)-\mathbf{\Psi}(\mathbf{x};k)\Big(-k^2\mathbf{\Phi}(\mathbf{x};k)\Big)\\ &= 0,\end{aligned} \tag{2.12}$$

which leaves us with

$$\iint\limits_{\mathcal{C}}\left(\mathbf{\Phi}(\mathbf{x};k)\frac{\partial\mathbf{\Psi}(\mathbf{x};k)}{\partial\mathbf{n}}-\mathbf{\Psi}(\mathbf{x};k)\frac{\partial\mathbf{\Phi}(\mathbf{x};k)}{\partial\mathbf{n}}\right)d\mathcal{C}=0. \tag{2.13}$$

The above equation, (2.13), is the foundation for deriving the Kirchhoff-Helmholtz integral equation.

2.1.4 The Kirchhoff-Helmholtz Integral Equation

The Kirchhoff-Helmholtz integral equation (KHIE) is the premise of many areas of acoustics. It states that the complete knowledge of acoustic sound pressure and velocity on the surface of a volume is sufficient to fully describe the pressure and velocity within that volume. The KHIE is often posed as a solution to both the interior and exterior problems. The interior problem is applicable to scenarios where acoustic sound fields are enclosed within boundaries and the exterior problem is applicable to radiation and scattering. The interior and exterior problem are shown in Figure 2.1 and Figure 2.2, respectively.

There are three surfaces considered in the exterior problem: the arbitrarily

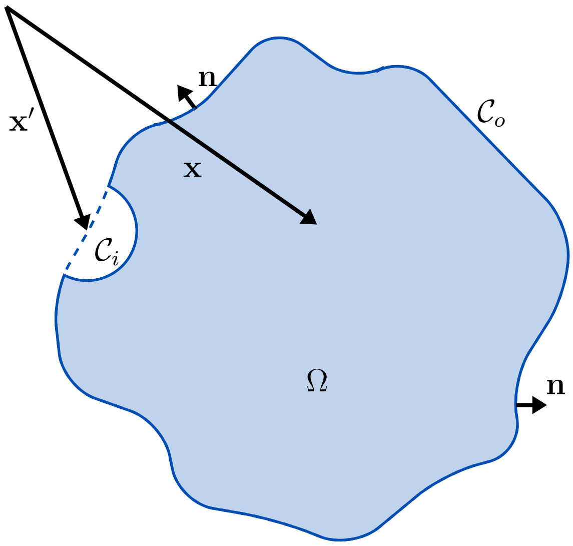

Figure 2.1: The shaded region depicts the source-free volume, Ω, for the interior domain. The volume is bounded by the surface $\mathcal{C}_o$ and $\mathbf{n}$ shows unit vectors normal to the surface. In particular, the figure illustrates the KHIE derivation when the evaluation point is located on the surface $\mathcal{C}_o$, as shown by $\mathcal{C}_i$.

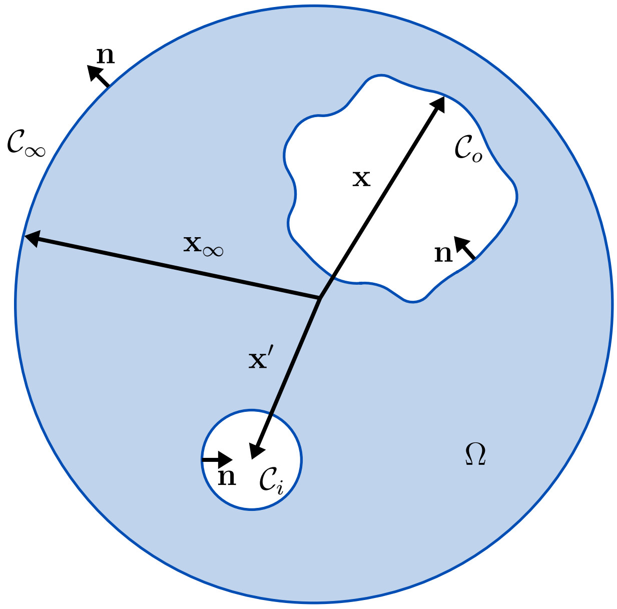

Figure 2.2: The shaded region depicts the source-free volume, Ω, for the exterior domain. The volume extends to infinity and encloses all evaluation points inside the surface $\mathcal{C}_\infty$. Unit vectors normal to the surface are shown by $\mathbf{n}$. These region definitions are useful for solving radiation and scattering problems.

shaped outer surface $\mathcal{C}_o$, which encloses the sound sources; the infinitesimally small sphere's surface $\mathcal{C}_i$, which contains the evaluation point; and the infinitely distant surface $\mathcal{C}_\infty$, which vanishes by the *Sommerfeld radiation condition.* For the exterior domain, the complete surface is,

$$\mathcal{C} = \mathcal{C}_o + \mathcal{C}_i + \mathcal{C}_\infty, \tag{2.14}$$

and is used with Green's theorem from (2.13).

The interior and exterior KHIE can be found using the inhomogeneous Helmholtz equation from (2.8) and Green's second identity from (2.11) as [2]

$$\breve{\alpha} P(\mathbf{x}'; k) = \iint\limits_{\mathcal{C}_o} \left(i\rho_0 ck G(\mathbf{x}, \mathbf{x}'; k) v_{\mathbf{n}}(\mathbf{x}; k) - P(\mathbf{x}; k) \frac{\partial G(\mathbf{x}, \mathbf{x}'; k)}{\partial \mathbf{n}} \right) d\mathcal{C}_o, \tag{2.15}$$

where

$$\breve{\alpha} = \underbrace{\begin{cases} 1, & \text{if } \mathbf{x}' \text{ is inside } \mathcal{C}_o \\ 1/2, & \text{if } \mathbf{x}' \text{ is on } \mathcal{C}_o \\ 0, & \text{if } \mathbf{x}' \text{ is outside } \mathcal{C}_o \end{cases}}_{\text{interior solution (Figure 2.1)}} = \underbrace{\begin{cases} 0, & \text{if } \mathbf{x}' \text{ is inside } \mathcal{C}_o \\ 1/2, & \text{if } \mathbf{x}' \text{ is on } \mathcal{C}_o \\ 1, & \text{if } \mathbf{x}' \text{ is outside } \mathcal{C}_o, \end{cases}}_{\text{exterior solution (Figure 2.2)}} \tag{2.16}$$

and $v_{\mathbf{n}}(\mathbf{x}; k)$ is the acoustic particle velocity in the normal direction, $\mathbf{n}$, to the bounding surface. We note that $\breve{\alpha}$ determines the result of (2.15) depending on the location of $\mathbf{x}'$, i.e. the soundfield is zero on one side of the boundary and is completely defined on the other side. The integral in (2.15) consists of a single layer potential and a double layer potential and the result in (2.16) is the jump relation. It is worth pointing out that (2.15) takes on half its value for points on the boundary.

2.1.5 Simple Source Formulation and Alternative Green's Functions

One potential disadvantage to using the KHIE from (2.15) directly is that it requires the complete knowledge of the acoustic propagation on the surface. That is to say, it requires both the sound pressure, $P(\mathbf{x};k)$, and normal particle velocity, $v_{\mathbf{n}}(\mathbf{x};k)$, on $\mathcal{C}_o$.

There are a few techniques which can be used to avoid this problem. They work by reducing the formulation of (2.15) so as to require only one of either $P(\mathbf{x};k)$ or $v_{\mathbf{n}}(\mathbf{x};k)$. The latter of the two techniques described in this section lead to Rayleigh's first and second integral [2], [123], [124].

Simple Source Formulation

One of the techniques is known as the *simple source formulation* and makes two assumptions; the first is that the pressure inside, $P_i(\mathbf{x};k)$, and outside, $P_o(\mathbf{x};k)$, the boundary surface are linked such that they are equal; and the second is that the gradient on boundary, inside and outside, are not equal. By subtracting the interior and exterior solutions of (2.15) given by (2.16) we get

$$P(\mathbf{x}';k) = \iint\limits_{\mathcal{C}_o} \left(\frac{\partial P_o(\mathbf{x};k)}{\partial \mathbf{n}} - \frac{\partial P_i(\mathbf{x};k)}{\partial \mathbf{n}} \right) G(\mathbf{x},\mathbf{x}';k)\, d\mathcal{C}_o, \tag{2.17}$$

such that $\frac{\partial P_o(\mathbf{x};k)}{\partial \mathbf{n}} \neq \frac{\partial P_i(\mathbf{x};k)}{\partial \mathbf{n}}$.

We set the difference in gradients, given by the normal derivatives, as

$$\mu(\mathbf{x};k) \equiv \frac{\partial P_o(\mathbf{x};k)}{\partial \mathbf{n}} - \frac{\partial P_i(\mathbf{x};k)}{\partial \mathbf{n}}, \tag{2.18}$$

so that (2.17) simply becomes

$$P(\mathbf{x}';k) = \iint\limits_{\mathcal{C}_o} \mu(\mathbf{x};k) G(\mathbf{x},\mathbf{x}';k)\, d\mathcal{C}_o. \tag{2.19}$$

In (2.19), $P(\mathbf{x}';k)$ is the pressure anywhere in space and (2.19) can be used for applications where the internal and external problem do not need to be considered together. The result that is (2.19) reduces the problem to one of either the internal or external domains in order to simplify the source distribution to $\mu(\mathbf{x};k)$.

A common method in soundfield reproduction is to define $P(\mathbf{x}';k)$ and use (2.19) to determine the signals for the sources on the boundary, which are $\mu(\mathbf{x};k)$. The simple source formulation is used in various reproduction approaches such as the LSO method and Ambisonics. This formulation is discussed further throughout the thesis for the generation of single or multiple zone soundfields.

Neumann-Green's Function

Another technique used to simplify the formulation of the KHIE is to consider an alternative Green's function, called the *Neumann-Green's function*, $G_{\mathcal{N}}(\mathbf{x},\mathbf{x}';k)$, which eliminates one of the terms in (2.15). The Neumann-Green's function is a sum of the Green's function, $G(\mathbf{x},\mathbf{x}';k)$, and a non-zero homogeneous solution, $\boldsymbol{\Psi}_h$ (obtained analogously to (2.9)), of (2.13). For the Neumann problem, we invoke the boundary condition

$$\frac{\partial G_{\mathcal{N}}(\mathbf{x},\mathbf{x}';k)}{\partial \mathbf{n}} = 0, \tag{2.20}$$

over the entire $\mathcal{C}_o$, which results in (2.15) simplifying to

$$\breve{\alpha} P(\mathbf{x}';k) = i\rho_0 ck \iint\limits_{\mathcal{C}_o} G_{\mathcal{N}}(\mathbf{x},\mathbf{x}';k) v_{\mathbf{n}}(\mathbf{x};k)\, d\mathcal{C}_o. \tag{2.21}$$

The most common method in soundfield reproduction for finding an appropriate $G_{\mathcal{N}}(\mathbf{x},\mathbf{x}';k)$ is to consider (2.21) for an infinite planar boundary. In practice this relates well to planar loudspeaker or microphone arrays as we will see in later chapters. If we consider $\mathcal{C}_o$ to be a planar boundary along two spatial dimensions, then we can also define a mirror image of the evaluation point $\mathbf{x}'$ about $\mathcal{C}_o$, which we denote as $\mathbf{x}'_i$. The mirror image is a solution to the homogeneous Helmholtz equation that can be added to the Green's function $G(\mathbf{x},\mathbf{x}';k)$ to give the Neumann-Green's function for a plane

$$\begin{aligned} G_{\mathcal{N}}(\mathbf{x},\mathbf{x}';k) &= \frac{1}{4\pi}\left(\frac{\exp(ik\|\mathbf{x}-\mathbf{x}'\|)}{\|\mathbf{x}-\mathbf{x}'\|} + \frac{\exp(ik\|\mathbf{x}-\mathbf{x}'_i\|)}{\|\mathbf{x}-\mathbf{x}'_i\|}\right) \\ &= \frac{1}{2\pi}\frac{\exp(ik\|\mathbf{x}-\mathbf{x}'\|)}{\|\mathbf{x}-\mathbf{x}'\|} \\ &= 2G(\mathbf{x},\mathbf{x}';k), \end{aligned} \tag{2.22}$$

which is true for all points on the plane, since $\|\mathbf{x}-\mathbf{x}'\| = \|\mathbf{x}-\mathbf{x}'_i\|$. Substituting (2.22) into (2.21) results in Rayleigh's first integral for a plane [2]

$$P(\mathbf{x}';k) = -2i\rho_0 ck \iint\limits_{\mathcal{C}_o} G(\mathbf{x},\mathbf{x}';k)v_{\mathbf{n}}(\mathbf{x};k)\,d\mathcal{C}_o \tag{2.23}$$

and since the normal points outwards from the desired half-space, we negate (2.21) prior to the substitution. The $\breve{\alpha}$ term is dropped because at all points its value arrives at unity. The simplified solution, that is (2.23), forms the basis of the WFS method which we will discuss further later in this chapter.

Dirichlet-Green's Function

While we have comprehensively covered two techniques to simplify the KHIE, and which are applicable to most soundfield reproduction methods, we cover a third for completeness and to aid in understanding derivations in later chapters. In this technique, we consider another alternative Green's function, called the *Dirichlet-Green's function*, $G_{\mathcal{D}}(\mathbf{x},\mathbf{x}';k)$, which instead eliminates $i\rho_0 ckG(\mathbf{x},\mathbf{x}';k)v_{\mathbf{n}}(\mathbf{x};k)$ from the KHIE in (2.15). Similar to the Neumann-Green's function, the Dirichlet-Green's function is a summation of the Green's function and a non-zero homogeneous solution of (2.13). For the Dirichlet problem, we invoke the boundary condition

$$G_{\mathcal{D}}(\mathbf{x},\mathbf{x}';k) = 0, \tag{2.24}$$

over the entire $\mathcal{C}_o$. For this case, we have (2.15) reduce to

$$\breve{\alpha} P(\mathbf{x}'; k) = -\iint\limits_{\mathcal{C}_o} P(\mathbf{x}; k) \frac{\partial G_{\mathcal{D}}(\mathbf{x}, \mathbf{x}'; k)}{\partial \mathbf{n}} \, d\mathcal{C}_o. \tag{2.25}$$

Once again, considering a planar boundary as in the section above, we have a mirror image and the Dirichlet-Green's function is expressed as

$$G_{\mathcal{D}}(\mathbf{x}, \mathbf{x}'; k) = \frac{1}{4\pi} \left(\frac{\exp(ik\|\mathbf{x} - \mathbf{x}'\|)}{\|\mathbf{x} - \mathbf{x}'\|} - \frac{\exp(ik\|\mathbf{x} - \mathbf{x}_i'\|)}{\|\mathbf{x} - \mathbf{x}_i'\|} \right), \tag{2.26}$$

for all points on the plane. Substituting (2.26) into (2.25) results in Rayleigh's second integral for a plane [124]

$$P(\mathbf{x}'; k) = 2 \iint\limits_{\mathcal{C}_o} P(\mathbf{x}; k) \Big(\|\mathbf{x} - \mathbf{x}'\|^{-1} + ik \Big) \cos(\varphi) \, G(\mathbf{x}, \mathbf{x}'; k) \, d\mathcal{C}_o, \tag{2.27}$$

where φ is the angle between $(\mathbf{x} - \mathbf{x}')$ and $\mathbf{n}$.

The two Rayleigh integrals are used in many acoustics problems and are the fundamental concept of WFS and SDM sound reproduction. While Rayleigh's integrals are generally applicable to planar arrays, they have also been extended to line arrays and circular arrays, as will be discussed later in this chapter. The idea of the mirror image source is also relevant in that it is a fundamental concept of reflections and reverberation [125], caused by rigid or partially absorbing boundaries, and will be covered later in the thesis. The simplified sources derived in this section form the basis for most state-of-the-art soundfield reproduction techniques.

2.1.6 The Acoustic Scattering Problem

The KHIE can be used to generalise the formulation of acoustic reflections from a body, known as *scattering*. The concept of this generalisation is straightforward; we consider a new region which contains a point source for an incident field, and this point source defines the acoustic radiation that becomes our scattered acoustic field.

The radiation from the new point source, which is contained in a sphere with surface $\mathcal{C}_p$ and radius ϵ, is obtained from the limit of the KHIE as $\epsilon \to 0$, with $d\mathcal{C}_p = \epsilon^2 d\Theta$ ($d\Theta \equiv \sin\theta \, d\theta \, d\phi$),

$$\begin{aligned}
&\lim_{\epsilon\to 0} \iint\limits_{\mathcal{C}_p} \left(G(\mathbf{x}, \mathbf{x}'; k) \frac{\partial \mathbf{\Phi}}{\partial \mathbf{n}} - \mathbf{\Phi} \frac{\partial G(\mathbf{x}, \mathbf{x}'; k)}{\partial \mathbf{n}} \right) d\mathcal{C}_p \\
&= G(\mathbf{x}_{\mathrm{p}}, \mathbf{x}'; k) \lim_{\epsilon\to 0} \iint\limits_{\mathcal{C}_p} \frac{\partial \mathbf{\Phi}}{\partial \mathbf{n}} \epsilon^2 d\Theta = \mathbf{\Phi}_0 G(\mathbf{x}_{\mathrm{p}}, \mathbf{x}'; k) \equiv P_p(\mathbf{x}'; k).
\end{aligned} \tag{2.28}$$

The total pressure field can be written as

$$P(\mathbf{x}'; k) = P_p(\mathbf{x}'; k) + P_s(\mathbf{x}'; k), \tag{2.29}$$

and so the KHIE for the scattering problem is

$$\breve{\alpha} P(\mathbf{x}'; k) = P_p(\mathbf{x}'; k) + \iint\limits_{\mathcal{C}_o} \left(i\rho_0 ck G(\mathbf{x}, \mathbf{x}'; k) v_{\mathbf{n}}(\mathbf{x}; k) - P(\mathbf{x}; k) \frac{\partial G(\mathbf{x}, \mathbf{x}'; k)}{\partial \mathbf{n}} \right) d\mathcal{C}_o. \tag{2.30}$$

This formulation is helpful for when $\mathcal{C}_o$ is planar as the geometry closely resembles that in typical rooms where rigid walls scatter acoustic waves. If the shape of scattering objects is known then (2.30) provides a good model for the scattered soundfield. In chapter 5, we will investigate the use of active planar arrays for suppressing acoustic scattering/reflections from room walls.

2.2 Soundfield Generation Using Loudspeakers

A system that generates a soundfield over an extended spatial region of interest, using a set of loudspeakers, is known as a soundfield reproduction system. There are numerous methods for designing such systems, which have been established, collectively, over the last century. In this section, we review the main approaches to soundfield reproduction: amplitude panning, least squares optimisation (LSO), wave field synthesis (WFS) and higher-order Ambisonics (HOA). Throughout this section we use S to denote a soundfield that is to be reproduced by practical discrete source/receiver distributions rather than P, which was used for continuous theoretical distributions in the previous section. Throughout the thesis these notations may change but the meaning will be made clear prior to their use and, generally, from the context.

2.2.1 Amplitude Panning

The concept of providing high quality spatial audio to listeners was made popular with the introduction of stereophony, which was developed in the early 1930's [6]. The idea behind stereophony was to provide a sense of spatial presence in sound reproductions by using a combination of delay and level differences [6], [7]. Amplitude panning is the most common approach to stereophony and is often used in commercialised surround sound systems due to the accurate localisation that it offers [8]–[10].

We start with amplitude panning for a stereophonic system. A gain for each of the two loudspeakers are specified as: W_1 for the left loudspeaker; and W_2 for the right loudspeaker. We call the angle to the left loudspeaker from the listener $-\phi_0$ and to the right loudspeaker $+\phi_0$. The angle to the virtual source from the listener we call θ_v. The soundfield reproduced from the stereophonic system in the x direction is given by

$$S(\mathbf{x}; k_x) = s(\mathbf{x}; k_x)(W_1 \mathrm{exp}(-ik_x x) + W_2 \mathrm{exp}(ik_x x)), \tag{2.31}$$

where $s(\mathbf{x}; k_x)$ is the spectrum of the desired virtual source and k_x is the wavenumber in the x direction. The *sine law* and *tangent law* of stereophony are amplitude panning laws governing the relationship between loudspeaker gains [126], [127] and are given by

$$\frac{1 - \frac{W_1}{W_2}}{1 + \frac{W_1}{W_2}} = \underbrace{\frac{\sin(\theta_\mathrm{v})}{\sin(\phi_0)}}_{\text{sine law}} \approx \underbrace{\frac{\tan(\theta_\mathrm{v})}{\tan(\phi_0)}}_{\text{tangent law}}, \tag{2.32}$$

where W_1/W_2 can be specified to find θ_v or vice versa.

Vector-based Amplitude Panning (VBAP)

An extension of stereophony to multiple channels is known as VBAP [128]. We denote the vector to the virtual source as

$$\mathbf{v} = \begin{bmatrix} \sin(\theta_{\mathrm{v}}) \\ \cos(\theta_{\mathrm{v}}) \end{bmatrix}, \tag{2.33}$$

the loudspeaker gains in vector form as

$$\mathbf{W} = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \tag{2.34}$$

and a matrix of vectors to each loudspeaker as

$$\mathbf{L} = \begin{bmatrix} \mathbf{l}_1 & \mathbf{l}_2 \end{bmatrix} = \begin{bmatrix} -\sin(\theta_{\mathrm{v}}) & \sin(\theta_{\mathrm{v}}) \\ \cos(\theta_{\mathrm{v}}) & \cos(\theta_{\mathrm{v}}) \end{bmatrix} \tag{2.35}$$

where $\mathbf{l}_1$ and $\mathbf{l}_2$ are the left and right loudspeaker vectors, respectively. The following equation gives the loudspeaker gains for the VBAP method

$$\mathbf{W} = \mathbf{L}^{-1}\mathbf{v}. \tag{2.36}$$

While the VBAP method is conceptually and mathematically straightforward, it lacks the depth of control of the soundfield over the spatial region. The VBAP only considers the sound along the vectors that it makes use of, i.e. it only considers the sound from the loudspeakers to a single point in space. For this reason, we continue on to more advanced soundfield control techniques which aim to cover the complete region of interest.

2.2.2 Least Squares Optimisation (LSO)

Another approach to soundfield reproduction is the LSO method, which was initially introduced in 1964 [42]. The aim of the method is to determine the loudspeakers weights that would reproduce a specified soundfield over an area with minimal error by finding the least square solution. The method has been investigated for determining the minimum number of required loudspeakers [43], [44] and, more recently, it has been used extensively for the active control of soundfields [45], [46], [51]. There has been work showing further benefits, such as improved performance in wide listening areas and better results on average when compared to WFS, as well as some concluding that LSO is better suited to discrete loudspeaker arrays [47], [49], [50]. As we will see in this section, the mathematical derivations are relatively simple, however, the matrix inversion that is a fundamental part of the method can lead to poor reproductions when that matrix is ill-conditioned [48]. Pseudo-inversion based on SVD, commonly used in MIMO inversion solutions, has been studied to reduce the effect of the problem in LSO-based techniques [50].

As was mentioned above, the LSO method aims to reproduce a soundfield, $S^{\mathrm{a}}(\mathbf{x};k)$, that matches, in a least squares sense, a desired soundfield, $S^{\mathrm{d}}(\mathbf{x};k)$. Let us start by assuming we have Z microphones and L loudspeakers. The microphone measurements produce a vector, $\mathbf{s}^{\mathrm{a}}(k)$, that describes $S^{\mathrm{a}}(\mathbf{x};k)$, and a vector, $\mathbf{s}^{\mathrm{d}}$, that describes $S^{\mathrm{d}}(\mathbf{x};k)$. The relationship between $\mathbf{s}^{\mathrm{a}}$ and the loudspeaker driving signals in vector form, $\mathbf{q}(k)$, is

$$\mathbf{s}^{\mathrm{a}}(k) = \mathbf{T}(k)\mathbf{q}(k), \tag{2.37}$$

where $\mathbf{T}(k)$ is a matrix of pairwise acoustic transfer functions, for each microphone and loudspeaker, whose size is $Z \times L$. The effects of reverberation on $\mathbf{T}(k)$ caused by room walls is investigated in [113].

The goal is to now minimise the error between $\mathbf{s}^{\mathrm{a}}$ and $\mathbf{s}^{\mathrm{d}}$. We would like to find a new solution for $\mathbf{q}(k)$ that would accomplish the minimisation of error. The minimisation is then

$$\min_{\mathbf{q}(k)} \left\| \mathbf{s}^{\mathrm{d}} - \mathbf{T}\mathbf{q} \right\|^2, \tag{2.38}$$

whose least square solution is

$$\widehat{\mathbf{q}}(k) = \mathbf{T}^{\dagger}\mathbf{s}^{\mathrm{d}}, \tag{2.39}$$

where $\widehat{\mathbf{q}}(k)$ are the new loudspeaker driving signals and $\{\cdot\}^{\dagger}$ is the Moore-Penrose pseudo-inverse [129]. When $\mathbf{T}(k)$ is full rank with linearly independent columns the Moore-Penrose pseudo inverse can be expressed simply as

$$\mathbf{T}^{\dagger} = \left(\mathbf{T}^{H}\mathbf{T}\right)^{-1}\mathbf{T}^{H}, \tag{2.40}$$

and for linearly independent rows it is

$$\mathbf{T}^{\dagger} = \mathbf{T}^{H}\left(\mathbf{T}\mathbf{T}^{H}\right)^{-1}, \tag{2.41}$$

where $\{\cdot\}^{H}$ is a Hermitian transposition (conjugate transpose).

The LSO method can be seen to match the individual responses at discrete locations over the soundfield in the least squares sense as described above. The responses at the discrete locations are the pressure signals received and so this method is also referred to as *pressure matching*. Any method that makes use of the LSO is susceptible to the ill-conditioning problem [48], which occurs in (2.39) from the inverse in (2.40) or (2.41). There has been some work on solving the ill-conditioning problem for soundfield reproduction scenarios, such as truncated singular value decomposition (SVD) [130], [131] and Tikhonov regularisation [132]. We will see later in this chapter that other soundfield reproduction techniques utilise the LSO, however, they perform the minimisation in a different domain and for this reason they are considered a separate approach.

2.2.3 Wave Field Synthesis (WFS)

WFS was first theorised in the late 1980's and early 1990's as a method of acoustical holography with the underlying theory based on the KHIE [52]–[57]. As we discussed previously in 2.1.4, the KHIE states that any soundfield can be described with the knowledge of the monopole pressure and pressure gradient on its enclosing boundary. The pressure gradient, or velocity vector, can be obtained using dipole sources or receivers, however, it is not always straightforward to implement them in practice. Loudspeaker responses are, in practice, very similar to monopole sources at low frequencies [133]. For this reason, WFS uses the formulation described in 2.1.5 to eliminate the necessity of the dipole term in the KHIE and rely entirely on the monopole term. However, by using (2.23), $\breve{\alpha}$ is dropped and the soundfield is non-zero where (2.16) previously specified it would be zero. WFS uses (2.23) to reproduce virtual wave fields within a, theoretically, continuous distribution of secondary monopole sources [58]–[61]. Non-smooth secondary source distributions and multiple parallel linear arrays have also been integrated into the WFS framework [60], [62].

Although WFS is based on (2.23), it is usually expressed in terms of a loudspeaker driving function, $Q(\mathbf{l};k)$, at loudspeaker position, $\mathbf{l} \in \mathcal{C}_o$. The driving functions are designed to reproduce a soundfield, $S^{\mathrm{a}}(\mathbf{x};k)$, which matches the desired soundfield, $S^{\mathrm{d}}(\mathbf{x};k)$, of the virtual source, so from (2.7) and (2.23) we have,

$$S^{\mathrm{a}}(\mathbf{x};k) = \iint\limits_{\mathcal{C}_o} Q(\mathbf{l};k) G(\mathbf{x},\mathbf{l};k)\, d\mathbf{l}, \tag{2.42}$$

and the loudspeaker driving function is

$$Q(\mathbf{l};k) = 2\frac{\partial S^{\mathrm{d}}(\mathbf{x};k)}{\partial \mathbf{n}}. \tag{2.43}$$

The virtual source soundfield is synthesised in the half-space using the loudspeaker driving function, $Q(\mathbf{l};k)$.

An approximate solution can be found for curved surfaces by applying the Kirchhoff approximation [134], [135]. The approximation holds as long as sound reproduced by secondary sources does not re-enter the volume, i.e. the formulation holds true only for convex secondary source geometries. If we assume that a curved surface can be modelled as a set of smaller planar surfaces whose width is much larger than the wavelength then we see that only part of the surface is active for a given virtual source soundfield. A window function, $\widehat{a}(\mathbf{l};k)$, is introduced to model the active parts of the surface [60], [135],

$$Q(\mathbf{l};k) = 2\widehat{a}(\mathbf{l};k)\frac{\partial S^{\mathrm{d}}(\mathbf{x};k)}{\partial \mathbf{n}}. \tag{2.44}$$

The derivation of the loudspeaker driving function in (2.44) is known as the *Rayleigh formulation* of WFS [61].

The WFS method has the advantage that it is applicable to arbitrary geometries and creates wide regions of accurate broadband reproduction below the spatial Nyquist frequency. Historically, WFS has been considered the solution for reproduction over a large area [69] and has been researched extensively for reproduction over a horizontal plane [56], [58], [61], [136]–[141]. We will see in chapter 5 that the wide reproduction region of WFS-based methods is useful when cancelling reflections off walls.

Spectral Division Method (SDM)

The spectral division method is a technique that uses the spatial Fourier transform to obtain the required loudspeaker driving function for soundfield reproduction [61], [63]. It has been shown that WFS constitutes an approximation of the exact solution given by the SDM [64]. Once again, assuming we have a planar array of loudspeakers with positions, $\mathbf{l} \in \mathcal{C}_o$, the spatial Fourier transform of (2.42) with respect to the dimensions of the array gives the reproduced SDM soundfield as,

$$\widetilde{S}^{\mathrm{a}}(k_{\mathbf{l}}, \mathbf{x}_{\mathbf{n}}, k) = \widetilde{Q}(k_{\mathbf{l}}, k)\widetilde{G}(k_{\mathbf{l}}, \mathbf{x}_{\mathbf{n}}, k), \tag{2.45}$$

where a tilde indicates a function in the spatial frequency domain obtained from a spatial Fourier transform, $k_{\mathbf{l}}$ is the wavenumber along the dimensions of the loudspeaker array and $\mathbf{x_n} \triangleq \mathbf{x} \cdot \mathbf{n}$ is the location in the normal direction to the loudspeaker array.

The forward spatial Fourier transform of $S^{\mathrm{a}}(\mathbf{x}; k)$ with respect to the plane of interest is

$$\widetilde{S}^{\mathrm{a}}(k_{\mathbf{l}}, \mathbf{x_n}, k) = \iint\limits_{\mathcal{C}_o} S^{\mathrm{a}}(\mathbf{x}; k) \exp(i k_{\mathbf{l}} \mathbf{l}) \, d\mathbf{l}, \tag{2.46}$$

where a positive exponent is used according to [61], [63]. $\widetilde{Q}(k_{\mathbf{l}}, k)$ and $\widetilde{G}(k_{\mathbf{l}}, \mathbf{x_n}, k)$ are found using the same Fourier transform.

Rearranging (2.45) for the driving function from the desired soundfield and we get

$$\widetilde{Q}(k_{\mathbf{l}}, k) = \frac{\widetilde{S}^{\mathrm{d}}(k_{\mathbf{l}}, \mathbf{x_n}, k)}{\widetilde{G}(k_{\mathbf{l}}, \mathbf{x_n}, k)} \tag{2.47}$$

which is a division in the spectral domain, hence the name *spectral division method.* An inverse spatial Fourier transform then yields,

$$Q(\mathbf{l}; k) = \frac{1}{4\pi^2} \iint\limits_{k_{\mathbf{l}} \in \mathbb{R}} \frac{\widetilde{S}^{\mathrm{d}}(k_{\mathbf{l}}, \mathbf{x_n}, k)}{\widetilde{G}(k_{\mathbf{l}}, \mathbf{x_n}, k)} \exp(-i k_{\mathbf{l}} \mathbf{l}) \, d k_{\mathbf{l}}, \tag{2.48}$$

which is the SDM loudspeaker driving function for soundfield reproduction. In order for $\widetilde{Q}(k_{\mathbf{l}}, k)$ and $Q(\mathbf{l}; k)$ to be defined, $\widetilde{G}(k_{\mathbf{l}}, \mathbf{x_n}, k)$ must not manifest any zeros. Due to the division, when $\widetilde{G}(k_{\mathbf{l}}, \mathbf{x_n}, k)$ is small the driving functions are large in comparison.

Since its inception only a decade ago [63] the SDM method has been extended to reproduce focused sources [142] and has been shown to perform as well as, and sometimes better than, WFS [64]. However, the WFS method, and therefore likely the SDM, suffers from detrimental spatial aliasing artefacts above the spatial Nyquist frequency more so than HOA [69] and will be discussed in the next section.

2.2.4 Higher Order Ambisonics (HOA)

Ambisonics was first introduced in 1973 [65], [66], where initial investigations looked at zeroth and first order harmonics, and was further extended to HOA in the 1990's and early 2000's [67]–[70], [72] with harmonic orders of two or greater. The fundamental idea of Ambisonics and HOA is that any soundfield can be described by a combination of soundfields resulting from the harmonic expansions of secondary source signals. As was discussed earlier in section 2.1.5, Ambisonics is an alternative solution to the wave equation in spherical coordinates using the Kirchhoff-Helmholtz integral equation [2]. In practice, a desired soundfield is matched to the expansion of either cylindrical (2D) or spherical (3D) harmonics up to a finite harmonic mode limit [68], [71], [74]. The finiteness of the mode results in predictable truncation errors [73]. The concept of matching the desired soundfield using the harmonic modes is known as the *mode-matching* approach [71].

We start with the notion that any soundfield can be described by a suitably weighted set of orthogonal basis functions. In a spherical coordinate system, i.e. $\mathbf{x} \equiv (r, \theta, \phi)$, the wave equation can be decomposed into an orthogonal set of spherical harmonics. Thus, any three dimensional soundfield can be described as [68], [143]

$$S^{\mathrm{d}}(\mathbf{x}; k) = \sum_{\bar{n}=0}^{\overline{N}} \sum_{\bar{m}=-\bar{n}}^{\bar{n}} E_{\bar{n}\bar{m}}(k) j_{\bar{n}}^{(1)}(k\mathbf{x}) Y_{\bar{n}\bar{m}}(\widehat{\mathbf{x}}), \tag{2.49}$$

where $\bar{m}$ is the mode, $\bar{n}$ is the order, $\overline{N}$ is the highest order, $j_{\nu}^{(1)}(\cdot)$ is the νth order spherical Bessel function of the first kind and $E_{\bar{n}\bar{m}}(k)$ are harmonic coefficients. The unit directional vector, $\widehat{\mathbf{x}}$, has the relation $\mathbf{x} = \|\mathbf{x}\|\widehat{\mathbf{x}}$. The spherical harmonics are given by

$$Y_{\bar{n}\bar{m}}(\widehat{\mathbf{x}}) = \sqrt{\frac{(2\bar{n}+1)}{4\pi} \frac{(\bar{n}-|\bar{m}|)!}{(\bar{n}+|\bar{m}|)!}} \mathcal{P}_{\bar{n}|\bar{m}|}(\cos(\theta)) \exp(i\bar{m}\phi) \tag{2.50}$$

and $\mathcal{P}_{\bar{n}\bar{m}}(\cdot)$ are the associated Legendre functions.

Now consider the actually reproduced soundfield due to a discrete loudspeaker array with a set of loudspeaker positions, $\mathbf{l}_l, l \in [\![L]\!]$, which is constrained to a fixed radius $R_\mathrm{c} = \|\mathbf{l}\|$,

$$S^\mathrm{a}(\mathbf{x};k) = \sum_{l\in[\![L]\!]} Q(\mathbf{l}_l;k)G(\mathbf{x},\mathbf{l}_l;k), \tag{2.51}$$

where the soundfield is found from the superposition of soundfields from individual loudspeakers. We express the driving functions, $Q(\mathbf{l};k)$, in terms of the spherical harmonics

$$Q\Big(\widehat{\boldsymbol{\phi}};k\Big) = \sum_{\bar{n}'=0}^{\infty}\sum_{\bar{m}'=-\bar{n}'}^{\bar{n}'} E'_{\bar{n}'\bar{m}'}(k)Y_{\bar{n}'\bar{m}'}\Big(\widehat{\boldsymbol{\phi}}\Big) \tag{2.52}$$

where $\mathbf{l} = R_\mathrm{c}\widehat{\boldsymbol{\phi}}$ and apply the addition theorem to the three dimensional Greens function [2]

$$G_{3\mathrm{D}}(\mathbf{x},\mathbf{l};k) = ik\sum_{\bar{n}=0}^{\infty}\sum_{\bar{m}=-\bar{n}}^{\bar{n}} j_{\bar{n}}^{(1)}(k\|\mathbf{x}\|)h_{\bar{n}}^{(1)}(kR_\mathrm{c})Y_{\bar{n}\bar{m}}(\widehat{\mathbf{x}})Y^*_{\bar{n}\bar{m}}\Big(\widehat{\boldsymbol{\phi}}\Big) \tag{2.53}$$

where $\{\cdot\}^*$ indicates complex conjugation and $h_\nu^{(1)}(\cdot)$ is the νth order spherical Hankel function of the first kind. The spherical harmonics have a key property in that they are orthogonal

$$\int Y^*_{\bar{n}\bar{m}}(\widehat{\mathbf{x}})Y_{\bar{n}'\bar{m}'}(\widehat{\mathbf{x}})\,d\widehat{\mathbf{x}} = \delta_{\bar{n}\bar{n}'}\delta_{\bar{m}\bar{m}'}, \tag{2.54}$$

where δ is the Kronecker delta function. We apply (2.54) when substituting (2.52) and (2.53) into (2.51) to arrive at

$$S^\mathrm{a}(\mathbf{x};k) = \sum_{\bar{n}=0}^{\overline{N}}\sum_{\bar{m}=-\bar{n}}^{\bar{n}} ikE'_{\bar{n}\bar{m}}(k)h_{\bar{n}}^{(1)}(kR_\mathrm{c})j_{\bar{n}}^{(1)}(k\mathbf{x})Y_{\bar{n}\bar{m}}(\widehat{\mathbf{x}}), \tag{2.55}$$

which is truncated to $\overline{M}$. After equating (2.49) and (2.55) we obtain the desired driving functions,

$$Q\Big(\widehat{\boldsymbol{\phi}};k\Big) = \mathcal{A}\sum_{\bar{n}=0}^{\overline{N}}\sum_{\bar{m}=-\bar{n}}^{\bar{n}} \frac{iE_{\bar{n}\bar{m}}(k)}{kh_{\bar{n}}^{(1)}(kR_\mathrm{c})}Y_{\bar{n}\bar{m}}\Big(\widehat{\boldsymbol{\phi}}\Big), \tag{2.56}$$

which is scaled by an approximation of the surface area of the sphere,

$$\mathcal{A} = 2\pi R^2\left(1-\cos\left(\frac{\Delta\phi_\mathrm{s}}{2}\right)\right), \tag{2.57}$$

where R is the radius of the desired reproduction region and $\Delta\phi_{\mathrm{s}}$ is the spacing between adjacent loudspeakers. The driving functions given in (2.57) can be used to reproduce the actual soundfield with (2.51).

While the Ambisonics approach does assume specific loudspeaker array geometries, such as circular or spherical, it also provides more accurate reproduction within the specified region when compared to WFS [69]. The designation of a specific reproduction region allows the HOA approach to continue to reproduce with high accuracy above the spatial aliasing frequency, however, this only occurs within the so called *sweet spot* which shrinks in size with increasing frequency [135]. It has also been concluded that HOA results in less detrimental spatial aliasing artefacts above the aliasing frequency [69]. Traditionally, HOA is based on the amplitude panning approach, which requires that the loudspeakers be placed far away so that the wave propagation can be approximated as plane waves and is not always practically feasible. The techniques presented in this section truncate the Bessel functions and so do not require the same large distances between the source and the listener, which leads to a more practical implementation for small areas. The Ambisonics driving functions have also been analytically derived specifically for plane wave reproduction [144].

In the next section, we will discuss the concept of personal sound and how the methods presented so far have been extended in the literature to provide individual zones of bright (loud) and quiet audio content. The idea of personal sound is further explored through the perception and privacy of reproduced multiple zone soundfields.

2.3 Personal Sound

The ability to provide personalised sound to individuals in shared environments has been of significant interest to researchers in recent years [24], [25], [31], [36], [80], [86], [88], [91], [93], [99], [101], [113], [145]–[153]. There are many different aspects to personal sound, including, but not limited to: the perceived loudness, annoyance, quality and spatial impression of shared sound; the intelligibility of speech and personal speech privacy; and the numerous approaches used to reproduce personal sound. In this section, we cover the broader concept of personal sound zones, ways in which we can produce personal sound and current literature on human perception and speech privacy.

2.3.1 Concept of Sound Zones

The idea of personal sound is quite broad in that it can be provided by various means. The final goal of a personal sound system is to provide a completely isolated acoustic environment for an individual or small groups of individuals.

One of the easiest methods to accomplish this feat is to use passive absorbers, whether they be specifically tested acoustic fibre panels or merely room walls filled with thermal insulation. While the use of physical passive absorbers seems convenient and is the most common method, there are many drawbacks. For instance, there is the lost ability to freely move about the global space; there is lost ability to communicate to others through the walls; spaces are usually confined; effective passive absorbers can further reduce available space; and regular room walls are ineffective at isolating sounds.

Another popular method that is used to obtain personal sound is through the use of headphones and earphones. While quite effective it is not always feasible to wear them for long periods of time and there can be significantly less spatial impression when using current technology devices, although there is ongoing work to improve this aspect with approaches based on head-related transfer functions (HRTFs) [154]. Headphones and earphones have seen further improvements in areas of personalisation such as with active noise cancelling devices, which are used to cancel unwanted sounds from being heard. A drawback of the use of headphones for personal sound is that they make it more difficult to hold conversations as they passively block most sounds for the person wearing them. The potential for a personal sound system to provide the flexibility of physical movement in a space, visibility across that space and the ability to hold comfortable conversations with others in close vicinity, is the interest that researchers have come to acquire recently.

The process of creating multiple spatially separated zones of sound for individuals has become known as *multizone soundfield reproduction*. The physical basis for the idea is that of constructive and de-constructive sound waves. An array of loudspeakers is used, not unlike discussed previously throughout this chapter, to synchronously generate many wave fields in ways that construct in a single area to create a loud zone of sound that is commonly termed the *bright zone* and to de-construct elsewhere in the space to create what is known as a *quiet zone*. This concept does not require the use of physical barriers or earphones/headphones and allows persons to move freely about the entire shared environment. The artificial induction of sound within a particular environment can lead to numerous undesired effects. The perceptual quality and spatial localisation capability can be influenced when using such systems as they are heavily restricted in their freedom to reproduce sound, i.e. they are restricted to reproduction within spatial areas. The spatial restrictions placed upon the systems can also lead to imperfect reproductions where sound may leak audibly from one area to another, which can result in leaked information, thus developing a privacy issue.

In the next several sections, we will discuss the mathematical background of several methods that have been developed to reproduce multizone soundfields for providing personal sound zones. A discussion on the human perception of sound and how it relates to personal sound follows with background on intelligibility and privacy at the end.

2.3.2 Multiple Zone Reproduction Techniques

The ability to reproduce more than a single zone of sound in a wider region of interest is possible using various techniques. A brief overview of the prevailing approaches is given for those methods that aim to reproduce a bright zone and a quiet zone (also referred to as a dark zone). To reproduce two bright zones, with different audio programmes, the methods discussed are often simply designed to reproduce the bright zones superimposed on other quiet zones. The only technique here that does not aim to specifically quieten the dark zone is the oldest method, the beam-forming approach, which we will discuss first.

Beam-forming

The oldest technique that was specifically designed to address the personal sound zone problem was partially based on a beam-forming approach in 1997 [24], even though the idea of beam-forming had existed earlier in antenna array design [76]. Since then the beam-forming approach has seen various improvements and evaluation studies. Beam-forming methods have been compared to the acoustic contrast control (ACC) method [95] and further extended to super-directive beam-forming techniques [77], beam-width control [155], MIMO optimisation [78] and unique parametric loudspeaker arrays [156], [157].

Two classical types of beam-forming are *delay-and-sum* beam-forming and *filter-and-sum* beam-forming. For a discrete array, the delay-and-sum beam-former can be expressed simply as [1]

$$S(\mathbf{x};k) = \sum_{l\in[\![L]\!]} \underbrace{W_l \exp\Big(-i\Big(\omega(\tau_0+\tau_l) + \boldsymbol{k}^{\mathsf{T}}(\mathbf{b}-\mathbf{l}_l)\Big)\Big)}_{Q(\mathbf{l}_l;k)} G(\mathbf{x},\mathbf{l}_l;k), \tag{2.58}$$

where W_l is a real valued weight that scales the loudspeaker response, $\boldsymbol{k}$ is the desired wavenumber vector, $\mathbf{b}$ is the desired bright zone position, τ_l is the time delay and τ_0 causal time delay. The maximum response is when τ_l compensates for the propagation delay, $\boldsymbol{k}^{\mathsf{T}}(\mathbf{b}-\mathbf{l}_l)$. For a derivation of the filter-and-sum beam-former the reader is referred to [1].

A common approach to beam-forming when using microphone arrays is to point a null in the direction of the interferer to improve the signal to noise ratio. There are also formulations that aim to optimise statistical attributes, such as variance, whilst constraining the optimisation for other some other criteria, such as a distortionless response. An alternative beam-former approach for personal sound zones was introduced in 2002 [85] which optimises the acoustical brightness in specified areas. The second of the two optimisation approaches presented in [85] is described next.

Acoustic Contrast Control (ACC)

One of the earliest techniques, which is still compared to today, is the ACC approach to multizone soundfield reproduction and was first published in 2002 [85]. The method aims to maximise the contrast in energy ratios between the bright zone and the quiet zone. For the derivation of the ACC method we will use vector notation. We express the soundfield in vector notation as

$$\mathbf{s}(k) = \mathbf{T}(k)\mathbf{q}(k) \tag{2.59}$$

where $\mathbf{s}(k)$ is soundfield column vector for all points in the reproduction region, $\mathbf{T}(k)$ is a matrix of transfer functions between the control points in the reproduction region and the loudspeakers, and $\mathbf{q}(k)$ is a column vector of the driving functions for each loudspeaker. We call the complete reproduction region $\mathbb{D}$, the region containing all the points in the bright zone we call $\mathbb{D}_\mathrm{b}$ and the region containing all the points in the quiet zone we call $\mathbb{D}_\mathrm{q}$. The size of the regions are,

$$\mathfrak{d}_\mathrm{b} \triangleq \int_{\mathbb{D}_\mathrm{b}} 1\, d\mathbf{x}, \quad \forall \mathbf{x} \in \mathbb{D}_\mathrm{b}, \tag{2.60}$$

$$\mathfrak{d}_\mathrm{q} \triangleq \int_{\mathbb{D}_\mathrm{q}} 1\, d\mathbf{x}, \quad \forall \mathbf{x} \in \mathbb{D}_\mathrm{q}. \tag{2.61}$$

From this we define new vectors and matrices as

$$\mathbf{s}_{\mathfrak{d}_\mathrm{b} \times 1}(k) = \mathbf{T}_{\mathfrak{d}_\mathrm{b} \times L}(k)\mathbf{q}_{L \times 1}(k), \tag{2.62}$$

$$\mathbf{s}_{\mathfrak{d}_\mathrm{q} \times 1}(k) = \mathbf{T}_{\mathfrak{d}_\mathrm{q} \times L}(k)\mathbf{q}_{L \times 1}(k). \tag{2.63}$$

The cost function that is to be maximised for the ACC method is then given by

$$\mathcal{J}_{\mathrm{ACC}} = \frac{\mathbf{s}^{H}_{\mathfrak{d}_{\mathrm{b}}\times 1}\mathbf{s}_{\mathfrak{d}_{\mathrm{b}}\times 1}}{\mathbf{s}^{H}_{\mathfrak{d}_{\mathrm{b}}\times 1}\mathbf{s}_{\mathfrak{d}_{\mathrm{b}}\times 1} + \mathbf{s}^{H}_{\mathfrak{d}_{\mathrm{q}}\times 1}\mathbf{s}_{\mathfrak{d}_{\mathrm{q}}\times 1}} \tag{2.64}$$

$$= \frac{\mathbf{q}^{H}_{L\times 1}\mathbf{T}^{H}_{\mathfrak{d}_{\mathrm{b}}\times L}\mathbf{T}_{\mathfrak{d}_{\mathrm{b}}\times L}\mathbf{q}_{L\times 1}}{\mathbf{q}^{H}_{L\times 1}\Big(\mathbf{T}^{H}_{\mathfrak{d}_{\mathrm{b}}\times L}\mathbf{T}_{\mathfrak{d}_{\mathrm{b}}\times L} + \mathbf{T}^{H}_{\mathfrak{d}_{\mathrm{q}}\times L}\mathbf{T}_{\mathfrak{d}_{\mathrm{q}}\times L}\Big)\mathbf{q}_{L\times 1}}, \tag{2.65}$$

which is a ratio of the energy in the bright zone to the total energy in both zones.

The driving function vector that maximises $\mathcal{J}_{\mathrm{ACC}}$ is then given by

$$\mathcal{J}_{\mathrm{ACC}}\mathbf{q}_{L\times 1} = \Big(\mathbf{T}^{H}_{\mathfrak{d}_{\mathrm{b}}\times L}\mathbf{T}_{\mathfrak{d}_{\mathrm{b}}\times L} + \mathbf{T}^{H}_{\mathfrak{d}_{\mathrm{q}}\times L}\mathbf{T}_{\mathfrak{d}_{\mathrm{q}}\times L}\Big)^{-1}\Big(\mathbf{T}^{H}_{\mathfrak{d}_{\mathrm{b}}\times L}\mathbf{T}_{\mathfrak{d}_{\mathrm{b}}\times L}\Big)\mathbf{q}_{L\times 1}, \tag{2.66}$$

which is an eigenvalue problem, where the maximised cost function, $\mathcal{J}_{\mathrm{ACC}}$, is given by the eigenvector that corresponds to the largest eigenvalue.

Pressure Matching (PM)

The pressure matching approach, which was presented by Poletti in 2008 [158], is straightforward to compute if we recall from earlier, $\mathbf{s}^{\mathrm{d}}$ and (2.39). We can then simply redefine $\mathbf{s}^{\mathrm{d}}$ so that the measurements from each zone are weighted separately in the LSO then computing the LSO will result in multizone soundfield reproduction loudspeaker driving functions. We redefine the measurement vector as

$$\mathbf{s}^{\mathrm{d}} = \begin{bmatrix} \mathbf{s}^{\mathrm{d}}_{\mathbb{D}_{\mathrm{b}}} \\ \mathbf{s}^{\mathrm{d}}_{\mathbb{D}_{\mathrm{q}}} \end{bmatrix}, \tag{2.67}$$

where $\mathbf{s}^{\mathrm{d}}_{\mathbb{D}_{\mathrm{b}}}$ is the desired soundfield in the bright region and $\mathbf{s}^{\mathrm{d}}_{\mathbb{D}_{\mathrm{q}}}$ is the desired soundfield in the quiet region. When using (2.39) for the multizone scenario the matrix of transfer functions also needs to be redefined as,

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{\mathbb{D}_{\mathrm{b}}} \\ \mathbf{T}_{\mathbb{D}_{\mathrm{q}}} \end{bmatrix}. \tag{2.68}$$

It is worth noting that both $\mathbf{s}^{\mathrm{d}}_{\mathbb{D}_\mathrm{b}}$ and $\mathbf{s}^{\mathrm{d}}_{\mathbb{D}_\mathrm{q}}$ can be specified as arbitrary soundfields, including quiet zones. The new loudspeaker driving functions, $\widehat{\mathbf{q}}(k)$, are given by (2.39) using (2.67) and (2.68).

Planarity Control (PC)

The design of the PC method was motivated by the need for a combination of soundfield synthesis methods and energy control methods and addresses the issues the ACC method exhibits with the control of the velocity component of the soundfield [95]–[100]. The ACC method we discussed above considers a maximisation of acoustic contrast but does not consider the directional components of the soundfield. The maximisation that is performed in the ACC method can create unpredictable distributions in pressure, on the other hand, soundfield synthesis methods providing smoother pressure distributions but at the cost of lower acoustic contrast between zones [96]. The cost function formulated for the PC method optimises the attenuation into the quiet zone and the reproduction of the plane wave into the bright zone [96], [98], [101]. The PC method does this with an additional constraint defined by a steering angle matrix that weights the system to consider the directional components of the soundfield. The metric known as *planarity* is defined as the ratio between the energy contribution from the largest plane wave component direction and the total energy of the plane wave components for a given soundfield.

We start with the plane wave energy distribution, $\mathcal{E}_\rho$, of the soundfield for each direction, ρ. The planarity of the bright zone soundfield is then given by

$$\mathcal{P}_{\mathbb{D}_\mathrm{b}} = \frac{\sum_\rho \mathcal{E}_\rho \widehat{\mathbf{u}}_\rho \cdot \widehat{\mathbf{u}}_{\rho_\mathrm{max}}}{\sum_\rho \mathcal{E}_\rho}, \tag{2.69}$$

where $\widehat{\mathbf{u}}$ is a unit vector in the direction indicated by its subscript and $\rho_\mathrm{max} = \arg\max_\rho \mathcal{E}_\rho$. The energy distribution in vector form is

$$\boldsymbol{\mathcal{E}} = \begin{bmatrix}\mathcal{E}_1 & \mathcal{E}_2 & \cdots & \mathcal{E}_J\end{bmatrix}^\mathsf{T} \tag{2.70}$$

and by using (2.62) we arrive at the relationship

$$\boldsymbol{\mathcal{E}} = \frac{1}{2}\|\mathbf{H}_{\mathbb{D}_{\mathrm{b}}}\mathbf{s}_{\mathfrak{d}_{\mathrm{b}}\times 1}(k)\| \tag{2.71}$$

where $\mathbf{H}_{\mathbb{D}_{\mathrm{b}}}$ is a steering matrix of size $J \times \mathfrak{d}_{\mathrm{b}}$ used in the optimisation. A weighting term is used to focus energy in specified directions and is expressed as

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_1 & 0 & 0 & 0 \\ 0 & \gamma_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \gamma_J \end{bmatrix}, \tag{2.72}$$

where $0 \leq \gamma_\rho \leq 1$.

The planarity control optimisation cost function is then defined as an extension of the ACC method from (2.66) with

$$\mathcal{J}_{\mathrm{PC}}\mathbf{q}_{L\times 1} = -\Big(\mathbf{T}^H_{\mathfrak{d}_{\mathrm{b}}\times L}\mathbf{H}^H_{\mathbb{D}_{\mathrm{b}}}\mathbf{\Gamma}\mathbf{H}_{\mathbb{D}_{\mathrm{b}}}\mathbf{T}_{\mathfrak{d}_{\mathrm{b}}\times L}\Big)^{-1}\Big(\mathbf{T}^H_{\mathfrak{d}_{\mathrm{q}}\times L}\mathbf{T}_{\mathfrak{d}_{\mathrm{q}}\times L}\mathbf{q}_{L\times 1} + \lambda_{\mathrm{cond.}}\mathbf{q}_{L\times 1}\Big), \tag{2.73}$$

where $\lambda_{\mathrm{cond.}}$ is a Lagrange multiplier that is initialised based on the matrix condition number. The optimised driving functions used to reproduce the multizone soundfield are found with (2.73). While the PC method has significant advantages over the ACC method, in that it controls the direction of propagating sound waves, it is only defined to control plane wave fronts. Nonetheless, the PC method offers a good trade-off between the separation in sound pressure levels between zones and the error of the plane wave shape in the bright zone.

Harmonic Expansion (HE)

The Harmonic Expansions (HE) based approaches to soundfield reproduction that we discussed in section 2.2.4 have been successfully extended to the multizone case [102]–[108]. The extension to the multizone case is based on spatially filtering and translating the soundfield coefficients in a global coordinate system. The coefficient translation theorem spatially relocates the soundfield of a particular zone, defined by its coefficients, to an alternative global coordinate location [103]. To avoid unwanted leakage from one zone to another additional angular windowing can be applied to the coefficients after translation [103]. As an alternative to the angular windowing, spatial band stop filters that use the higher order spatial harmonics of a zone to cancel undesired effects of its lower order harmonics on other zones can be applied [104].

For simplicity, let us consider the two dimensional equivalent of (2.49) as [2]

$$S^{\mathrm{d}}(r_o, \theta_o; k) = \sum_{\bar{m}=-\overline{M}}^{\overline{M}} E_{\bar{m}}(k) \underbrace{\mathcal{J}_{\bar{n}}^{(1)}(kr_o)\exp(i\bar{m}\theta_o)}_{V(r_o,\theta_o;k)}, \tag{2.74}$$

where $r_o = \|\mathbf{x}\|$, $\theta_o = \cos^{-1}(\widehat{\mathbf{x}} \cdot \widehat{\mathbf{u}}_{\mathrm{o}})$, the global mode limit is $\overline{M}$, $\mathcal{J}_{\nu}^{(1)}(\cdot)$ is a νth order cylindrical Bessel function of the first kind and $\widehat{\mathbf{u}}_{\mathrm{o}}$ is the origin unit vector. In this work the global mode limit is given by $\overline{M} = \lceil kR \rceil$.

Let us consider $E_{\overline{m}}(k)$ to be the coefficients in global coordinates. We then denote the coefficients for the bright zone as $E_{\overline{m}}^{(\mathrm{b})}(k)$ and the coefficients for the quiet zone as $E_{\overline{m}}^{(\mathrm{q})}(k)$. Expressing the coefficient in vector form we have

$$\mathbf{E}(k) = \left[E_{-\overline{M}}(k) \quad \cdots \quad E_{\overline{M}}(k) \right]^{\mathsf{T}} \tag{2.75}$$

$$\mathbf{E}^{(\mathrm{z})}(k) = \left[E_{-\overline{M}_b}^{(\mathrm{b})}(k) \quad \cdots \quad E_{\overline{M}_b}^{(\mathrm{b})}(k) \quad E_{-\overline{M}_q}^{(\mathrm{q})}(k) \quad \cdots \quad E_{\overline{M}_q}^{(\mathrm{q})}(k) \right]^{\mathsf{T}}, \tag{2.76}$$

where $\mathbf{E}(k)$ is the vector of global coefficients and $\mathbf{E}^{(\mathrm{z})}(k)$ is the vector of concatenated zone coefficients. The mode limit for the bright zone, $\overline{M}_b = \lceil kr_{\mathrm{b}} \rceil$, is obtained using the radius of the bright zone, r_{b}, and the mode limit for the quiet zone, $\overline{M}_q = \lceil kr_{\mathrm{q}} \rceil$, is obtained using the radius of the quiet zone, r_{q}.

Using (2.75) and (2.76) we can write the system of simultaneous equations,

$$\mathbf{E}^{(\mathrm{z})}(k) = \mathbf{V}(k)\mathbf{E}(k), \tag{2.77}$$

where

$$\mathbf{V}(k) = \begin{bmatrix} V^{(\mathrm{b})}_{-\overline{M}_b+\overline{M}} & \cdots & V^{(\mathrm{b})}_{-\overline{M}_b-\overline{M}} \\ \vdots & \ddots & \vdots \\ V^{(\mathrm{b})}_{\overline{M}_b+\overline{M}} & \cdots & V^{(\mathrm{b})}_{\overline{M}_b-\overline{M}} \\ V^{(\mathrm{q})}_{-\overline{M}_q+\overline{M}} & \cdots & V^{(\mathrm{q})}_{-\overline{M}_q-\overline{M}} \\ \vdots & \ddots & \vdots \\ V^{(\mathrm{q})}_{\overline{M}_q+\overline{M}} & \cdots & V^{(\mathrm{q})}_{\overline{M}_q-\overline{M}} \end{bmatrix}, \tag{2.78}$$

the mode limit for the bright zone is $\overline{M}_b$, the quiet zone mode limit is $\overline{M}_q$ and $V^{(\mathrm{z})}_{\overline{M}} \triangleq V(r_z,\theta_z;k)$ for a mode limit of $\overline{M}$ from (2.74). The global coefficients can then be solved with

$$\mathbf{E}(k) = \mathbf{V}^{\dagger}(k)\mathbf{E}^{^{(\mathrm{z})}}(k). \tag{2.79}$$

In the cylindrical harmonic expansion method the loudspeaker signals can be found using the following expansion

$$Q(\mathbf{l};\phi_l) = \sum_{\bar{m}=-\overline{M}}^{\overline{M}} \frac{2}{i\pi\mathcal{H}^{(1)}_{\bar{m}}(kR_{\mathrm{c}})} E_{\bar{m}}(k)\exp(i\bar{m}\phi_l)\Delta\phi_{\mathrm{s}}. \tag{2.80}$$

This particular method of harmonic expansion is used through this thesis with the expansion coefficients computed using the method described below in the next section. For a derivation of multizone soundfield reproduction using the spherical harmonic expansion method the reader is referred to [159].

The HE based methods compute the coefficients to reproduce the soundfield using both the pressure and vector components of the soundfield which results in good reproduction quality and high acoustic contrast. The drawback of the method is that reproduction errors occur from the truncation of the modes to the given mode limit. However, the error induced from truncation is well understood and analytical derivations of the error have been published [73], [106].

Orthogonal Basis Expansion (OBE)

The OBE approach to multizone soundfield reproduction is a further improvement on existing state-of-the-art techniques and was pioneered by Jin et al. in 2013 [109]. The OBE method controls both the pressure and velocity of the soundfield over an entire region similar to the HE coefficient translation method and the PC method. The benefit of OBE over the PC method is that it is defined for any arbitrary soundfield and not constrained to plane waves. The OBE method further improves on the original HE based methods by relieving the constraint on zones of quiet. The relief is given with zone based weights which are used in a modified Gram-Schmidt process. Another improvement of the OBE approach over the HE based methods is its potential to make use of sparse basis functions [112], [114].

The concept of optimising multizone soundfield reproductions with relative weights for each zone was introduced in [109]. This concept was further extended to the HE based multizone approach in 2014 where the prioritised control of regions was introduced [107]. The OBE approach in particular has been shown as a viable method of reproduction in real-world environments [112], [113] and has been extended to reverberant rooms using sparse methods [110], [112], [114], which makes it a good choice for practical applications. As we will see in the next chapter, extended frequency dependent zone weights are useful for perceptually controlling multizone soundfield reproductions.

We begin with the notion that any arbitrary soundfield function, $S(\mathbf{x};k)$, can be expressed as the summation of a weighted series of basis functions,

$$S^{\mathrm{a}}(\mathbf{x};k) = \sum_{j\in[\![J]\!]} E_j(k) F_j(\mathbf{x};k), \tag{2.81}$$

where $\{F_j\}_{j\in[\![J]\!]}$, the expansion coefficients are $E_j(k)$ and J is the number of basis functions. The white square brackets are defined to be a compact notation for indices with $[\![A]\!] \triangleq \{x : x \in \mathbb{N}_0, x < A\}$. The goal is to find the expansion coefficients that minimise the difference between the actual soundfield and the desired soundfield,

$$\min_{E_j} \left\| \sum_j E_j(k) F_j(\mathbf{x}; k) - S^{\mathrm{d}}(\mathbf{x}; k) \right\|^2, \tag{2.82}$$

which can be done by solving the weighted inner product

$$E_j(k) = \left\langle S^{\mathrm{d}}(\mathbf{x}; k), F_j(\mathbf{x}; k) \right\rangle_w = \int_{\mathbb{D}} S^{\mathrm{d}}(\mathbf{x}; k) {F_j}^*(\mathbf{x}; k) w(\mathbf{x})\, d\mathbf{x}, \tag{2.83}$$

where, for any X, $\|X\|^2_{(\mathrm{w})} = \langle X, X \rangle_w$.

The zone weighting function, $w(\mathbf{x})$, is designed with a weight for the relative importance of the reproduction at each point in space. The zone weighting function can be defined as

$$w(\mathbf{x}) = \begin{cases} w_{\mathrm{b}}, & \mathbf{x} \in \mathbb{D}_{\mathrm{b}} \\ w_{\mathrm{q}}, & \mathbf{x} \in \mathbb{D}_{\mathrm{q}} \\ w_{\mathrm{u}}, & \mathbf{x} \in \mathbb{D}_{\mathrm{u}} \end{cases} \tag{2.84}$$

Next, we wish to find the set of orthogonal basis functions, $\{F_j\}_{j \in [\![J]\!]}$. We can do this by implementing an orthogonalisation on a set of planewaves that arrive from a set of discrete angles,

$$P_h(\mathbf{x}; k) = \exp(ik\mathbf{x} \cdot \boldsymbol{\rho}_h), \tag{2.85}$$

where $\boldsymbol{\rho}_h \equiv (1, \rho_h)$, $\rho_h = (h-1)\Delta\rho$ and $\Delta\rho = 2\pi/J$. Note that these plane waves can be combined to describe any soundfield and the functions are not limited to plane waves.

A modified Gram-Schmidt process is used to give the orthogonalised basis functions and which also contains the relative zone weighting function. The modified Gram-Schmidt process is

$$F_j(\mathbf{x}; k) = P_j(\mathbf{x}; k) - \sum_{p \in [\![j-1]\!]} \frac{\langle P_j(\mathbf{x}; k), F_p(\mathbf{x}; k) \rangle_w}{\langle F_p(\mathbf{x}; k), F_p(\mathbf{x}; k) \rangle_w} F_p(\mathbf{x}; k) \tag{2.86}$$

The Gram-Schmidt process then results in

$$F_j(\mathbf{x};k) = \sum_{h\in[\![J]\!]} \mathbf{R}_{hj} P_h(\mathbf{x};k), \tag{2.87}$$

where $\mathbf{R}_{hj}$ is the (h,j)th element of the lower triangular matrix, $\mathbf{R}$. Substituting (2.87) in (2.81), yields

$$S^{\mathrm{a}}(\mathbf{x};k) = \sum_{h\in[\![J]\!]} \mathcal{W}_h P_h(\mathbf{x};k), \tag{2.88}$$

where $\mathcal{W}_h = \sum_{j\in[\![J]\!]} E_j \mathbf{R}_{hj}$ are the plane-wave coefficients used to construct the actual soundfield. Using the HE described above, we can replace the expansion coefficients in (2.80) with

$$E_{\bar{m}}(k) = \sum_{h} \mathcal{W}_h i^{\bar{m}} \exp(-i\bar{m}\rho_h). \tag{2.89}$$

Finally, we can obtain the driving functions using (2.80) which are used to reproduce the multizone soundfield.

As briefly mentioned above, later in chapter 3 we will show how the multizone spatial weighting can be used to reduce the error in the bright zone by considering human perception of sound in the quiet zone. The human perception of sound is well understood and has been studied for many decades, we will discuss relevant aspects of this field in the following section.

2.3.3 Human Perception

There are many aspects to soundfields, and sound in general, that extend beyond the limit of human hearing, such as the dynamic range, frequency, noise level and ability to localise sources of sound. In many cases it is not necessary to perfectly reconstruct the physical characteristics of audio scenes as those who listen to the result cannot distinguish between the digital reconstruction and the original. For example, some of the mostly widely used digital compression techniques, such as MPEG Audio Layer III (MP3) [160], [161] and JPEG [162], [163], use human perceptual models to increase compression ratios so that changes are minimally perceivable. In this section, we review background on key components of perceptual acoustic models for humans, how they relate to multizone soundfield reproduction and link forward to relevant chapters in the thesis.

The Hearing Threshold and Equal Loudness Level

The human auditory system is exceptionally sensitive to pressure fluctuations in fluid mediums, however, there is a limit at which humans have difficulty perceiving sounds, called the *absolute hearing threshold* [1], [164], [165]. The absolute hearing threshold is where a just-noticeable-difference (JND) in the level of a test tone in a quiet environment is heard. The threshold in quiet is frequency dependant and has been well established with functions that provide a good approximation for the limit at different frequencies [164], [166]. This threshold of human hearing is a good definition for the level that one might call “quiet”, as opposed to complete silence defined by zero energy of a soundfield. We will see in chapter 3 the relevance for this threshold in multizone soundfield reproductions.

A further extension of the hearing threshold is to equal loudness levels, where each frequency is perceived to be of the same apparent sound pressure level, called loudness [166]–[168]. The equal loudness curves are useful for when one wishes to reproduce sounds that are heard equally across all frequency bands. However, the functions that define loudness for tones and for wideband noise are not identical [165]. The reference level for equal loudness is often the threshold of hearing for normal persons at 0 phon, where a phon is a measure of loudness for tones. Mapping sound pressure levels to loudness can be done using the mapping functions provided in [166]. There are several options for expressing wideband sounds in the loudness scale which include: using the unweighted root-mean-square (RMS) level over the audio frequency range; using an A-weighted signal level; or using loudness defined in sones, which is a more accurate perceived loudness scale for sounds with more than one frequency component [168].

Psychoacoustic Masking and Spreading Functions

The human auditory system is very good at hearing differences in sound levels and hearing a wide range of frequencies. However, the detection of sound components that are close to neighbouring ones can be a difficult task for human hearing. This difficulty exists in both the frequency domain and time domain. Sounds at particular frequencies mask the discernible presence of nearby frequency components and this characteristic is generally called *psychoacoustic frequency masking*. Whereas, the masking of nearby frequency components in time is known as *psychoacoustic temporal masking*.

Understanding how auditory masking behaves allows psychoacoustic models to be developed that can provide accurate representation of the human auditory system. For instance, the level of masking on neighbouring frequencies tapers off the further they are from the reference frequency component. This spread in masking level has been modelled extensively over the last century and the psychoacoustic models that describe it are known as *spreading functions* [164], [169]. The knowledge of the auditory masking behaviour and the psychoacoustic models that followed have provided benefits for many real-world applications [160], [161].

One of the most popular spreading functions that is still used in the ISO/IEC MPEG Psychoacoustic Model 2 is given by [164], [170]

$$\begin{aligned} 10\log_{10} SF(dz) = 15.8111389 &+ 7.5(1.05dz + 0.474) \\ &- 17.5\sqrt{1 + (1.05dz + 0.474)^2} \\ &+ 8\min\Big(0, \big((1.05dz - 0.5)^2 - 2(1.05dz - 0.5)\big)\Big), \qquad (2.90) \end{aligned}$$

where dz is the difference between the maskee and the masker frequencies in the Bark scale. (2.90) is based on the Schroeder spreading function [170], [171] obtained from Zwicker's data [172]. This spreading function has the advantage that, when compared to other spreading functions, it is not dependent on the SPL of the masker, which results in faster computation. In chapter 3, we will show how this spreading function can be used to reduce spatial error in multizone soundfield reproductions. There are various other spreading functions that have been established over the last several decades; the reader is referred to [164] for more details.

Speech Quality

In any case, applications of audio systems are sought to be of high quality for the end user. The calibre of speech processing systems, in particular, rely on speech quality assessments, which are a result of human speech perception and a process of assessment [173]. For this reason, speech quality only exists due to the subjects whom provide the assessment and these measures of quality are called *subjective* speech quality measures. It is not always feasible to perform subjective quality assessments as they can be expensive and time consuming. There are, however, algorithms that are designed to approximate the results that would have been obtained using a subjective quality assessment, which are called *objective* speech quality measures. There are two main types of quality assessments; those that require a reference signal; and those that do not require a reference signal. The former is often called an *intrusive* measure when referring to objective tests and the latter is called a *non-intrusive* measure.

In general, there are many aspects to speech quality that may affect the results of an assessment and/or the perceived quality of a system to an end user. Some of these aspects are loudness, listening effort, naturalness and intelligibility. There are also numerous factors that degrade speech quality, such as reverberation (echoes), crosstalk and background noise. In section 2.3.4, we further discuss speech intelligibility and its relationship to speech privacy.

There are several types of intrusive objective measures that can be categorised into two different classes; those which are based on spectral comparisons; and those which are based on psychoacoustic models. The simplest of these measures are the signal-to-noise ratio (SNR) measure [174] and the segmental SNR (SSNR) [175]. Spectral distance measures, such as the Itakura-Saito (IS) distance and cepstral distance, can also be used as quality measures [176]. An advantage of spectral based methods is that better alignment of speech signals allows for distances to be easily calculated.

The more advanced psychoacoustic based models for speech quality measures have been widely adopted in recent years. One of the most popular methods is the perceptual evaluation of speech quality (PESQ) measure [177], [178], which has since been extended to wideband speech [179]. The PESQ measure consists of several speech processing components, including a psychoacoustic model, and was originally designed to assess a range of network-based speech degradations, such as from codecs and packet loss. It has since been used for a wide variety of speech quality assessments. In 2011, the perceptual objective listening quality assessment (POLQA) was published, which addresses some of the shortcomings of the PESQ method, such as the perceived quality at different presentation levels [180]. The performance of POLQA in a wide range of applications is still an area of ongoing research.

Later, in chapter 4, we provide methods for controlling aspects of speech quality in multizone soundfield reproductions based on intrusive objective speech quality measures. For an overview of subjective tests and non-intrusive speech quality measures the reader is referred to [173].

Spatial Sound Perception

Humans are highly capable of detecting nuances in auditory cues such as the interaural time difference (ITD), interaural level difference (ILD) and pinnae based spectral changes. The human brain is capable of processing these cues, which are received at both ears, to extract meaningful information from otherwise noisy signals. The cognitive processes that extract the information perform by suppressing perceived reverberation, localising sound sources and suppressing unwanted sound sources [168].

Binaural noise suppression is the process of separating the desired signal from the undesired parts by using the perceptual mechanism of source localisation. The binaural suppression of noise is also termed *binaural release from masking* or *binaural release from masking* and it is a perceptual effect that can allow humans to naturally improve the intelligibility of speech in noisy environments. However, this effect is only possible when sound and noise sources arrive from different spatial directions and it is less effective when there are many sound sources contributing to a more diffuse field, a phenomenon commonly known as the *cocktail party problem* [4], [181]. We cover later, in chapter 4, applications where it may be desired to maintain, or induce, masking effects (e.g. to improve speech privacy) and derive methods to collocate speech and noise sources, which would inherently hinder binaural release from masking.

The area of research involved with the perception of modern spatial audio reproductions, particularly multizone approaches, is still a topic of ongoing research. There have been studies looking to base soundfield reproductions on perceptual models [182], [183] and numerous others have investigated the perceptual effects of spatial audio reproductions [100], [145], [184]–[187]. The spatial aspects of intelligibility have also been investigated for improving teleconferencing applications using WFS [34].

2.3.4 Acoustic Privacy

The privacy of information in general is often sought after, whether it is textual, visual or acoustic. The de facto standard for acoustic privacy is passive isolation, although, there has been recent popularity in sound masking systems, which produce noisy soundfields across shared spaces in order to mask or hide other sound. Speech is the most common method for humans to transfer information from one person to another and the amount of information conveyed is typically gauged with a measure of intelligibility. Speech privacy can be thought of as the special case of reducing the information imparted to a third-party listener, which our discussion will lead to next.

Speech Intelligibility

There are many approaches to measuring the intelligibility of speech and it is often a case for it to be maximised in noisy environments, so as to improve communication between a talker and a listener [188]. The mutual information between a received message and the original message is a good basis for a measure of intelligibility. In a similar nature to speech quality, there are several categories of technical approaches to measuring speech intelligibility [188]: those operating at the level of individual words [189], [190]; measures based on a system of auditory models [191], [192]; and those operating on short-term spectra [193]–[199].

A popular intelligibility metric known as the short-time objective intelligibility measure (STOI) [199] has been shown as a good measure for time-frequency weighted noisy speech. The loudspeaker driving functions (filters), which have been described earlier in this chapter, are generally derived as time-frequency weights, especially so for applications of active control or temporal masking. For this reason, the STOI measure is suitable for use in many soundfield reproduction scenarios. The STOI algorithm has recently been further enhanced to the extended STOI (ESTOI) algorithm [200], which is suitable for temporally modulated noise sources. A recent study [201] has evaluated 12 existing monaural intrusive instrumental intelligibility metrics showing that STOI and ESTOI perform best for time-frequency weighted signals like those commonly used for soundfield reproduction. The evaluation study also shows that a recent method called the speech intelligibility in bits (SIIB) by Van Kuyk et al. [202] has the highest overall performance. The SIIB is an information theoretic measure that is based on the mutual information shared between the original and degraded speech signal.

The enhancement of speech intelligibility in communication channels has long been of interest [188], [203], [204], however, more recently there has been work on enhancement in the spatial domain [34], [205]. Crespo et al. show that signals leaking from one zone to another (crosstalk) in multiple zone scenarios, which can cause degradations in quality and intelligibility, may benefit from optimisation frameworks designed to model noise, reverberation and crosstalk to enhance intelligibility [205]. The problem of crosstalk between zones is also closely related to speech privacy and is treated in-depth later in chapter 4.

Speech Privacy

As briefly mentioned above, the leakage of speech between spatial regions can lead to two possibilities: that the speech is mixed through a crosstalk process leading to less intelligible speech; or that the speech leaks to an area that was intended to contain signals that happen to not degrade the leaked speech. In the latter scenario, there is potential for information carried by the speech to be heard by listeners whom it was not intended for, resulting in a speech privacy issue.

There has been considerable work on speech privacy in open plan and closed room spaces [206]–[209] and several speech privacy standards have been published [210], [211]. The two main metrics for speech privacy that are used in ASTM E1130 and ASTM E2638 are the articulation index (AI) and the speech privacy class (SPC). It has been shown by Gover et al. that the SPC provides more accurate results than AI for high privacy situations [209]. However, the SPC is based on the signal-to-noise ratio (SNR) and does not consider various other aspects of speech intelligibility as were discussed above. Recent work has proposed the control of the speech transmission index (STI) for speech privacy enhancement in simulated conditions [212].

While speech privacy has been considered for large spaces, the mathematical basis for most current methods fails to specifically address speech privacy in the spatial domain. In chapter 4, we derive spatial field metrics for speech privacy, accompanied by speech quality and privacy control methods. Techniques for the active control of speech transmission over open spaces and the suppression of reflected speech in closed rooms is proposed in chapter 5.

2.4 Summary

In this chapter, we have covered background theory of fundamental acoustic wave propagation and it's relationship to soundfield reproduction methods. Mathematical descriptions of state-of-the-art soundfield reproduction techniques for single zone and multiple zone systems have been derived to help understand the subtleties of each approach. We have derived expressions for amplitude panning, least squares optimisation, wave field synthesis and higher order Ambisonics methods. We then discussed the concept of personal sound and the reproduction of sound zones. After the discussion of sound zones we derived the expressions for beam-forming, acoustic contrast control, pressure matching, planarity control, harmonic expansion and orthogonal basis expansion techniques of multizone soundfield reproduction. We have shown links between some of the multizone soundfield reproduction techniques and particular single zone approaches. A discussion and review of human perception and its connection to personal sound followed with links to sound masking and speech quality. Finally, we provided a view on acoustic privacy with an emphasis on speech intelligibility, its link to information theory and related studies examining performance in the spatial domain.

Throughout this chapter, we have linked forward to later chapters and discussed the relationship between current literature and the work provided through the thesis. In the next chapter, we will cover the use of psychoacoustic frequency masking and spreading functions for reduced error in multizone soundfield reproductions and describe an approach for efficiently computing the zone-based weights required to implement the proposed psychoacoustic methods. In the chapters that follow, we will then provide sophisticated techniques for controlling several aspects of multizone soundfield reproductions for personal sound, such as the control of quality, intelligibility, zone leakage, object reflections and spatial aliasing artefacts.

Chapter 3

Perceptually Weighted Multizone Soundfields

Overview: *In this chapter, we propose and evaluate an efficient approach for practical reproduction of multizone soundfields for speech sources. The reproduction method, based on a previously proposed approach, utilises weighting parameters to control the soundfield reproduced in each zone. An interpolation scheme is proposed for predicting the weighting parameter values of the multizone soundfield model that otherwise requires significant computational effort. We also propose a method for the reproduction of multizone speech soundfields using perceptual weighting criteria. Psychoacoustic models are used to derive a space-time-frequency weighting function to control leakage of perceptually unimportant energy from the bright zone into the quiet zone. An efficient codebook implementation is described, which uses predetermined weights based on desired soundfield energy in the zones. We perform simulations to gauge the performance of the methods. We show that the interpolation scheme can significantly reduce computation time with little error in the reproduced soundfield when compared to reproduction without interpolated weighting parameters. The perceptual impact on the quality of the speech reproduced using the interpolation method is also shown to be negligible. We also show that the perceptual weighting technique is capable of improving the spatial mean squared error for reproduced speech* *in the bright zone. Results indicate that the perceptual model can lead to a significant reduction in the spatial error within the bright zone whilst requiring significantly less loudspeaker signal power for cases where zones occlude each other. By using soundfield codebooks determined using the proposed approaches, practical reproduction of dynamically weighted multizone soundfields of wideband speech could be achieved in real-time.*

3.1 Introduction

Spatial audio reproduction gives listeners a full experience of the acoustic environment, including the sound source, and has been further extended to multizone soundfield reproduction, which provides audio in spatially separated regions from a single set of loudspeakers, originally proposed in [24]. They may also be used for suppressing, or cancelling, audio arriving from outside a targeted listening zone [110]. The multizone approach has many applications such as the creation of personal sound zones in multi-participant teleconferencing, entertainment/cinema and vehicle cabins where personal sound zones are optimised to provide one, or many, listener(s) with individual acoustic material [25].

Many existing approaches to multizone sound field reproduction attempt to completely suppress leakage between zones (interzone interference), which can result in either: loudspeaker signal amplitudes that are too large; or levels in zones that are too low. A method allowing weighted control between zones was introduced in [109]. The approach uses an orthogonal basis expansion which reduces the problem to the reconstruction of a set of basis wave fields and allows each zone to be weighted according to the importance of its reproduction. This weighting improves the practical feasibility of the system by relaxing the requirement of completely quiet zones outside the target bright zone. The theory in [106] was extended in [107] to include a similar weighting criteria as used in [109].

In order to maintain the perception of individual sound zones it is necessary to minimise the perceived interzone interference, which consequently maximises the apparent acoustic separation of the zones. This is difficult to achieve in situations where a desired soundfield in the bright zone is obscured by or directed to another zone, as the system requires reproduction signals many times the amplitude of what is reproduced within any zone. This is known as the *multizone occlusion problem* [25], [158], [213] and has been dealt with in various ways, such as the control of planarity [99], orthogonal basis planewaves [109] and alleviated zone constraints [107], [109].

Requiring large signals in relation to the reproduced zones means the system is inefficiently directing its energy for the multizone reproduction, with most sound energy present in unattended regions. This may be undesirable at times where listeners commute between sound zones and could put unnecessary strain on loudspeaker drivers. More recent work has focused on alleviating the constraint such that the interference (or leakage) is allowed into other zones and allows for leakage control with a weighting function [107], [109]. Allowing the sound to leak into other zones can improve the practical feasibility of the system but decrease the individuality of zones.

While [109] assumes the same weight for each frequency, dynamically deriving the weights can be used to control the reproduction accuracy of individual frequency components within the bright and quiet zones. For example, the weightings can be based on the perceptual importance of particular frequencies in the zones in an effort to improve the overall perceived sound quality. However, this results in increases in computational complexity. To reduce this complexity and create a more practical solution, we propose in this chapter the interpolation of spatial components of the reproduction along different domains, such as the weighting domain and frequency domain.

The control of acoustic components to enhance the perception of a signal has been researched thoroughly for applications such as compression [164]. The relationship between the quality in the bright zone and interference in other zones has been subjectively tested [100], however, the occlusion problem is not directly addressed and the planarity control does not directly address human perception. Hence, perceptual models are employed in this chapter in order to enhance the experience in personal sound zones, especially where the occlusion problem is present. Leaked sound energy is treated as unwanted noise in other zones and controlled such that it is perceptually less noticeable as indicated by established psychoacoustic models.

In general, multizone soundfield reproduction systems may be implemented for an arbitrary number of zones (in this chapter we simplify the problem to two zones) where in each zone a different soundfield may be desired. When there is perfect reproduction using the system, i.e. no error and no interference, a perceptual frequency masking threshold can be defined using each of the desired zones frequency spectra, below which all interference from other zones will not be perceived. In practice, limitations of the soundfield reproduction techniques when using particular reproduction geometry will result in total leakage in zones rising above the perceptual masking threshold. The goal of multizone soundfield perceptual masking is to then adjust the reproduction across all zones to minimise the sum of the leakage contributed from each other zone that is present above the perceptual masking threshold.

When performing perceptual frequency masking, however, the multizone solution becomes dependent on the acoustic power at neighbouring frequencies and requires updates to the reproduction loudspeaker filters as the reproduced content changes. Hence, requiring regular updates to the filters at regular time intervals. For the case where the desired reproduction sound level in one zone is silence, there is no potential for frequency masking within that zone and therefore more effort is required by the system to prevent leakage into this zone. As the number of zones of silence increases out of the total number of zones, there becomes less masking available for the whole perceptually weighted reproduction to make use of. For a reproduction with a single bright zone and where all other zones are silent, the problem reduces to a standard non-perceptual multizone problem.

To investigate the efficacy of a perceptual reproduction method, a system is synthesised with varying linear interpolation distances by using different resolution lookup tables (LUTs) for storing pre-computed loudspeaker weights and soundfield values. The synthesis comprises reproducing wideband zones where individual zones are weighted uniformly over space with weights that are in the centre of interpolation regions, optimised to minimise the error between the reproduced spectra and the desired spectra. The approach is validated by comparing the reproduced zone signals from the interpolation method with signals reproduced without interpolation using Mean Squared Error (MSE) and Perceptual Evaluation of Speech Quality (PESQ) [177] measures. The method is extended with psychoacoustic models and analysed with sound pressure level and spatial soundfield error.

In section 3.2, we begin with an explanation of the weighted multizone soundfield method used in this work and discuss the proposed dynamically weighted multizone approach. The interpolation method is described in section 3.3 and the psychoacoustic models are introduced in section 3.4. The results from evaluations of the proposed approaches are given in section 3.5 and conclusions outlined in section 3.6.

3.2 Weighted Multizone Wideband Soundfields

The multizone soundfield reproduction layout used in this chapter is shown in Figure 3.1 and contains a reproduction region, $\mathbb{D}$, with a radius of R. The reproduction region consists of three regions called the bright, quiet and unattended zones which are denoted as $\mathbb{D}_{\mathrm{b}}$, $\mathbb{D}_{\mathrm{q}}$ and $\mathbb{D} \cap (\mathbb{D}_{\mathrm{b}} \cup \mathbb{D}_{\mathrm{q}})'$, respectively. The centres of $\mathbb{D}_{\mathrm{b}}$ and $\mathbb{D}_{\mathrm{q}}$ have a distance of r_z from the centre of $\mathbb{D}$ and each of these zones has a radius of r. Loudspeakers are positioned with a distance of R_l from the centre of $\mathbb{D}$ on an arc subtending an angle of ϕ_{L}. The loudspeakers start at angle ϕ and reproduce plane-wave speech soundfields in $\mathbb{D}_{\mathrm{b}}$ with an angle of θ.

In the orthogonal basis expansion method of weighting multizone soundfields [109], a spatial weighting filter, $w(\mathbf{x})$, is used to control the reproduction of sound within each of the zones. This approach can be used with space-time-frequency dependent weighting functions, $w(\mathbf{x}; n, k)$, which allows the weighting functions to be adapted based on the signal characteristics of the target soundfield. We denote w_{b}, w_{q} and w_{u} as the weights for $\mathbf{x}_{\mathrm{b}} \in \mathbb{D}_{\mathrm{b}}$, $\mathbf{x}_{\mathrm{q}} \in \mathbb{D}_{\mathrm{q}}$ and $\mathbf{x}_{\mathrm{u}} \in \mathbb{D} \cap (\mathbb{D}_{\mathrm{b}} \cup \mathbb{D}_{\mathrm{q}})'$,

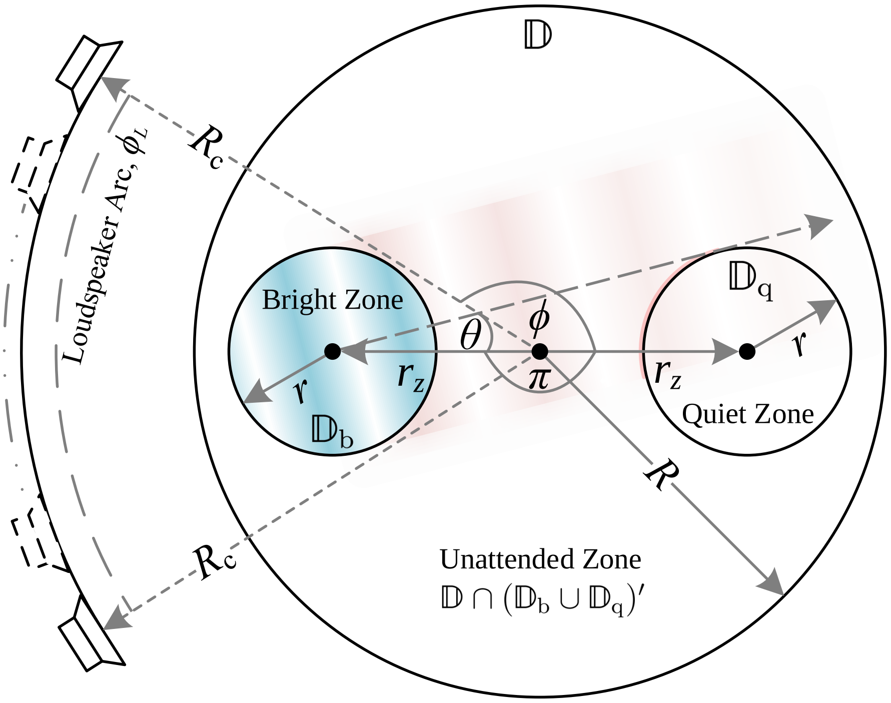

Figure 3.1: A weighted multizone soundfield reproduction layout is shown. The shading depicts the desired bright zone soundfield partially directed towards the quiet zone causing the occlusion problem.

respectively. The time domain reproduced soundfield pressure, $\hat{p}_w(\mathbf{x};n)$, at any point in the reproduction region can be obtained using an inverse discrete Fourier transform on the spatial short-time frequency domain soundfield values,

$$\hat{p}_w(\mathbf{x};n) = \mathrm{Re}\left\{ \frac{1}{K} \sum_{m \in [\![K]\!]} S_w^\mathrm{a}(\mathbf{x};n,k_m) \exp\Big(icnk_m / 2\hat{f} \Big) \right\}, \tag{3.1}$$

where $m \in [\![K]\!]$ is the frequency index of the set of frequencies, k_m, and $\hat{f}$ is the maximum temporal frequency, which is typically half the sampling frequency (Nyquist frequency) in numerical implementations. In (3.1), $S_w^\mathrm{a}(\mathbf{x};n,k)$ is a zone weighted reproduced soundfield, which is derived as a function of a desired soundfield, $S^\mathrm{d}(\mathbf{x};n,k)$, and a weighting function, $w(\mathbf{x};n,k)$ using the approaches outlined in chapter 2 [109]. Here, $\mathbf{x}$ is a given position, n is a point in time and k is a specific wavenumber. $S_w^\mathrm{a}(\mathbf{x};n,k)$ is summed for K different sinusoidal components. In this chapter, $k = 2\pi f/c$ and $c = 343\,\mathrm{m\,s^{-1}}$.

The weighting associated with (3.1), $w(\mathbf{x}; n, k)$, allows independent weighting of soundfield components in space and time. It is then possible to define the reproduced space-time-frequency domain signal for a particular input as,

$$\widehat{Y}_w(\mathbf{x}; n, k) = S_w^{\mathrm{a}}(\mathbf{x}; n, k) Y(n, k) \tag{3.2}$$

where $\widehat{Y}_w(\mathbf{x}; n, k)$ is the time-frequency signal at an arbitrary location, $\mathbf{x}$, in the reproduction region, $\mathbb{D}$, and $Y(n, k)$ is obtained from the discrete short-time Fourier transform of the windowed frame of input $y(n)$. Using overlap-add reconstruction we can obtain the time-domain signal at any point in $\mathbb{D}$ where a different weighting function can be used for each space-time-frequency. The weighting function can now be used to control the leaked content into the quiet zone in the space-time-frequency domain.

3.3 A Priori Soundfield Synthesis & Weighting

It is computationally demanding to construct a weighted multizone soundfield using the methods discussed in the previous section, and in chapter 2, due to the QR factorisation involved for all time-frequency components (e.g. a three second audio file sampled at 16 kHz may require at least approximately 48×10^3 independent soundfield syntheses, one for each time-frequency sample over the entire field of interest). It is not uncommon for the syntheses to be repeated, which results in redundant computation. To make good use of the repeated computations, the loudspeaker weights and soundfield pressure samples can be synthesised and stored for later referral. Interpolation of smooth, preferably monotonic, functions can further reduce computation and error caused by truncated modes. We propose using Look-Up Tables (LUTs) (codebooks) to store matrices of pre-computed weighted soundfield values to be used for a particular multizone setup or wideband reproduction; an example pressure magnitude matrix is shown in Figure 3.2. The reproduced soundfield and the required weighting is linearly related to the content being reproduced.

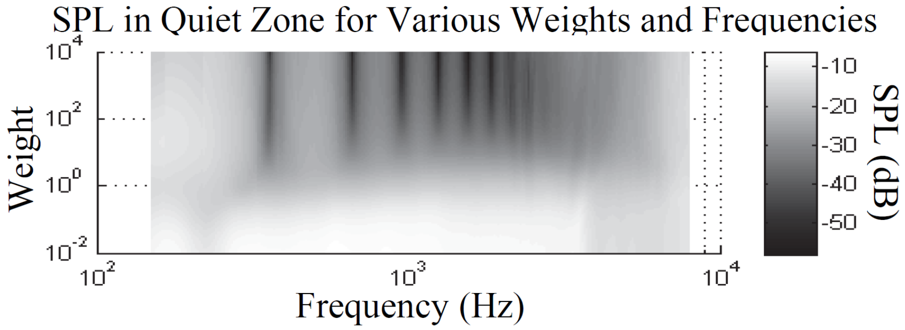

Figure 3.2: Example high resolution matrix of values for absolute sound pressure levels in the quiet zone.

The LUTs are defined as matrices of soundfield reproduction values for a particular range of frequencies and weights. We assume all zone weightings are fixed except for those in the zone of interest, for instance, w_{q}. The relationship between the zone weight and the soundfield synthesis is not straightforward. To simplify the explanations, we denote a soundfield that has a varying weight in $\mathbb{D}_{\mathrm{q}}$ as a function of the weight for that zone, w_{q}, as $S_w^{\mathrm{a}}(\mathbf{x}_{\mathrm{q}}, w_{\mathrm{q}}; n, k)$. The relationship between the soundfield and the loudspeaker signals is described in section 2.3.2. The LUT for varying w_{q} is then,

$$\mathbf{S}_{K\times Z}^{(w)}(\mathbf{x}_{\mathrm{q}}) = \begin{bmatrix} S_w^{\mathrm{a}}\Big(\mathbf{x}_{\mathrm{q}}, w_{\mathrm{q}}^{(\min)}; n, k^{(\min)}\Big) & \cdots & S_w^{\mathrm{a}}\Big(\mathbf{x}_{\mathrm{q}}, w_{\mathrm{q}}^{(\max)}; n, k^{(\min)}\Big) \\ \vdots & \ddots & \vdots \\ S_w^{\mathrm{a}}\Big(\mathbf{x}_{\mathrm{q}}, w_{\mathrm{q}}^{(\min)}; n, k^{(\max)}\Big) & \cdots & S_w^{\mathrm{a}}\Big(\mathbf{x}_{\mathrm{q}}, w_{\mathrm{q}}^{(\max)}; n, k^{(\max)}\Big) \end{bmatrix}, \tag{3.3}$$

with K frequencies in the range $\{k^{(\min)}, \ldots, k^{(\max)}\}$, and Z zone weights in the range $\{w_{\mathrm{q}}^{(\min)}, \ldots, w_{\mathrm{q}}^{(\max)}\}$. The set of frequencies is logarithmically spaced as it closely resembles the spacing of the Bark scale [164] and the set of weights is logarithmically spaced to provide large control ranges in sound pressure level (SPL) covering the human hearing range. The matrix of soundfield values can be interchanged with the equivalent loudspeaker driving weights.

To evaluate the spatial error and perceptual effects of quantising and interpolating soundfield values, we provide a comparison between two LUTs (see section 3.5). We apply the MSE measure to the interpolated values of lower and higher resolution

LUTs,

$$\epsilon_{\mathbf{S}} = \frac{1}{K'Z'} \sum_{[\![K']\!]} \sum_{[\![Z']\!]} \Big(\widetilde{\mathbf{S}}^{(w)}_{K' \times Z'} - \mathbf{S}^{(w)}_{K' \times Z'} \Big)^2, \tag{3.4}$$

where $\epsilon_{\mathbf{S}}$ is the MSE for an interpolated LUT, $\widetilde{\mathbf{S}}^{(w)}_{K' \times Z'}$, relative to the highest resolution LUT, $\mathbf{S}^{(w)}_{K' \times Z'}$, K' is the highest number of frequencies in a LUT and Z' is the highest number of weights in a LUT. The interpolated LUT, $\widetilde{\mathbf{S}}^{(w)}_{K' \times Z'}$, is a matrix of size $K' \times Z'$ obtained from interpolation of a smaller matrix, $\mathbf{S}^{(w)}_{K \times Z}$. In this chapter, we perform bicubic interpolation on the regular grid in the logarithmic scale.

3.4 Psychoacoustic Weighting Models

The weighting function can be used to control the energy leaked between zones by relating the weights to the desired reproduced signal. The leaked audio spectrum can be designed such that it is masked by another spectrum in the same zone. From this idea, we propose psychoacoustic modelling of the weighting function to reduce the perceptual effect of sound leakage in the quiet zone.

3.4.1 The Hearing Threshold

The benefit of using zone weighting is that the hard constraint of zero energy in the quiet zone is alleviated and sound energy may be allowed to leak into the quiet zone. Doing so, however, will result in the quiet zone having an increased level of soundfield energy, which can be less than ideal if the increased energy level is perceivable.

Due to the human threshold of hearing in quiet, we redefine the quiet zone to one that is perceivably of zero sound to humans. This then allows weighted multizone systems to remain perceptually quiet whilst simultaneously relieving constraints on the soundfield reproduction. The threshold in quiet has been well established with frequency dependent functions that provide a good approximation [164], [166], which we covered in chapter 2.

Using the space-time-frequency domain weighting established above, it is possible to apply the threshold in quiet approximation to (3.2) where $w(\mathbf{x}; n, k)$ is chosen so that the output in the quiet zone, $\widehat{Y}_w(\mathbf{x}_\mathrm{q}; n, k)$, is as close to the threshold in quiet as possible. Then, using the LUT, $w(\mathbf{x}; n, k)$ can be chosen to minimise the difference,

$$\min_{w_\mathrm{q}} \Big(\widehat{Y}_w(\mathbf{x}_\mathrm{q}; n, k) - \mathcal{Y}(\mathbf{x}_\mathrm{q}; n, k) \Big), \tag{3.5}$$

where $\mathcal{Y}(\mathbf{x}_\mathrm{q}; n, k)$ is a space-time-frequency dependent function describing the perceptual criteria. In this work SPL in dB is relative to the threshold of hearing $p_r = 20\,\mu\mathrm{Pa}$.

3.4.2 Spreading Functions to Reduce Multizone Error

The work on weighted multizone reproductions in [109] reveals that larger zone weighting suppresses the quiet zone at the expense of increased error in the bright zone. The soundfield in the bright zone is less erroneous when the zone weighting is small for the quiet zone; a benefit of allowing energy to leak. When the constraint on the quiet zone is such that minimal energy will leak, then the error in the bright zone increases.

The spatial errors shown in Figure 3.3 are measured with [109]

$$\epsilon_\mathrm{b}(n, k) = \frac{\int_{\mathbb{D}_\mathrm{b}} \left| S^\mathrm{d}(\mathbf{x}; n, k) - S^\mathrm{a}_w(\mathbf{x}; n, k) \right|^2 d\mathbf{x}}{\int_{\mathbb{D}_\mathrm{b}} \left| S^\mathrm{d}(\mathbf{x}; n, k) \right|^2 d\mathbf{x}}, \tag{3.6}$$

where $\epsilon_\mathrm{b}(n, k)$ is the spatial error in the bright zone and $S^\mathrm{d}(\mathbf{x}; n, k)$ is the desired soundfield. Jin et al. [109] reported that for $k = 2\pi(2\,\mathrm{kHz})/c$ the spatial error is greater than $-5\,\mathrm{dB}$ when the quiet zone is occluded by the bright zone and has a large weight (equivalent to $w_\mathrm{q} = 10$), however, the spatial error is less than $-20\,\mathrm{dB}$ when the weight is alleviated (equivalent to $w_\mathrm{q} = 0.1$).

For applications where secondary content is superimposed over the quiet zone for a second user to consume, thus making it no longer truly quiet, it is possible to significantly improve the reproduction error. In Figure 3.3 it is shown that using a

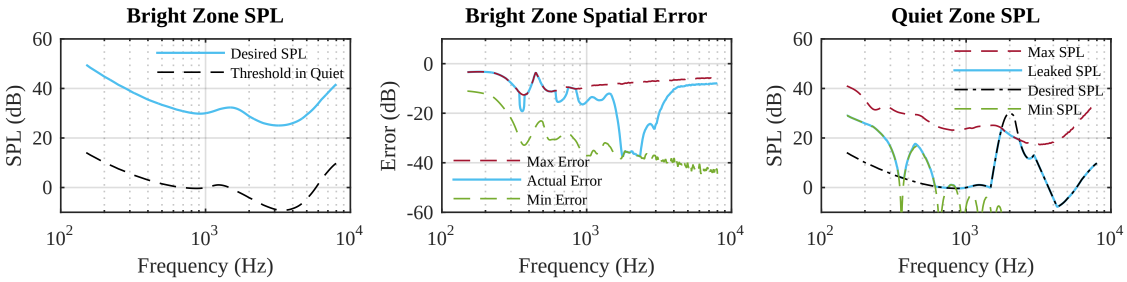

Figure 3.3: Multizone soundfield reproduction with perceptual weighting in the quiet zone. The desired bright zone signal is an equal loudness curve at 30 phon [166] and a 2 kHz masker signal at 30 dB SPL is present in the quiet zone. The red and green dashed lines show the worst and best case scenarios, respectively. The bright zone error is calculated using (3.6). The “Leaked SPL” shows the result after controlling the interzone interference with w_q.

spreading function to mask apparent sounds, in the target quiet zone, can reduce the error of the reproduction in the bright zone. This result is due to the sound energy at particular frequencies leaking into the quiet zone with no perceptual effect. If the target quiet zone contains many different frequency components, then the majority of bright zone energy could be allowed to leak into the quiet zone unnoticed and, thus, reduce the error in the bright zone substantially.

3.5 Results

This section describes the evaluations and results of the proposed interpolation techniques and psychoacoustic zone weighting methods that are outlined above.

3.5.1 Evaluation Setup

We evaluated the multizone soundfield layout of Figure 3.1 with $r = 0.3\,\mathrm{m}$, $r_z = 0.6\,\mathrm{m}$, $R = 1\,\mathrm{m}$, $R_c = 1.5\,\mathrm{m}$, $\theta = \sin^{-1}(r/2r_z) \approx 14.5°$ and $\pi \approx 3.141\,59\,\mathrm{rad}$. The setup was chosen similar to [109] and θ was chosen such that the reproduced planewave would interfere with approximately half the quiet zone. This choice of angle results in a slight occlusion problem where the range of weighting control is larger than for no occlusion and full occlusion. Signals sampled at 16 kHz were converted to the time-frequency domain using a Hamming window (with 50 % overlap) and discrete

Fourier transform (DFT) of length 1024. The LUTs were built and evaluated for reproductions with $L = 65$, $\phi = \theta + \pi/2$ and $\phi_{\mathrm{L}} = \pi$. The evaluation setup has an aliasing frequency of approximately 1.9 kHz. We will see later in chapter 4 that a the proposed redefined and reformulated zone-based spatial aliasing frequency calculation provides a much higher frequency than was previously possible with existing techniques that use setups similar to this chapter.

The tables were built with the soundfield pressures for all $\mathbf{x} \in \mathbb{D}_{\mathrm{b}} \cap \mathbb{D}_{\mathrm{q}}$ and averaged over $\mathbb{D}_{\mathrm{b}}$ and $\mathbb{D}_{\mathrm{q}}$. Each soundfield zone consisted of 2724 spatial samples and the soundfield zone pressure was approximated from the mean over the zone. The zone weights were chosen as $w_{\mathrm{b}} = 1.0$ and $w_{\mathrm{u}} = 0.05$ following [109] and the variable weight was w_{q}. The effect of w_{q} on the input signal was evaluated using (3.2) and (3.1).

Without interpolating the LUTs, the highest frequency resolution was 512 frequencies up to $\hat{f} = 8\,\mathrm{kHz}$ (based on the 1024 length DFT) and 256 different zone weighting values, which resulted in negligible reconstruction error. Each table was built for logarithmically spaced resolutions, consecutively halving, and decreasing in resolution down to 16 frequencies and 8 weights. In this work, we used $w_{\mathrm{q}} \in \{10^{-2}, \ldots, 10^{4}\}$ which extends the range used in [109]. The error between the different LUTs was evaluated using (3.4), where the highest resolution for frequencies was $K' = 512$ and for weights was $Z' = K'/2$. The set of frequency and weight resolutions to be evaluated were $K = \{16, 32, 64, 128, 256\}$ and $Z = K/2$, respectively. The proposed interpolation approach was evaluated using PESQ [177] to estimate the perceptual quality of the reproduced soundfields.

Speech samples for the evaluation were taken from the TIMIT corpus [214] where 20 speech segments, of approximately 3 s in length, were chosen randomly. The random choice was constrained to a final male to female speaker ratio of 1 : 1. The reference signal for the PESQ algorithm was the original speech signal. PESQ values were obtained for the reproduced speech soundfields using the different resolution LUTs and then mapped to the PESQ Mean Opinion Score (MOS) [215]. These reproductions used $w_\mathrm{q} = \{10^{-0.5}, 10^{0.5}, 10^{1.5}, 10^{2.5}\}$ such that they existed primarily in the centre of the interpolation regions. This allowed the highest resolution LUT to be evaluated, however, due to the computational complexity, was limited to four different weights.

Using (3.5) for the perceptual weighting, w_q was chosen to match the quiet zone to a given level, $\mathcal{Y}(\mathbf{x}_\mathrm{q}; n, k)$. In this chapter, we chose $\mathcal{Y}(\mathbf{x}_\mathrm{q}; n, k)$ to be the threshold in quiet using the ISO226 standard [166] with additional masking curves using the ISO/IEC MPEG Psychoacoustic Model 2 spreading function [164] defined in (2.90).

3.5.2 Interpolation Method Evaluation Results

Figure 3.4 shows an example LUT for the bright and quiet zone samples. There is a significant contrast in SPL levels between the bright and quiet zones and spatial aliasing above 1.9 kHz can be seen in the quiet zone LUT. The increase in zone weighting can be seen to decrease the SPL in the quiet zone below the aliasing frequency. The bright zone LUT pressure level remains consistent around 0 dB regardless of the zone weight and is less susceptible to spatial aliasing. The horizontal discontinuities in Figure 3.4 are due to the truncated modes.

Analysing the MSE between the different interpolation distances (Figure 3.5) indicates that the lower resolution LUTs cause little error whilst requiring significantly less computational effort than those of the higher resolution. The labels show the relative decrease in the number of reproduced soundfields, which is up to 1024 times less than, at 0.10 %, the number of computations of the highest resolution LUT. An MSE of −85 dB is comparable to high end audio systems and can be provided by the low resolution LUT. The general trend is that an increase in the interpolation distance increases the MSE.

In Figure 3.6, the increased MSE caused by larger interpolation distances has no significant impact on the perceptual quality. The maximum mapped MOS is indicated by the red line. Figure 3.6 does show, however, a slight increase in the variation of the PESQ MOS, as indicated by the 95 % confidence interval markers,

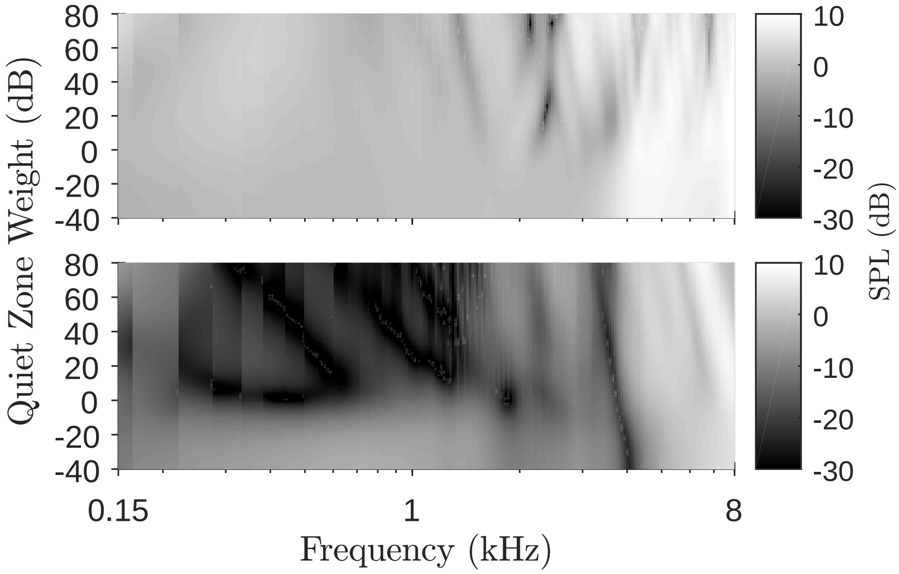

Figure 3.4: LUT from the aliasing setup for the bright zone (top) and quiet zone (bottom).

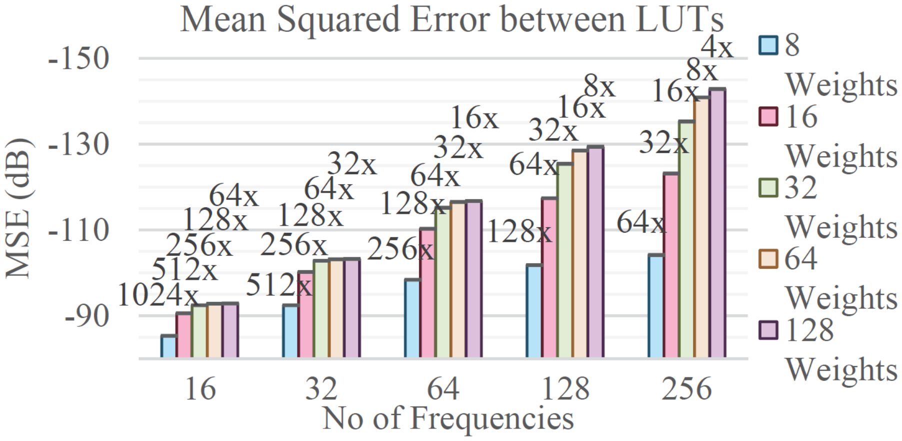

Figure 3.5: MSE between different LUT resolutions. Labels show the relative complexity decrease from $\mathbf{A}_{u'v'}$.

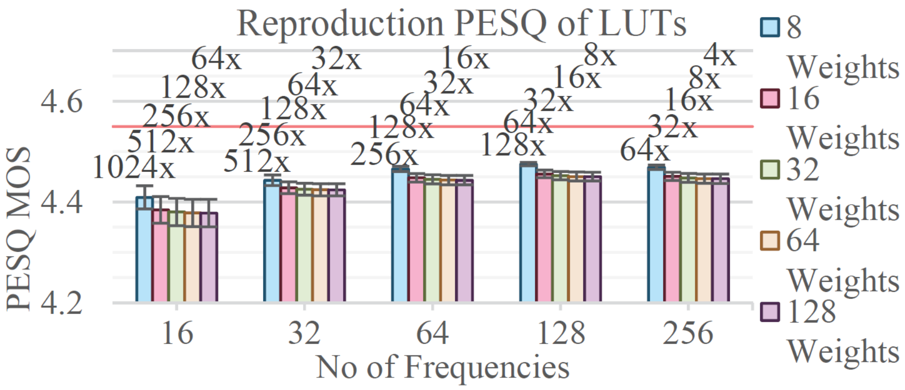

Figure 3.6: PESQ MOS between weighted speech files reproduced by different LUTS with 95 % confidence intervals. Labels show the relative complexity decrease from $\mathbf{A}_{u'v'}$. Red line indicates maximum mapped PESQ MOS.

where larger interpolation distances are required. This shows that interpolating the zone weighted soundfield values has an insignificant perceptual effect on the reproduction and decreases the computational complexity by up to 1024 times.

3.5.3 Reduced Bright Zone Error from Psychoacoustic Masking

The error induced from the multizone reproduction of the speech soundfields is, again, gauged using the MSE of the reproduced speech with reference to the original speech. To obtain an approximation of the reproduced speech the mean of the simulated spatial pressure samples, obtained with the approach of section 3.4, are used over $\mathbb{D}_\mathrm{b}$ and $\mathbb{D}_\mathrm{q}$.

Upon analysing the spatial MSE of different reproduced speech segments, it becomes apparent from 3.5 and 3.7 that the majority of the error measured in the bright zone from the reproduction is spatial error. The sampling theory used to obtain the reproduced speech does not use spatial information, however, (3.6) can be used to evaluate the spatial error or, alternatively, the measure of planarity could be used [99]. The application of perceptual criteria is then a natural reasoning for the reduction of spatial error in the multizone reproduction.

The maximum improvement in MSE of the speech in the bright zone is $-10.5\,\mathrm{dB}$,

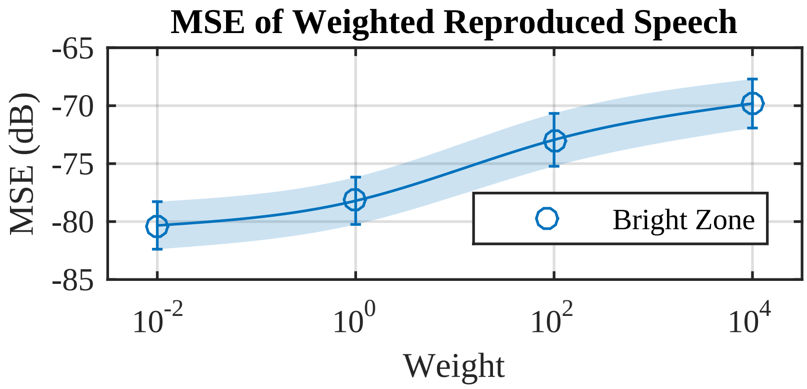

Figure 3.7: Shows the MSE of reproduced speech signals in the bright zone for different uniform weighting functions (w_q).

from -69.8 dB for $w_\mathrm{q} = 10^4$ to -80.3 dB for $w_\mathrm{q} = 10^{-2}$, and can be seen in Figure 3.7. Even though there is a difference of -10.5 dB, the error in the reproduced speech is minimal. However, a maximum improvement in spatial error for the bright zone, $\epsilon_\mathrm{b}(n, k)$, averaged for all frequencies is -24.0 dB, from -7.4 dB for $w_\mathrm{q} = 10^4$ to -31.5 dB for $w_\mathrm{q} = 10^{-2}$, and can be seen in Figure 3.3.

A reduction in spatial error is depicted in Figure 3.8 where the perceptual weighting uses $w_\mathrm{q} = 10^{-2}$ instead of $w_\mathrm{q} = 10^4$, which gives a smaller difference between the desired soundfield and reproduced soundfield. Recall that a 2 kHz masker signal in the quite zone can allow the spatial error in the bright zone to be reduced, as was shown earlier in Figure 3.3. In Figure 3.8, the magnitude difference is calculated from $\left|S^\mathrm{d}\right| - \left|S^\mathrm{a}_w\right|$ and the phase difference from $\arg(S^\mathrm{d}/S^\mathrm{a}_w)$. The equivalent improvement in $\epsilon_\mathrm{b}(n, k)$ and required loudspeaker power due to the perceptual weighting is -28 dB and 65% less, respectively.

3.6 Conclusions and Contributions

In this chapter, we proposed a method for building multizone soundfields for speech signals that allows dynamic control of the weighting between zones. We have proposed a method for reducing the computational effort involved when dynamically weighting zones for speech signals. A novel method for perceptually weighting multizone speech

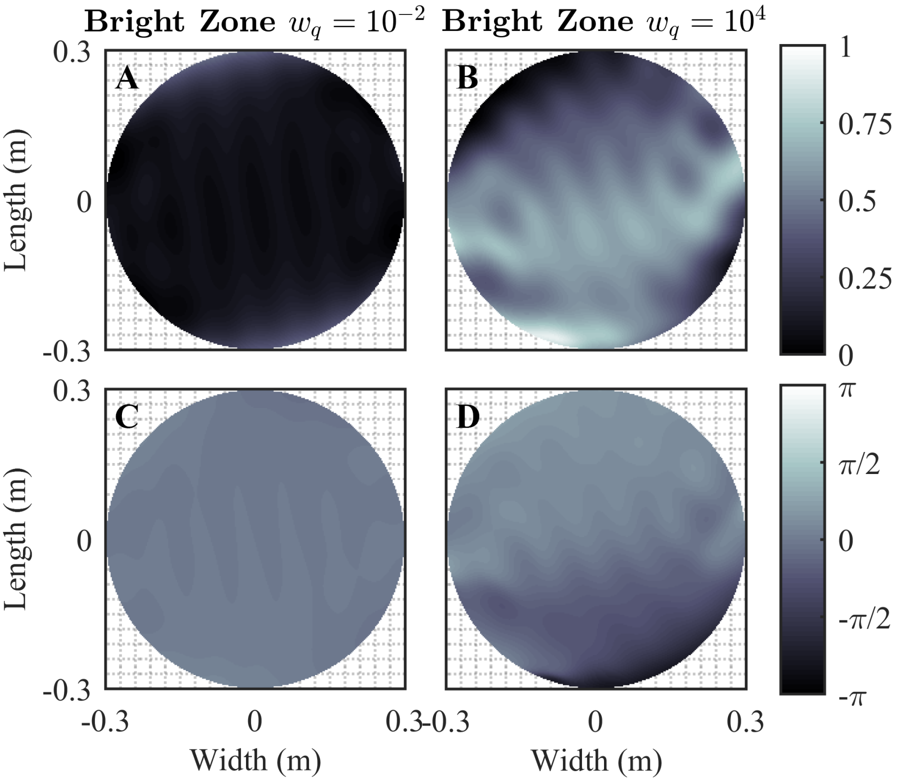

Figure 3.8: Difference between the desired soundfield and actual weighted soundfield for $f = 2\,\text{kHz}$. A and B show the magnitude difference and C and D show the phase difference. A and C are for $w_\text{q} = 10^{-2}$ and B and D are for $w_\text{q} = 10^{4}$.

soundfields is proposed, which can improve error in bright zones, especially when the occlusion problem is present. The LUT based method has been evaluated and shows indiscernible impact on perceptual quality of reproductions and decreased computational complexity. The interpolation scheme evaluations show PESQ MOS values of 4.4 and MSE of -85 dB are achievable at 1024 times less soundfield syntheses. Perceptual weighting is shown to improve the MSE for reproduced speech in the bright zone from -69.8 dB to -80.3 dB and significantly reduce the spatial error on average from -7.4 dB to -31.5 dB whilst requiring less loudspeaker driving power.

In the next chapter, we further consider the acoustic quality in zones through perceptual measures. Acoustic privacy between zones is discussed and perceptual measures, such as speech intelligibility and speech quality, are used to enhance the privacy and maintain quality in personal sound zones.

Chapter 4

Multizone Soundfield Privacy and Quality Based Speech Maskers

Overview: *Reproducing zones of personal sound is a challenging signal processing problem which has garnered considerable research interest in recent years. We introduce in this work an extended method to multizone soundfield reproduction which overcomes issues with speech privacy and quality. Measures of Speech Intelligibility Contrast (SIC) and speech quality are used as cost functions in an optimisation of speech privacy and quality. Novel spatial and (temporal) frequency domain speech masker filter designs are proposed to accompany the optimisation process. Spatial masking filters are designed using multizone soundfield algorithms which are dependent on the target speech multizone reproduction. Combinations of estimates of acoustic contrast and long term average speech spectra are proposed to provide equal masking influence on speech privacy and quality. Spatial aliasing specific to multizone soundfield reproduction geometry is further considered in analytically derived low-pass filters. Simulated and real-world experiments are conducted to verify the performance of the proposed method using semi-circular and linear loudspeaker arrays. Simulated implementations of the proposed method show that significant speech intelligibility contrast and speech quality is achievable between zones. A range of Perceptual Evaluation of Speech Quality (PESQ) Mean Opinion Scores (MOS)* *that indicate good quality are obtained while at the same time providing confidential privacy as indicated by SIC. The simulations also show that the method is robust to variations in the speech, virtual source location, array geometry and number of loudspeakers. Real-world experiments confirm the practical feasibility of the proposed methods by showing that good quality and confidential privacy are achievable.*

4.1 Introduction

Personal sound zones, such as the individual sound environments provided to listeners by means of spatial multizone soundfield reproduction, without the need for physical barriers or headphones, have gained significant interest of researchers in recent years [25], [100], [112]. Some applications of personal sound zoning systems include vehicle cabin entertainment/communication systems, multi-participant teleconferencing, cinema surround sound systems and personal audio in restaurants/cafés [25], [37], [216]. In some cases, it is desirable to maintain quiet areas by cancelling or suppressing audio from adjacent zones. Quiet areas may be desired so that, for example, vehicle satellite navigation instructions may be heard by drivers without disturbing passengers or so that someone may read/work in silence while someone else listens to a talk show or news in the same room [217]. Limitations exist in the majority of work with multizone soundfield reproduction systems where sound is audible (and likely intelligible) for listeners in designated quiet zones and/or the perceived quality in target reproduction zones is degraded from interference caused by other zones [100], [112], [218].

Multizone soundfield systems attempt to eliminate audio spatially leaked between zones [25], [85]–[87], [219]. Multizone soundfield reproductions constraining quiet zones to zero energy may result in uncontrolled regions containing sounds many times the amplitude of the target bright zone. Techniques that improve performance in these situations optimise over spatial regions with planarity [99], basis plane-waves [109] and reduced constraints [107], [109]. In chapter 3, we showed that spatial weighting of importance for each zone [107], [109] can be used to control the amount of leakage and improve the performance of the multizone reproduction system.

Multizone soundfield reproductions designed for single frequency (mono-frequent) soundfields have been extended to wideband soundfields including speech [92]. Recent research has investigated the perceptual quality of multizone soundfields [100] and in chapter 3 we proposed methods to improve the quality using psychoacoustic models. In this chapter, we address open questions on the perception of leakage and what this means for speech privacy amongst zones.

Reproducing personal sound in public spaces, such as open-offices, brings concerns regarding privacy between zones. Existing methods do not specifically address the problem of information leaking between zones and may lead to the ability of users to deduce what content is being reproduced in other zones, e.g. in private teleconference meetings. Good speech privacy requires that the leaked speech signal is not intelligible [208], [209]. Although research has shown how to synthesise and reproduce wideband speech soundfields in multiple zones, state-of-the-art methods still lack the acoustic contrast between zones to provide speech privacy [100], [112], [218]. For reproduction of speech at a level of 60 dBA in a target bright zone, state-of-the-art methods can provide a quiet zone level down to ≈ 35 dBA for zones large enough to fit a human listener and for sound arriving from any direction. However, in order to provide speech privacy in a quiet room, a consistent acoustic contrast of ≈ 60 dBA may be required, which would maintain a quiet zone level below the threshold of hearing (≈ 0 dBA). In simulated reverberant rooms, a room impulse response may be manipulated to control privacy [212]. The level of acceptable interference while in different listening scenarios has also been studied and has shown that, in some scenarios, experienced listeners have an acceptability threshold of less than -40 dB [220]. Most measurements of speech intelligibility, and thus privacy, are based on the mutual information conveyed between a speaker and listener [188]. In this chapter, we will show how the mutual information between different zones in a multizone reproduction scenario can be controlled, for the goal of maintaining speech privacy, by using spatially synthesised masking.

Theoretically, with many loudspeakers, any soundfield can be synthesised to meet specific requirements. However, in practice, a reduced number of loudspeakers introduces deleterious phenomena such as spatial aliasing.

The fundamental problem of spatial aliasing in discretised soundfield reproductions has been investigated in [221] and shows that, in a multizone scenario, spatial aliasing can be considered another contributor to zone leakage. Analytical definitions have been formulated for the occurrence of aliasing in zoned soundfields [106], [109], [221] which can be used to account for its particular contribution to leakage. Another contributor to leakage is that caused by current multizone soundfield methods, where constraints on power and spatial error reduce acoustic contrast. It has been shown that acoustic contrast, and hence leakage, is frequency dependent [82], [97], [99], [109], [112] with most multizone soundfield synthesis and reproduction techniques, however, in chapter 3 we showed that the leakage can be partly controlled per frequency. Frequency dependent leakage leads to an unknown spectral distortion of the audio content across different spatial regions.

In this chapter, a novel method consisting of several stages for improving speech privacy in personal sound zones is proposed. The proposed measure, Speech Intelligibility Contrast (SIC), which is based on mutual information between spatial regions, is used to maximise speech privacy in multizone soundfield reproductions. Optimisations are formulated to maximise SIC and instrumental measures of subjective quality after extending the reproduction method used in the previous chapter from two dimensional (2-D) to three dimensional (3-D) wave equations.

Novel multizone soundfield dependent spatial and spectral masker filters are also incorporated in the method. The spatial masker filter is designed as a multizone soundfield filter which is dependent on the multizone soundfield reproduction scenario of the speech in the target bright zone. The spectral masker filters are designed as a combination of *a priori* estimates of the acoustic contrasts of both the masker signal and target speech signal multizone soundfield reproductions. Further, spectral shaping filters are designed to reduce the effects of aliasing, caused by discretised loudspeaker spacings, specifically on multizone soundfield reproductions. A combination of the proposed filters is used in masking the leaked speech in the quiet zone whilst leaving the target bright zone speech unimpaired.

The extended methods are analysed and evaluated to ensure a practical, systematic and robust procedure to improving speech privacy in personal sound zones. Experimental results are presented for both simulations and a real-world implementation using practical numbers of loudspeakers.

4.2 Weighted Multizone Speech Soundfields

This section overviews the soundfield synthesis and reproduction from the weighted orthogonal basis expansion [109], [112] and spherical harmonic expansion [2], [71], [159], [222] methods, respectively. The methods described later in this chapter rely on general properties (and combinations of properties) of multizone soundfield reproductions, such as acoustic contrast, loudspeaker layout, zone geometry and target zone soundfield wave fronts. The multizone techniques that can be used with the proposed methods are not limited to those described in this section, however, the descriptions in this section are given to facilitate the reader in understanding the proposed methods.

4.2.1 Notation, Definitions and Multizone Setup

Throughout this chapter, the following notations are used: time-domain functions and their frequency-domain function transformation are represented in lowercase and uppercase italics, respectively. Vectors and matrices are represented by lowercase and uppercase bold face, respectively. The set of all real numbers is $\mathbb{R}$, $\mathbb{R}^+ \triangleq \{x : x \in \mathbb{R}, x \geq 0\}$, the set of all natural numbers starting at zero is $\mathbb{N}_0$, sets of indices are given by $[\![X]\!] \triangleq \{x : x \in \mathbb{N}_0, x < X\}$ and the unit imaginary number is $i = \sqrt{-1}$.

A personal sound zone system is depicted in Fig. 4.1 where the reproduction

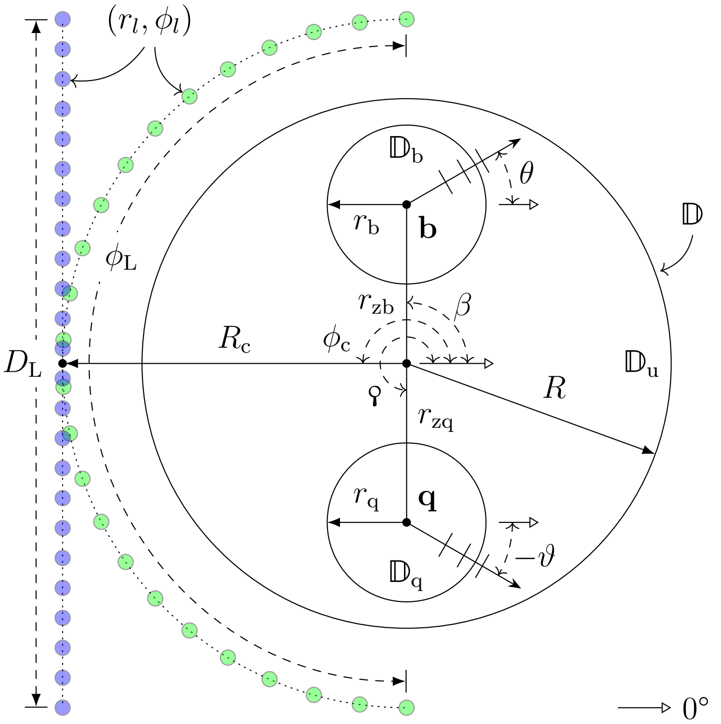

Figure 4.1: A multizone soundfield reproduction layout is shown for a semi-circular (green) and linear (blue) loudspeaker array.

region, $\mathbb{D}$, of radius R contains three sub-regions denoted by $\mathbb{D}_{\mathrm{b}}$, $\mathbb{D}_{\mathrm{q}}$ and $\mathbb{D}_{\mathrm{u}} = \mathbb{D} \setminus (\mathbb{D}_{\mathrm{b}} \cup \mathbb{D}_{\mathrm{q}})$ called the bright, quiet and unattended zone, respectively. The radius of $\mathbb{D}_{\mathrm{b}}$ and $\mathbb{D}_{\mathrm{q}}$ are r_{b} and r_{q}, respectively and have centre points $\mathbf{b} \equiv (r_{\mathrm{zb}}, \beta)$ and $\mathbf{q} \equiv (r_{\mathrm{zq}}, ϙ)$, respectively. Two separate loudspeaker geometries are shown for L loudspeakers with array centres at an angle of ϕ_{c} and distance R_{c} with the lth loudspeaker position $\mathbf{l}_l \equiv (r_l, \phi_l), l \in [\![L]\!]$. The semi-circular array is concentric with $\mathbb{D}$, has a radius R_{c} and subtends an angle ϕ_{L}. The linear array is of length D_{L}. The loudspeakers are assumed to behave like omnidirectional point sources for simplicity. The angle of a desired point-source or plane-wave in $\mathbb{D}_{\mathrm{b}}$ is θ and in $\mathbb{D}_{\mathrm{q}}$ is ϑ. The wavenumber is given by $k = 2\pi f/c$, where f is frequency and c is the speed of sound propagation through a medium. In this work, c is assumed to be constant and therefore, f and k are interchangeable within a multiplicative constant, $2\pi/c$.

4.2.2 Multizone Soundfield Reproduction Method

Any arbitrary soundfield can be described by a set of plane-waves arriving from every angle [2], including speech soundfields. A soundfield function, $S(\mathbf{x}; k)$, that fulfils the wave equation, where $\mathbf{x} \in \mathbb{D}$ is an arbitrary spatial sampling point, can be defined with an additional spatial weighting function, $w(\mathbf{x})$, as shown in the orthogonal basis expansion approach [109], [112] to multizone soundfield reproduction. This weighting function allows for relative importance between zones to be specified for reproduction. The weighted soundfield function used in this work can be written as

$$S(\mathbf{x}; k) = \sum_{h \in [\![J]\!]} \mathcal{W}_h \, P_h(\mathbf{x}; k), \tag{4.1}$$

where, for a given weighting function, the coefficients, $\mathcal{W}_h$, are for a set of plane-wave soundfields, $P_h(\mathbf{x}; k)$, and J is the number of basis plane-waves [109].

The frequency domain complex loudspeaker weights used to reproduce the soundfield over the plane[1] are [109], [143], [159]

$$W_l(k) = \frac{\Delta\phi_{\mathrm{s}}}{2\pi i k} \sum_{\bar{m}=-\overline{M}}^{\overline{M}} \sum_{h \in [\![J]\!]} \frac{i^{\bar{m}} \exp(i\bar{m}(\phi_l - \rho_h))}{\mathscr{h}_{\bar{m}}^{(1)}(r_l k)} \mathcal{W}_h \tag{4.2}$$

where $\overline{M} = \lceil kR \rceil$ is the maximum mode order (also known as the mode truncation length) [109], $\mathscr{h}_{\nu}^{(1)}(\cdot)$ is a νth-order spherical Hankel function of the first kind, $\rho_h = 2\pi(h-1)/J$ are the plane-wave angles, ϕ_l is the angle of the lth loudspeaker from the horizontal axis and $\Delta\phi_{\mathrm{s}}$ is the angular spacing of the loudspeakers. Here, $\mathcal{W}_h$ is chosen to minimise the difference between the desired soundfield and the actual soundfield [109].

To reproduce plane-wave speech soundfields, the set of loudspeaker signals can be found by applying $W_l(k)$ to the speech in the frequency domain and inverse transforming the signal back to the time domain. The set of framed loudspeaker

[1] Since the loudspeakers lie on a plane, an integration over elevation is carried out on the orthonormal spherical harmonics to simplify (6.4) and remove the dependence on elevation [2, Ch. 8].

signals in the time-frequency domain are given by[2]

$$Q_l(a,k) = W_l(k)\,Y(a,k) \tag{4.3}$$

where $Y(a,k)$ is the discrete Fourier transform of the ath overlapping windowed frame, from a total of A frames, of the input speech signal, $y(n)$. Each loudspeaker signal, $q_l(n)$, is reconstructed by performing overlap-add reconstruction with inverse transformed $Q_l(a,k)$ and the synthesis window. The synthesis and analysis windows are chosen such that they sum to a constant value for an overlap-add process. This results in the loudspeaker signals, which will reproduce the multiple zones.

Filtering each of the loudspeaker signals with their respective 3-D acoustic transfer function (ATF)[3] [2],

$$T(\mathbf{x},\mathbf{l};k) = \frac{\exp(ik\|\mathbf{x}-\mathbf{l}\|)}{4\pi\|\mathbf{x}-\mathbf{l}\|}, \tag{4.4}$$

and summing to give the superposition will result in the actual speech soundfield,

$$P^{(\mathrm{sp})}(\mathbf{x};a,k) = \sum_{l\in[\![L]\!]} Q_l(a,k)\,T(\mathbf{x},\mathbf{l}_l;k), \tag{4.5}$$

where $Q_l(a,k)$ is the time-frequency domain transform of $q_l(n)$ and $Q_l(k)$ is the frequency domain transform of $q_l(n)$.

The sound pressure in the time domain for each frame can be observed as

$$p(\mathbf{x};a,n) = \mathrm{Re}\left\{K^{-1}\sum_{m\in[\![K]\!]} P^{(\cdot)}(\mathbf{x};a,k_m)\exp\left(\frac{icnk_m}{2\hat{f}}\right)\right\}, \tag{4.6}$$

where $\mathrm{Re}\{\cdot\}$ returns the real part of its argument, $P^{(\cdot)}$ is any given soundfield function, $k_m \triangleq 2\pi\hat{f}m/cK$ and $\hat{f}$ is the maximum frequency. Performing overlap add on $p(\mathbf{x};a,n)$ then results in the pressure signal $p(\mathbf{x};n)$.

[2] Note that recomputing $W_l(k)$ for each frame, a, is required for moving virtual sources and/or zones.

[3] 2-D system models have also been shown to provide reasonable acoustic contrast in real-world environments [112]. 2-D ATFs are given by $T_{\mathrm{2D}}(\mathbf{x},\mathbf{l};k) = \frac{i}{4}\mathcal{H}_0^{(1)}(k\|\mathbf{x}-\mathbf{l}\|)$ [2].

The soundfield can now be evaluated at any given point in the reproduction region for different input signals and $p(\mathbf{x}; n)$ can be observed in the bright zone and quiet zone in order to estimate the behaviour of the system. From (4.6) it is possible to analyse the speech intelligibility and quality in different zones in order to control the soundfield reproduction as described in the next section.

4.3 Speech Privacy and Intelligibility Contrast

This section discusses the relationship between speech privacy and intelligibility, and proposes the Speech Intelligibility Contrast (SIC) measure for improving privacy in personal sound zones. Two optimisations are provided as methods to control multizone soundfield reproductions to improve speech privacy where the latter of the described methods also yields quality control in reproductions.

4.3.1 The Speech Intelligibility Contrast (SIC)

It is noted that the relation between speech privacy and intelligibility is highly correlated. Two different privacy measures, the Speech Privacy Class (SPC) for closed spaces and the Articulation Index (AI) for open plan spaces, are published as standards ASTM E2638 [211] and ASTM E1130 [210], respectively. The SPC has been shown to be a good measure for high privacy scenarios [209] and with the two standard measures (SPC and AI) highly correlated to speech intelligibility, it is reasonable to maximise an intelligibility contrast measure to obtain speech privacy. A measure of intelligibility contrast has the benefit, over SPC and AI, of providing accurate estimations of speech privacy in different scenarios, such as reverberant rooms [223] and with time-frequency weighted noisy speech [199].

The basis of many objective intelligibility measures is an analysis of spectral band powers which have been shown to be highly correlated with subjective measures. A clean speech (talker) signal, $y_T(n)$, and a degraded speech (listener) signal, $y_L(n)$, with a high signal to noise ratio (SNR) will also attain high mutual information [188]. In this work, $\mathcal{I}_{\mathcal{M}}(y_L; y_T)$ is used to denote the intelligibility for two signals, $y_T(n)$ and $y_L(n)$. A proxy of the mutual information, such as that provided by the Short-Time Objective Intelligibility (STOI) [199] or Speech Transmission Index (STI) [223], is denoted by the measure, $\mathcal{M}$. The soundfield intelligibility, $\mathcal{I}_{\mathcal{M}}(p(\mathbf{x};\cdot); y) \in \{0, \dots, 1\} \subsetneq \mathbb{R}$, of a signal, $y(n)$, at some spatial point, $\mathbf{x} \in \mathbb{D}$, is measured using the pressure signal, $p(\mathbf{x}; n)$.

We define the SIC as

$$\mathrm{SIC}_{\mathcal{M}} = \mathfrak{d}_{\mathrm{b}}^{-1} \int_{\mathbb{D}_{\mathrm{b}}} \mathcal{I}_{\mathcal{M}}(p(\mathbf{x};\cdot); y)\, d\mathbf{x} - \mathfrak{d}_{\mathrm{q}}^{-1} \int_{\mathbb{D}_{\mathrm{q}}} \mathcal{I}_{\mathcal{M}}(p(\mathbf{x};\cdot); y)\, d\mathbf{x}, \tag{4.7}$$

where $\mathfrak{d}_{\mathrm{b}} \triangleq \int_{\mathbb{D}_{\mathrm{b}}} 1\, d\mathbf{x}$ and $\mathfrak{d}_{\mathrm{q}} \triangleq \int_{\mathbb{D}_{\mathrm{q}}} 1\, d\mathbf{x}$ are the areas (sizes) of $\mathbb{D}_{\mathrm{b}}$ and $\mathbb{D}_{\mathrm{q}}$, respectively, and $\mathrm{SIC}_{\mathcal{M}}$ has a restricted domain such that $\mathcal{I}_{\mathcal{M}}, \forall \mathbf{x} \in \mathbb{D}_{\mathrm{b}}$ is greater than or equal to $\mathcal{I}_{\mathcal{M}}, \forall \mathbf{x} \in \mathbb{D}_{\mathrm{q}}$. The following subsection provides two methods to maximise $\mathrm{SIC}_{\mathcal{M}}$.

4.3.2 Privacy and Quality Control

To maximise the SIC, $\mathcal{I}_{\mathcal{M}}$ must be maximised at all points in $\mathbb{D}_{\mathrm{b}}$ whilst maintaining a minimum valued $\mathcal{I}_{\mathcal{M}}, \forall \mathbf{x} \in \mathbb{D}_{\mathrm{q}}$. In general, the higher the mean SNR of $p(\mathbf{x}; n)$ over $\mathbb{D}_{\mathrm{b}}$ the better, so reducing the mean SNR of $p(\mathbf{x}; n)$ over $\mathbb{D}_{\mathrm{q}}$ naturally becomes the criteria to increase $\mathrm{SIC}_{\mathcal{M}}$. To maximise the SIC, noise is added to the arbitrary loudspeaker signals, $q_l(n)$, that are used to reproduce $p(\mathbf{x}; n)$. It is assumed that $q_l(n)$ are designed to reproduce a mean amplitude of $p(\mathbf{x}; n)$ over $\mathbb{D}_{\mathrm{b}}$ greater than that of $p(\mathbf{x}; n)$ over $\mathbb{D}_{\mathrm{q}}$. A constrained optimisation is then formulated which is dependent on the reproduced signals in the quiet and bright zones as

$$\arg\max_{G} \ \mathrm{SIC}_{\mathcal{M}}, \ \text{subject to:}\, G \in \mathbb{R}^{+}, \tag{4.8}$$

where the optimal noise levels, G, of $q_l(n)$ are found.

A private personal sound zone system would ideally support high perceptual quality in the bright zone whilst preserving maximum SIC. A trade-off between privacy and target quality is apparent when adjusting G of $q_l(n)$, as doing so reduces the quality of $p(\mathbf{x};n), \forall \mathbf{x} \in \mathbb{D}_\mathrm{b}$ due to the addition of error to $p(\mathbf{x};n), \forall \mathbf{x} \in \mathbb{D}$. Using a similar notation to $\mathcal{I}_\mathcal{M}$, the quality of $p(\mathbf{x};n), \forall \mathbf{x} \in \mathbb{D}_\mathrm{b}$ (the reproduction of $y(n)$) is any speech quality assessment model, $\mathcal{B}_{\acute{\mathcal{M}}}(p(\mathbf{x};\cdot);y) \in \{0,\dots,1\} \subsetneq \mathbb{R}$, for a given measure, $\acute{\mathcal{M}}$, which is scaled to match that of $\mathcal{I}_\mathcal{M}$.

Now a new optimisation can be defined as

$$\begin{aligned} \underset{G}{\arg\max} \quad & \left(\mathrm{SIC}_\mathcal{M} + \frac{\lambda}{\eth_\mathrm{b}} \int_{\mathbb{D}_\mathrm{b}} \mathcal{B}_{\acute{\mathcal{M}}} \, d\mathbf{x} \right), \\ \text{subject to:} \quad & G \in \mathbb{R}^+, \\ & \mathcal{I}_\mathcal{M} \geq \mathcal{B}_{\acute{\mathcal{M}}}, \; \forall \mathbf{x} \in \mathbb{D}_\mathrm{b}, \end{aligned} \tag{4.9}$$

where the optimal noise levels, G, are defined in section 4.4 and the importance of quality in the optimisation is controlled with the weighting parameter, $\lambda \in \mathbb{R}^+$.

The multi-stage process proposed in this chapter aims to optimally choose the value of G to satisfy (4.9) whilst also constraining the amount of energy leaked between zones and meeting constraints due to spatial aliasing resulting from the use of a limited number of loudspeakers. The next section describes the spatial and spectral sound masker design approaches proposed in this work.

4.4 Spatial and Spectral Sound Masking

In this section, a method for improving speech privacy between spatial zones in multizone soundfield reproduction scenarios is described. The intelligibility between $y_T(n)$ and $y_L(n)$ can be reduced by reducing the ratio of the leaked pressure to the reproduced masker (i.e. the SNR) as described in section 4.3.2. The optimisations, also formulated in section 4.3.2, are realised by using spatially and spectrally weighted noise maskers. Spatial filters are defined using the multizone soundfield reproduction approach and vary depending on the target multizone speech soundfield, loudspeaker layout and zone geometry. Spectral shaping is described in the form of weighted predicted acoustic contrast ratios which are also dependent on the multizone reproduction of the target speech.

4.4.1 Spatial Sound Masking

To optimise the criteria in (4.9) a maximum mean SNR of $p(\mathbf{x}; n)$ over $\mathbb{D}_{\mathrm{b}}$ and minimum mean SNR of $p(\mathbf{x}; n)$ over $\mathbb{D}_{\mathrm{q}}$, is required. To achieve this, a time-domain Gaussian noise mask, $u(n)$, is projected into the spatial domain over $\mathbb{D}$ such that its reproduction becomes a multizone soundfield reproduction scenario. In this work, constraints are applied to the multizone reproduction of $u(n)$, which is quiet in $\mathbb{D}_{\mathrm{b}}$ and a plane-wave field in $\mathbb{D}_{\mathrm{q}}$, in order to simplify the optimisation of (4.8) and (4.9). The constraints are

$$\vartheta = \cos^{-1}\left(\frac{\overrightarrow{\mathbf{pq}} \cdot \hat{\mathbf{u}}_{\mathrm{o}}}{\|\overrightarrow{\mathbf{pq}}\|}\right), \tag{4.10}$$

so that the masker source is collocated with the leakage of the target bright zone soundfield reproduction (see section 4.5.2 for definitions of $\overrightarrow{\mathbf{pq}}$ and $\hat{\mathbf{u}}_{\mathrm{o}}$), and a new weighting function, $\hat{w}(\mathbf{x})$, is constrained to an importance of 0.05, 1 and 100 in $\mathbb{D}_{\mathrm{u}}$, $\mathbb{D}_{\mathrm{q}}$ and $\mathbb{D}_{\mathrm{b}}$, respectively [109]. The collocation of the masker source with the leakage is arranged such that the direction of propagation of the masker and the leakage are the same in order to provide the most effective spatial masking. The remainder of the multizone reproduction is the same as used to generate $Q_l(a, k)$ for the speech signal.

The goal is to solve (4.8) and (4.9) or, equivalently, to control the mean SNR of $p(\mathbf{x}; n)$ over $\mathbb{D}_{\mathrm{q}}$ by finding another set of loudspeaker signals that would reproduce $u(n)$ in $\mathbb{D}_{\mathrm{q}}$ only. To do this, $u(n)$ is transformed to the frequency domain, framed as $U(a, k)$ and used in replacement of the input signal, $Y(a, k)$, in (4.3) to give

$$\widehat{Q}_l(a, k) = \widehat{W}_l(k)\, U(a, k), \tag{4.11}$$

where the masker loudspeaker signals, $\widehat{Q}_l(a, k)$, are found after new loudspeaker weights are derived from (6.4) as $\widehat{W}_l(k)$. Superposition gives the resulting masker soundfield as

$$P^{(\mathrm{m})}(\mathbf{x}; a, k) = \sum_{l \in [\![L]\!]} \widehat{Q}_l(a, k)\, T(\mathbf{x}, \mathbf{l}_l; k). \tag{4.12}$$

The masker soundfield is then added to the speech soundfield

$$P^{(\mathrm{sp,m})}(\mathbf{x};a,k) = P^{(\mathrm{sp})}(\mathbf{x};a,k) + \bar{G}P^{(\mathrm{m})}(\mathbf{x};a,k) \tag{4.13}$$

$$= \sum_{l\in[\![L]\!]} Q'_l(a,k)\, T(\mathbf{x},\mathbf{l}_l;k), \tag{4.14}$$

where $Q'_l(a,k)$ are the new loudspeaker signals and $\bar{G}$ is the relative gain adjustment given by the root mean square (RMS) value from all L loudspeaker signals,

$$\bar{G} \triangleq \frac{G}{K}\left(\frac{1}{L}\sum_{l\in[\![L]\!],m\in[\![K]\!]} \left|Q_l(k_m)\right|^2\right)^{1/2}. \tag{4.15}$$

Then, $\mathrm{SIC}_{\mathcal{M}}$ is obtained from (4.7) after $p(\mathbf{x};n)$ is found from (4.6) using $P^{(\mathrm{sp,m})}(\mathbf{x};a,k)$. Now $\mathrm{SIC}_{\mathcal{M}}$ can be used to optimise G from (4.15) through (4.13) with (4.8). Alternatively, though, similarly, $\mathrm{SIC}_{\mathcal{M}}$ and $\mathcal{B}_{\acute{\mathcal{M}}}$ can be used to optimise G with (4.9). The optimisation problem can now be analysed by measuring $\mathcal{I}_{\mathcal{M}}$ for $\mathbf{x}\in\mathbb{D}_\mathrm{b}\cap\mathbb{D}_\mathrm{q}$, $\mathcal{B}_{\acute{\mathcal{M}}}$ for $\mathbf{x}\in\mathbb{D}_\mathrm{b}$, $\mathrm{SIC}_{\mathcal{M}}$ and for various $G\in\mathbb{R}^+$.

4.4.2 Long Term Average Speech Spectrum

The average magnitude spectrum of speech has been well documented and is known as the Long-Term Average Speech Spectrum (LTASS) [224], [225]. In order to accurately mask the speech that is leaked into the quiet zone, the spectrum of the masker should closely match the spectrum of the leakage. At any measurement point in a speech soundfield the spectral shape will, on average, consist of the speech magnitude spectrum and spectral shaping caused by the system response. Speech Shaped Noise (SSN) is an appropriate masking signal for the speech component of leaked content. To obtain SSN, framed Guassian noise is shaped to the LTASS as $U^{(\mathrm{sp})}(a,k)$ where $^{(\mathrm{sp})}$ denotes filtering for the speech spectrum. The magnitude response of the LTASS filter, $H^{(\mathrm{sp})}(k)$, can be approximated by either table 2 of [224], table 1 of [225] or by finding the mean sound pressure level (SPL) for a set of speech samples, e.g.,

$$\left|H^{(\mathrm{sp})}(k)\right|^2 = \frac{2}{BN^2} \sum_{b\in[\![B]\!]} \left| \sum_{n\in[\![N]\!]} h_b^{(\mathrm{sp})}(n) \exp\left(\frac{-icnk}{2\hat{f}}\right) \right|^2 \tag{4.16}$$

where $h_b^{(\mathrm{sp})}(n) \in \mathbb{R}$ is the bth non-overlapping frame from the sequence of $\widehat{N}$ speech samples, B is the number of frames and $B = \lceil \widehat{N}/N \rceil$. The SSN, $U^{(\mathrm{sp})}(a,k)$, can then be used in (4.11) to obtain $Q_l'(a,k)$ from (4.14) via (4.13) and (4.12).

4.4.3 A Priori Reproduction Spectrum Estimation

Even though the multizone reproduction system aims to match the desired input signal spectrum in the bright zone it does not guarantee that the quiet zone spectrum that is leaked remains the same shape. In fact, the spectrum of the quiet zone will vary significantly depending on many factors, such as the geometrical positioning of zones, virtual sources and secondary sources, and the type of reproduction technique used.

It is possible, however, to form an *a priori* estimate of the leaked spectrum by either knowing or estimating the inverse of the underlying acoustic contrast in the system. The inverted acoustic contrast can be found by either the ratio of energies between zones or by assuming a uniform (temporal) frequency spectrum in the bright zone. The system magnitude response in $\mathbb{D}_\mathrm{q}$ can be estimated using the soundfield $P^{(\mathrm{sp})}(\mathbf{x};a,k)$ reproduced from $Q_l(a,k)$, as

$$\left|H^{(\mathrm{q})}(k)\right| = \frac{1}{A} \sum_{a\in[\![A]\!]} \left(\frac{\mathfrak{d}_\mathrm{b} \int_{\mathbb{D}_\mathrm{q}} \left|P^{(\mathrm{sp})}(\mathbf{x};a,k)\right| d\mathbf{x}}{\mathfrak{d}_\mathrm{q} \int_{\mathbb{D}_\mathrm{b}} \left|P^{(\mathrm{sp})}(\mathbf{x};a,k)\right| d\mathbf{x}} \right)^{1/2}, \tag{4.17}$$

where $^{(\mathrm{q})}$ denotes a filter for the leaked quiet zone spectrum.

In practical reproductions it may be unnecessary to shape the noise spectrum above some aliasing frequency, k_u, as the leakage would boost high frequencies which can be seen later in Fig. 4.6. A more practical filter can be approximated as,

$$\left|H^{(\mathrm{q}')}(k)\right| = \begin{cases} \left|H^{(\mathrm{q})}(k)\right|, & k < k_{\mathrm{u}} \\ \left|H^{(\mathrm{q})}(k_{\mathrm{u}})\right|, & k \geq k_{\mathrm{u}} \end{cases}, \tag{4.18}$$

which ensures no shaping above k_{u}.

The leakage spectrum filter, $H^{(\mathrm{q}')}(k)$, can be used alongside the LTASS filter from (4.16) to obtain a good approximation of the leaked speech spectrum. The Gaussian noise, $U^{(\mathrm{sp},\mathrm{q}')}(a,k)$, shaped to $H^{(\mathrm{sp})}(k)$ and $H^{(\mathrm{q}')}(k)$, then matches accurately the leaked speech in the quiet zone up to the aliasing frequency and can then be used in (4.11) to obtain $Q_l'(a,k)$ from (4.14) via (4.13) and (4.12).

4.4.4 Secondary Leakage

Leakage between zones is a feature of multizone reproductions regardless of the target reproduction signal. When reproducing a multizone masking soundfield which matches the leaked speech in the target quiet zone there will also be leakage of the masker back into the target bright zone, we term this the *secondary leakage.* The shape of the secondary leakage may detrimentally influence both $\mathrm{SIC}_{\mathcal{M}}$ and $\mathcal{B}_{\acute{\mathcal{M}}}$ which shows the importance of the masker spectrum in the optimisation of (4.9). Ideally, a spectrum which influences both $\mathrm{SIC}_{\mathcal{M}}$ and $\mathcal{B}_{\acute{\mathcal{M}}}$ to equal extent, or to satisfy (4.9), is needed.

In this work we propose the use of a secondary leakage filter, $H^{(\mathrm{b})}(k)$, to determine a masker spectrum which has equal influence on $\mathrm{SIC}_{\mathcal{M}}$ and $\mathcal{B}_{\acute{\mathcal{M}}}$. As seen from the target quiet zone, the leaked spectrum back into the target bright zone is estimated using the soundfield $P^{(\mathrm{m})}(\mathbf{x};a,k)$ reproduced from $\widehat{Q}_l(a,k)$, and the secondary leakage spectrum is found as

$$\left|H^{(\mathrm{b})}(k)\right| = \frac{1}{A} \sum_{a \in [\![A]\!]} \left(\frac{\mathfrak{d}_{\mathrm{q}} \int_{\mathbb{D}_{\mathrm{b}}} \left|P^{(\mathrm{m})}(\mathbf{x};a,k)\right| d\mathbf{x}}{\mathfrak{d}_{\mathrm{b}} \int_{\mathbb{D}_{\mathrm{q}}} \left|P^{(\mathrm{m})}(\mathbf{x};a,k)\right| d\mathbf{x}} \right)^{1/2}, \tag{4.19}$$

where $^{(\mathrm{b})}$ denotes a filter for the secondary leakage spectrum.

Following the same reasoning for (4.18), the secondary leakage filter that ensures no shaping above k_u is

$$\left|H^{(\mathrm{b}')}(k)\right| = \begin{cases} \left|H^{(\mathrm{b})}(k)\right|, & k < k_\mathrm{u} \\ \left|H^{(\mathrm{b})}(k_\mathrm{u})\right|, & k \geq k_\mathrm{u} \end{cases}, \tag{4.20}$$

which is used to obtain the masker spectrum which has controllable influence on intelligibility and quality as

$$\begin{aligned} \left|H^{(\mathcal{IB})}(k)\right| &= \exp\left((1-\grave{\lambda}) \ln\left(\left|H^{(\mathrm{sp})}(k)\right| \left|H^{(\mathrm{q}')}(k)\right| \right) + \grave{\lambda} \ln\left(\frac{\left|H^{(\mathrm{sp})}(k)\right|}{\left|H^{(\mathrm{b}')}(k)\right|} \right) \right) \\ &= \left|H^{(\mathrm{sp})}(k)\right| \frac{\left|H^{(\mathrm{q}')}(k)\right|^{1-\grave{\lambda}}}{\left|H^{(\mathrm{b}')}(k)\right|^{\grave{\lambda}}} . \end{aligned} \tag{4.21}$$

It is worth noting that $\grave{\lambda} = 0$ results in $|H^{(\mathcal{IB})}(k)| = |H^{(\mathrm{sp})}(k)||H^{(\mathrm{q}')}(k)|$ and when $\grave{\lambda} = 1$ the result is $|H^{(\mathcal{IB})}(k)| = |H^{(\mathrm{sp})}(k)|/|H^{(\mathrm{b}')}(k)|$. The influence of the spectrum on intelligibility over quality can be controlled with the parameter $\grave{\lambda} \in \{0, \ldots, 1\} \subsetneq \mathbb{R}$, unlike λ, which does not control the shape of the spectrum.

The spectral maskers in this section have been derived for a single target speech signal. The methods are also applicable for cases where separate speech signals in each zone are desired, however, because the leaked speech between zones is not controlled, further reductions in quality may occur. Methods for controlling the leaked spectrum between zones, which may then improve quality, have been proposed in chapter 3.

4.5 Reducing Loudspeakers and Aliasing

A fundamental issue with wideband soundfield synthesis is the high number of secondary sources required for alias free reproduction of speech or music. In this section the consequent effect of aliasing on multizone soundfields is described and an analytical approach to reduce the effect is presented.

4.5.1 Grating Lobe Motivated Masker Filtering

For a sound zoning system to remain practical it should be possible for a small number of loudspeakers to provide high SIC and quality. A fundamental problem with the reduction in the number of loudspeakers is spatial aliasing which gives rise to grating lobes (the aliasing lobes that replicate the energy of the main lobe) capable of impeding the different zones and cannot be spatially controlled with soundfield synthesis.

Since filtering the target bright zone signal will knowingly alter the quality of the reproduced content it is sensible to shape only the portion of the (temporal) frequency spectrum of the masker signal without spatial aliasing artefacts. If the masker signal is dominant at frequencies where its grating lobes directly impede the target bright zone then the quality will be significantly reduced. Band-limiting the masker signal, $u(n)$, by applying a low-pass, denoted by $^{(\mathrm{lp})}$, filter, $H^{(\mathrm{lp})}(k)$, with a cutoff frequency of k_u (some aliasing frequency) will eliminate this effect, however, the masker signal will then not be able to mask speech in the stopband. Any low-pass filter can be used, for instance, a Chebyshev Type I [226] is

$$\left|H^{(\mathrm{lp})}(k)\right| = \left(1 + \left(\varepsilon\, \mathcal{T}_{\breve{n}}(k/k_\mathrm{u})\right)^2\right)^{-1/2}, \tag{4.22}$$

where $\mathcal{T}_{\breve{n}}(\cdot)$ is a Chebyshev polynomial [226] of the first kind with order $\breve{n}$ and ε is the maximum allowable passband ripple. The noise signal, $u(n)$, is filtered with $H^{(\mathrm{lp})}(k)$ to obtain $U^{(\mathrm{lp})}(a, k)$.

Fortunately, the frequency spectrum of speech is dominant at lower frequencies [224], [225] and so the majority of information leaked can still be masked effectively from the low-pass filtered masker signal. To perform the spatially weighted masking, (4.11) is used with noise signal $U^{(\mathrm{lp})}(a, k)$ and $Q'_l(a, k)$ is found from (4.14) via (4.13) and (4.12).

4.5.2 Grating Lobe Prediction

The grating lobes can be accurately predicted if the loudspeaker array and zone geometry is known. The next two sub-subsections provide an analytical approach to finding the frequency where grating lobes touch the quiet zone for both circular and linear loudspeaker arrays.

Circular Array Grating Lobes

For a maximum mode order of $\overline{M}' = \lceil kR' \rceil$ [68], [109], where R' is the radius of the smallest circle (concentric with $\mathbb{D}$) encompassing all zones, and by using the part circle method [109], [143], it is possible to formulate an approximation for the upper frequency limit, k_u, at which aliasing will begin to occur. The minimum number of required loudspeakers is given by [106], [109]

$$L \geq \left\lceil \frac{\phi_\mathrm{L}\left(2\overline{M}+1\right)}{2\pi} \right\rceil + 1, \tag{4.23}$$

substituting the truncation length, $\overline{M}'$, and rearranging gives

$$\hat{k}_\mathrm{u} = \frac{2\pi(L-1) - \phi_\mathrm{L}}{2R'\phi_\mathrm{L}}, \tag{4.24}$$

however, this provides the frequency where the centre of the grating lobe is at least R' from the centre of the reproduction, not accounting for zone positions. In many cases it is possible to use a frequency higher than $\hat{k}_\mathrm{u}$ where the grating lobes do not travel through the quiet zone. That is to say, aliasing artifacts can be tolerated in $\mathbb{D}_\mathrm{u}$ depending on relative locations of the zones thus redefining the aliasing to that occurring in $\mathbb{D}_\mathrm{q}$, not $\mathbb{D}$. The aim is to find a new $\hat{k}_\mathrm{u}$ by deriving a replacement for $2R'$. To aid the derivations, Figure 4.2 shows a circular array with auxiliary values.

Here, the work in [221] is extended to the multizone reproduction scenario to define a zone based limit for the grating lobe. Similar to the work in [221], a point,

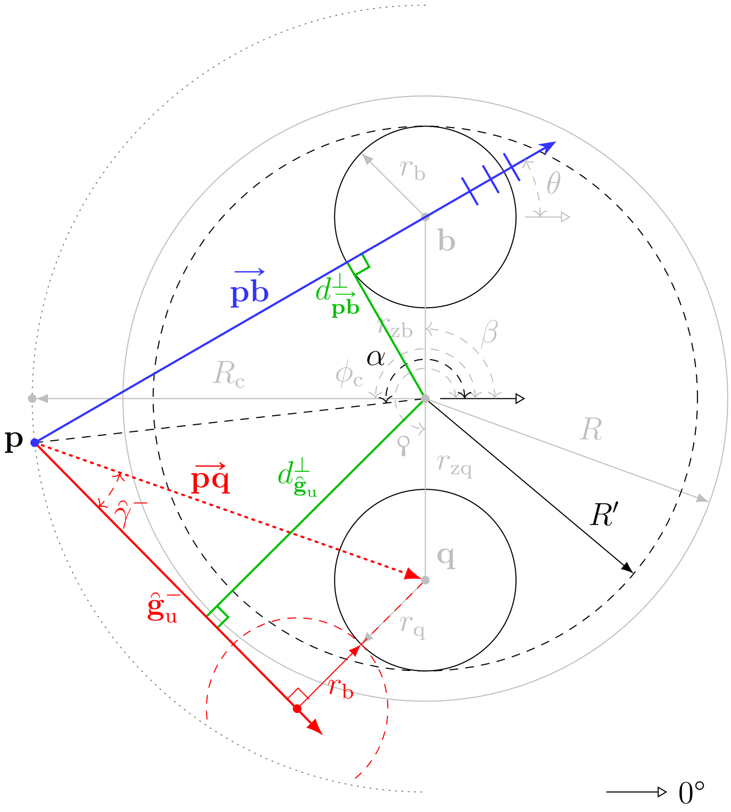

Figure 4.2: Auxiliary entities of a circular array multizone soundfield reproduction layout. The plane-wave vector ($\overrightarrow{\mathbf{pb}}$) is blue, the grating lobe limit ($\hat{\mathbf{g}}_u^{-}$ is shown) found using (4.32) is red and the frequency limit ($\hat{k}_u'$) is computed with (4.35) using the perpendicular distances ($d^{\perp}_{\hat{\mathbf{g}}_u}$ and $d^{\perp}_{\overrightarrow{\mathbf{pb}}}$) that are shown in green.

$\mathbf{p}$, is positioned on the loudspeaker arc at distance R_c and with angle

$$\alpha = \theta - \sin^{-1}\left(\frac{d^{\perp}_{\overrightarrow{\mathbf{pb}}}}{R_\text{c}}\right) + \pi, \tag{4.25}$$

where $\mathbf{p}$ is the origin for grating lobes as shown in Figure 4.2 and

$$d^{\perp}_{\overrightarrow{\mathbf{pb}}} = \Big| r_\text{zb} \sin(\beta - \theta) \Big|. \tag{4.26}$$

The first spectral repetition of grating lobes have a width equal to the bright zone diameter [221]. The outer-most tangent from the origin of a circle of radius $r_\text{b} + r_\text{q}$ at $\mathbf{q}$ which intersects $\mathbf{p}$, corresponds to the centre of a grating lobe whose edge touches $\mathbb{D}_\text{q}$. Vector notation is used when finding the tangent and hence the newly defined aliasing frequency.

The rotated grating lobe vector, $\overrightarrow{\mathbf{pq}}$, points from the circular array grating lobe origin, $\hat{\mathbf{p}}$ (at angle α and radius R_c), to the quiet zone origin, $\mathbf{q}$ (at angle $\qoppa$ and radius r_zq), and is given by

$$\hat{\mathbf{p}} = R_\text{c} \cdot \mathring{\mathbf{R}}(\alpha) \cdot \widehat{\mathbf{u}}_\text{o}, \tag{4.27}$$

$$\mathbf{q} = r_\text{zq} \cdot \mathring{\mathbf{R}}(\qoppa) \cdot \widehat{\mathbf{u}}_\text{o}, \tag{4.28}$$

$$\overrightarrow{\mathbf{pq}} = \mathbf{q} - \hat{\mathbf{p}}, \tag{4.29}$$

where $\widehat{\mathbf{u}}_\text{o}$ is a unit column vector at the origin and

$$\mathring{\mathbf{R}}(\delta) \triangleq \begin{bmatrix} \cos(\delta) & -\sin(\delta) & 0 \\ \sin(\delta) & \cos(\delta) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{4.30}$$

is a rotational matrix for a given angle δ.

The vector $\overrightarrow{\mathbf{pq}}$ can be rotated about $\mathbf{p}$ to equate the centre of the grating lobe. The maximum allowable angle of the grating lobe before impeding $\mathbb{D}_\text{q}$ is one of the two angles:

$$\hat{\gamma}^{\pm} = \pm \sin^{-1}\left(\frac{r_{\mathrm{b}} + r_{\mathrm{q}}}{\|\overrightarrow{\mathbf{pq}}\|}\right), \tag{4.31}$$

where $\|\cdot\|$ denotes the Euclidean norm. Therefore the grating lobe of the upper frequency limit due to aliasing is one of the two tangents[4]:

$$\hat{\mathbf{g}}_{\mathrm{u}}^{\pm} = \mathring{\mathbf{R}}\left(\hat{\gamma}^{\pm}\right) \cdot \overrightarrow{\mathbf{pq}}. \tag{4.32}$$

The perpendicular distance from $\hat{\mathbf{g}}_{\mathrm{u}}^{\pm}$ to the origin,

$$d^{\perp}_{\hat{\mathbf{g}}_{\mathrm{u}}^{\pm}} = \frac{\left|\hat{\mathbf{p}}^{\mathsf{T}} \cdot \left(\mathring{\mathbf{R}}\left(\frac{\pi}{2}\right) \cdot \hat{\mathbf{g}}_{\mathrm{u}}^{\pm}\right)\right|}{\left\|\hat{\mathbf{g}}_{\mathrm{u}}^{\pm}\right\|}, \tag{4.33}$$

where $\{\cdot\}^{\mathsf{T}}$ is a transposition of the vector, can be used to determine the correct tangent as

$$d^{\perp}_{\hat{\mathbf{g}}_{\mathrm{u}}} = \max\left(d^{\perp}_{\hat{\mathbf{g}}_{\mathrm{u}}^{+}}, d^{\perp}_{\hat{\mathbf{g}}_{\mathrm{u}}^{-}}\right). \tag{4.34}$$

The corresponding circular array aliasing frequency, $\hat{k}'_{\mathrm{u}}$, can then be found by replacing $2R'$ in (4.24) with $d^{\perp}_{\hat{\mathbf{g}}_{\mathrm{u}}} + d^{\perp}_{\overrightarrow{\mathbf{pb}}}$, as

$$\hat{k}'_{\mathrm{u}} = \max\left(\frac{2\pi(L-1) - \phi_{\mathrm{L}}}{\left(d^{\perp}_{\hat{\mathbf{g}}_{\mathrm{u}}} + d^{\perp}_{\overrightarrow{\mathbf{pb}}}\right)\phi_{\mathrm{L}}}, \hat{k}_{\mathrm{u}}\right). \tag{4.35}$$

Linear Array Grating Lobes

Similar to the derivation for a circular array, the linear array solution uses the tangents from the origin of the grating lobe to a circle of radius $r_{\mathrm{b}} + r_{\mathrm{q}}$ at point $\mathbf{q}$. Fig. 4.3 shows a linear array with auxiliary values. For a linear array the point of origin of the grating lobe, $\mathbf{p}$, is found from the intersection of the unit plane-wave vector and the loudspeaker array unit vector.

[4] The two tangents stem from the sign of (4.31) and are denoted by ±.

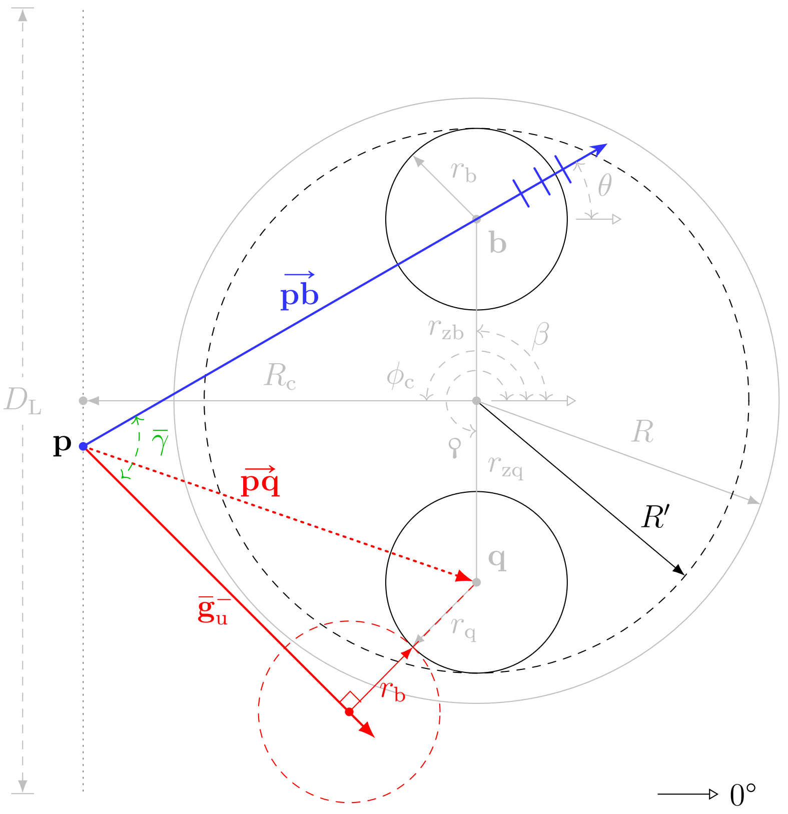

Figure 4.3: Auxiliary entities of a linear array multizone soundfield reproduction layout. The plane-wave vector ($\overrightarrow{\mathbf{pb}}$) is blue, the grating lobe limit ($\bar{\mathbf{g}}_{\mathrm{u}}^{-}$ is shown) found following section 4.5.2 is red and the frequency limit ($\bar{k}_{\mathrm{u}}$) is computed with (4.44) using the maximum allowable grating lobe angle ($\bar{\gamma}$) that is shown in green.

The centre point of $\mathbb{D}_\mathrm{b}$ and the loudspeaker array are

$$\mathbf{b} = r_\mathrm{zb} \cdot \mathring{\mathbf{R}}(\beta) \cdot \hat{\mathbf{u}}_\mathrm{o} \quad \text{and} \tag{4.36}$$

$$\mathbf{c} = R_\mathrm{c} \cdot \mathring{\mathbf{R}}(\phi_\mathrm{c}) \cdot \hat{\mathbf{u}}_\mathrm{o}, \tag{4.37}$$

respectively, and the solution for the intersection is

$$\mathbf{a}_1 = \begin{bmatrix} \cos(\theta) & \sin(\theta) & 0 \end{bmatrix}^\mathsf{T}, \tag{4.38}$$

$$\mathbf{a}_2 = \begin{bmatrix} \cos\Big(\phi_\mathrm{c} - \frac{\pi}{2}\Big) & \sin\Big(\phi_\mathrm{c} - \frac{\pi}{2}\Big) & 0 \end{bmatrix}^\mathsf{T}, \tag{4.39}$$

$$\begin{bmatrix} s_1 & s_2 \end{bmatrix}^\mathsf{T} = \begin{bmatrix} \mathbf{a}_1 & -\mathbf{a}_2 \end{bmatrix}^\dagger \cdot (\mathbf{c} - \mathbf{b}), \tag{4.40}$$

where † denotes a Moore-Penrose pseudoinverse. The intersecting point for the linear array is then

$$\bar{\mathbf{p}} = \mathbf{b} + s_1\mathbf{a}_1 = \mathbf{c} + s_2\mathbf{a}_2. \tag{4.41}$$

Inserting $\bar{\mathbf{p}}$ in replacement of $\hat{\mathbf{p}}$ in (4.29) yields a new $\overrightarrow{\mathbf{pq}}$ for a linear array and $\overrightarrow{\mathbf{pb}} = \mathbf{b} - \bar{\mathbf{p}}$.

Using $\overrightarrow{\mathbf{pq}}$ for a linear array in (4.31) and (4.32) yields $\bar{\mathbf{g}}_\mathrm{u}^\pm$ for a linear array. The maximum of the two angles between $\bar{\mathbf{g}}_\mathrm{u}^\pm$ and $\overrightarrow{\mathbf{pb}}$ gives the maximum allowable grating lobe angle for a linear array by

$$\psi^\pm = \cos^{-1}\left(\frac{\bar{\mathbf{g}}_\mathrm{u}^\pm \cdot \overrightarrow{\mathbf{pb}}}{\|\bar{\mathbf{g}}_\mathrm{u}^\pm\| \cdot \left\|\overrightarrow{\mathbf{pb}}\right\|} \right), \tag{4.42}$$

$$\bar{\gamma} = \max\Big(\psi^+, \psi^-\Big). \tag{4.43}$$

The linear array aliasing frequency [227, eq. (5.61)] is then

$$\bar{k}_\mathrm{u} = \frac{2\pi(L-1)}{D_\mathrm{L}(\sin(\bar{\gamma} - \Theta) + \sin(\Theta))}, \tag{4.44}$$

where $\Theta = |\pi - |\theta| - |\phi_\mathrm{c}||$, $\theta \in \{-\pi, \ldots, \pi\}$ and $\phi_\mathrm{c} \in \{-\pi, \ldots, \pi\}$.

The upper cut-off frequency from aliasing is k_u and equates to either $\hat{k}_\mathrm{u}'$ or $\bar{k}_\mathrm{u}$

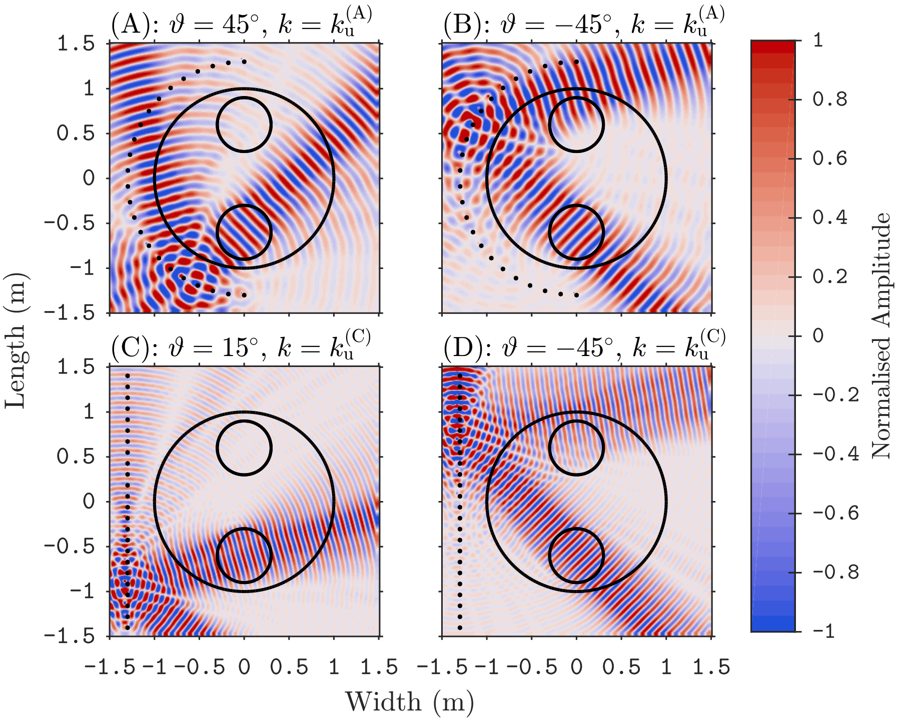

Figure 4.4: The real part of the masker soundfields at an aliasing frequency are shown to illustrate the impinging effect of grating lobes on $\mathbb{D}_{\mathrm{b}}$. The right column shows the masker grating lobe entering $\mathbb{D}_{\mathrm{b}}$ after ϑ is changed from the left column. Soundfields are shown for a semi-circular array and linear array in the top and bottom rows, respectively. The soundfield parameters are the same as those given in section 4.7.1 with $L = 24$. The angle of the wave front, ϑ, in $\mathbb{D}_{\mathrm{q}}$ is labelled for each plot and is chosen to best illustrate the interference in $\mathbb{D}_{\mathrm{b}}$. The superscript of k_{u} indicates which multizone setup in the figure was used to calculate k_{u}. The upper frequency limits are $k_{\mathrm{u}}^{(\mathrm{A})} = 35.6\,\mathrm{m}^{-1}$ and $k_{\mathrm{u}}^{(\mathrm{C})} = 59.6\,\mathrm{m}^{-1}$ corresponding to temporal frequencies of 1.94 kHz and 3.25 kHz, respectively.

depending on the loudspeaker array geometry. k_{u} can be used in a low-pass filter which can then be applied to any masker signal, $U(a, k)$, in (4.3).

4.5.3 Example Aliasing Artefacts

The impinging effect of the grating lobe into $\mathbb{D}_{\mathrm{b}}$ as ϑ is varied is shown in Figure 4.4. Other studies have investigated the effect of aliasing with differing numbers of loudspeakers [98]. For the cases in Figure 4.4 where k_{u} is found using the correct ϑ (i.e. in the first column) the mean energy in $\mathbb{D}_{\mathrm{b}}$ remains low. When the angle of the desired wave front in $\mathbb{D}_{\mathrm{q}}$ is moved, the energy in $\mathbb{D}_{\mathrm{b}}$ increases as the grating lobe traverses the zone which is undesirable and shows the importance of accurately computing k_{u}. Re-evaluating k_{u} for changes in ϑ will ensure the energy of the grating lobe in $\mathbb{D}_{\mathrm{b}}$ is kept low.

4.6 Reproduction Filtering

For a set of arbitrary magnitude responses, $\{|H^{(g)}(k)|\}_{g\in\mathbb{G}}$, where g denotes a particular filter, the complex symmetric linear-phase FIR filter can be found using a frequency-sampling method as

$$H^{(g)}(k) = \left|H^{(g)}(k)\right| \exp\left(\frac{-i\pi k}{2\Delta k}\right), \tag{4.45}$$

where $\Delta k \triangleq k_m/m$ is the wavenumber spacing. For the more general case where multiple magnitude responses are designed, their product will result in the complex cascaded filter bank

$$H^{(\mathbb{G})}(k) = \prod_{g\in\mathbb{G}} H^{(g)}(k), \tag{4.46}$$

where a window can be applied to the time transformed filter impulse response.

The arbitrary magnitude linear-phase FIR cascaded filter bank, $H^{(\mathbb{G})}(k)$, can now be applied to any system input signal, such as the speech, $Y(a,k)$, or the masker, $U(a,k)$, by frequency domain multiplication, e.g.,

$$U^{(\mathbb{G})}(a,k) = U(a,k)H^{(\mathbb{G})}(k), \tag{4.47}$$

which can then be used in (4.3) or (4.11) instead of $Y(a,k)$ or $U(a,k)$, respectively, to synthesise $Q_l'(a,k)$ from (4.14) via (4.13) and (4.12).

4.7 Results and Discussion

This section presents objective intelligibility results for the bright and quiet zones in anechoic reproduction environments and discusses the SIC and quality trade-off.

4.7.1 Experimental Setup

The geometrical layout of Fig. 4.1 is evaluated, where $r_{\mathrm{zb}} = r_{\mathrm{zq}} = 0.6\,\mathrm{m}$, $r_{\mathrm{b}} = r_{\mathrm{q}} = 0.3\,\mathrm{m}$, $R = 1.0\,\mathrm{m}$ and $\beta =$ ♀$/3 = 90\,°$. The loudspeaker arrays have $R_{\mathrm{c}} = 1.3\,\mathrm{m}$ and $\phi_{\mathrm{c}} = 180\,°$. The part circle loudspeaker array is an arc which subtends an angle of $\phi_{\mathrm{L}} = 180\,°$. The linear loudspeaker array has a length of $D_{\mathrm{L}} = (L-1)\Delta D_{\mathrm{L}}$ where $\Delta D_{\mathrm{L}} = 12.2\,\mathrm{cm}$ is the spacing between adjacent loudspeakers (designed to match Genelec 8010A loudspeakers). The values of $\theta = \{0\,°, 24.8\,°, 46.1\,°\}$ are used for the angle of the desired plane-wave virtual source in the bright zone for the part circle array and $\theta = \{0\,°, 24.8\,°, 42.7\,°\}$ are for the line array. Using (4.10), values of θ correspond to $\vartheta = \{-46.1\,°, -24.8\,°, 0\,°\}$ and $\vartheta = \{-42.7\,°, -24.8\,°, 0\,°\}$ for the part circle and line array, respectively. These angles are chosen such that the grating lobe for speech impedes $\mathbb{D}_{\mathrm{q}}$ at the same angle that the maskers grating lobe impedes $\mathbb{D}_{\mathrm{b}}$. The relationship is symmetrical about $\theta = 24.8°$ and three values are chosen.

A pseudo-random selection, constrained to have a male to female speaker ratio of $1 : 1$, was used to determine Twenty files from the TIMIT corpus [214] for the evaluation. Input speech signals with a sampling frequency of 16 kHz are framed with 50 % overlapping 64 ms windows and transformed using an FFT to the time-frequency domain. The loudspeaker signals, $Q_l'(a,k)$, are synthesised using the methods described in section 4.2 and section 4.4. The number of loudspeakers used for the simulated reproductions are $L = \{16, 24, 32, 114\}$ where, for the cases in this work, aliasing problems below 8 kHz are avoided in the reproduction using $L = 114$ for the semi-circular array [106], [109]. For the case when $L = 114$ for a linear array, $\Delta D_{\mathrm{L}} = 3.63\,\mathrm{cm}$ to prevent aliasing below 8 kHz and the speed of sound is $c = 343\,\mathrm{m\,s^{-1}}$. The noise masker gain levels, G, are varied ranging from $-40\,\mathrm{dB}$ to $20\,\mathrm{dB}$ in (4.15) for use in (4.13).

The anechoic reproductions are analysed with $\mathrm{SIC}_{\mathrm{STOI}}$ and $\mathcal{B}_{\mathrm{PESQ}}$ which evaluate the performance using the STOI [199] and Perceptual Evaluation of Speech Quality (PESQ) [177] measures, respectively. Thirty-two receivers are positioned randomly in each zone for recordings which are then analysed. Time-frequency weighted noisy speech, like the simulated recordings in this work, is well suited to the STOI measure. The PESQ measure is a good instrumental measure for quality of speech. The STOI and PESQ are measured in this work for each file and receiver combination using the clean, $y(n)$, and degraded, $p(\mathbf{x}; n)$, speech signals. A spatial average of the quality and intelligibility results over each zone is then performed following (4.7) and (4.9).

4.7.2 Soundfield Error and Planarity

The accuracy of the reproduced soundfield in $\mathbb{D}_\mathrm{b}$ is evaluated using the mean squared error (MSE) as defined in [112] and the planarity measure as defined in [98], [99]. Results for the MSE and planarity in the frequency domain are provided in Figure 4.5, where the target angle for the soundfield in the bright zone, θ, is varied from $-30°$ to $55°$. As the target bright zone angle is varied, the masker angle, ϑ, is computed using (4.10). The results show that the MSE in $\mathbb{D}_\mathrm{b}$ is consistently low below the aliasing frequency with an average error of $-30.3\,\mathrm{dB}$ for the semi-circular array and $-30.2\,\mathrm{dB}$ for the linear array. While the MSE increases above the aliasing frequency, it is still significantly low with an average of $-20.9\,\mathrm{dB}$ for the semi-circular array and $-24.0\,\mathrm{dB}$ for the linear array. It is also apparent that the planarity remains consistently high above the aliasing frequency, indicating that the shape of the wave front remains planar as the grating lobes impede $\mathbb{D}_\mathrm{b}$. The average planarity in $\mathbb{D}_\mathrm{b}$ above the aliasing frequency is 84.3% for the semi-circular array and 88.1% for the linear array. These results indicate that the spatial error is significantly low in the bright zone for a wide range of target bright zone angles when using the proposed methods.

4.7.3 Masker Filtering: Design and Comparison

The filters from (4.16), (4.18), (4.20) and (4.22) are $H^{(\mathrm{sp})}(k)$, $H^{(\mathrm{q}')}(k)$, $H^{(\mathrm{b}')}(k)$ and $H^{(\mathrm{lp})}(k)$, respectively, which are shown in Fig. 4.6 (A) along with the intermediate filters, $H^{(\mathrm{q})}(k)$ and $H^{(\mathrm{b})}(k)$, from (4.17) and (4.19), respectively. The LTASS is $H^{(\mathrm{sp})}(k)$, the leakage into $\mathbb{D}_\mathrm{q}$ is shaped by $H^{(\mathrm{q})}(k)$, the secondary leakage into

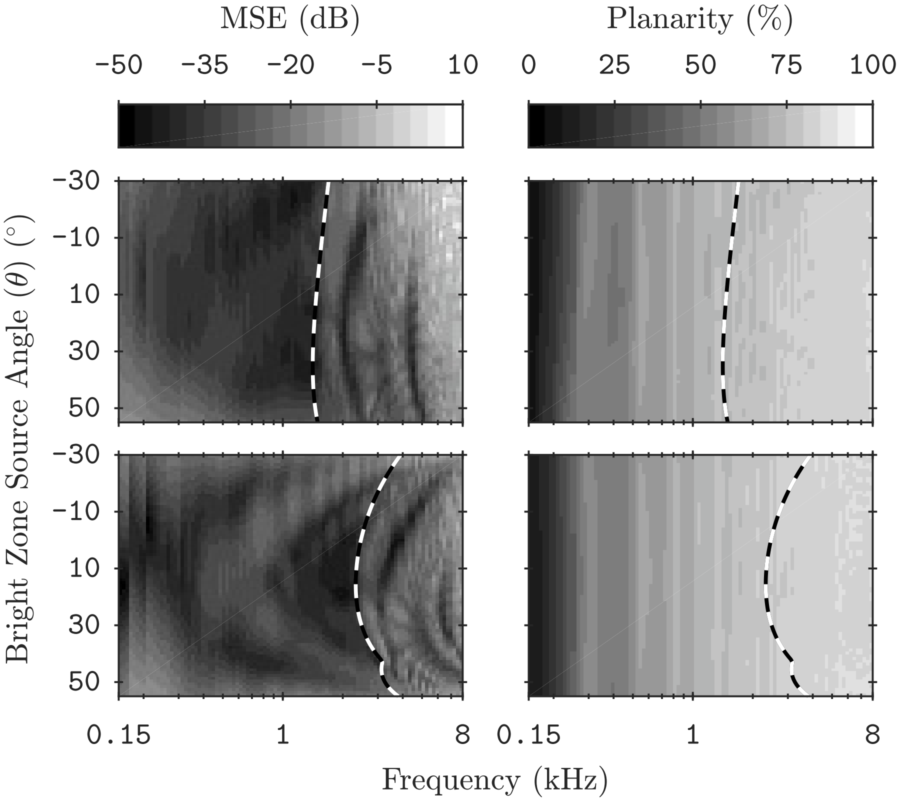

Figure 4.5: The MSE and planarity of $\mathbb{D}_\text{b}$ in the frequency domain are shown as θ is varied. Results for the semi-circular and linear loudspeaker array are given in the top and bottom rows, respectively, and the MSE and planarity are shown in the left and right column, respectively. As the target bright zone angle, θ, is varied the corresponding masker angle, ϑ, is found with (4.10). The number of loudspeakers is $L = 24$ and the masker gain is $G = -10\,\text{dB}$. The remainder of the setup is as described in section 4.7.1. The black and white dashed lines show the aliasing frequency as computed using the methods described in section 4.5.

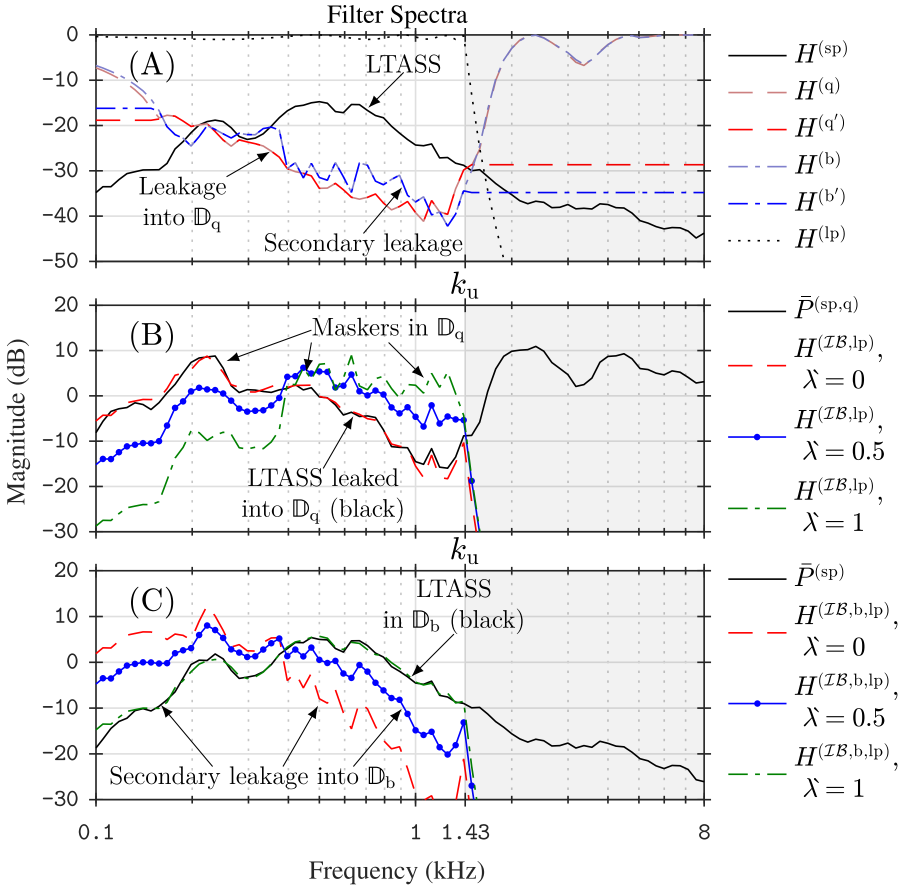

Figure 4.6: Example filter spectra are shown. Individual filter responses are displayed in A with comparisons to the average leaked pressure magnitude in $\mathbb{D}_q$ and $\mathbb{D}_b$ shown in B and C, respectively. Descriptive labels are provided for various spectra. Responses are averaged over 1/12th octave bands. The bandwidth of aliasing above k_u is shaded.

$\mathbb{D}_b$ is shaped by $H^{(b)}(k)$ and the low pass grating lobe filter is $H^{(lp)}(k)$. Using the experimental setup in section 4.7.1, a cascaded masker filter bank, $H^{(\mathbb{G})}(k)$, is obtained using (4.46) with $\mathbb{G} = \{\mathcal{IB}, \text{lp}\}$ and for $\lambda \in \{0.0, 0.5, 1.0\}$. Also shown is the spectrum of the proposed filtered masker in both $\mathbb{D}_q$ (Fig. 4.6 (B)) and in $\mathbb{D}_b$ (Fig. 4.6 (C)) for the various λ. The mean LTASS leaked over $\mathbb{D}_q$, denoted in this work as $\bar{P}^{(sp,q)}(k)$ and shown in Fig. 4.6 (B), is found using (4.6) and (4.16) with 32 virtual receivers and responses are averaged over the receiver positions. Similarly the mean LTASS over $\mathbb{D}_b$ is denoted as $\bar{P}^{(sp)}(k)$ and shown in Fig. 4.6 (C). It can be seen in Fig. 4.6 that $H^{(\mathbb{G})}(k)$ is a much closer match to the average leaked spectrum in $\mathbb{D}_q$ when $\lambda = 0.0$ and is closer to $\bar{P}^{(sp)}(k)$ when $\lambda = 1.0$. A trade-off between these two results is shown where $\lambda = 0.5$.

Table 4.1: Mean COSH distances, $\mathcal{E}^{(g)}_{\mathrm{COSH,z}}$, for different noise maskers and zones. Values are given in decibels and the smallest distance in each row is bold weight.

$\mathbb{G} =$	$\{\mathrm{wh}, \mathrm{lp}\}$	$\{\mathrm{p}, \mathrm{lp}\}$	$\{\mathcal{IB}, \mathrm{lp}\}$	$\{\mathcal{IB}, \mathrm{lp}\}$	$\{\mathcal{IB}, \mathrm{lp}\}$
			$\lambda = 0.0$	$\lambda = 0.5$	$\lambda = 1.0$
$\mathcal{E}^{(\mathbb{G})}_{\mathrm{COSH,b}}$	-6.02	3.7	2.58	-11.1	$\mathbf{-36.2}$
$\mathcal{E}^{(\mathbb{G})}_{\mathrm{COSH,q}}$	-7.21	-15.1	$\mathbf{-21.2}$	-9.99	2.4
Mean	-6.6	-1.38	-2.9	$\mathbf{-10.5}$	-3.52

To measure the accuracy of the filters with respect to the leaked spectrum to be masked, a symmetrical variant of the Itakura-Saito (IS) [228] distance is used, the hyperbolic cosine (COSH) spectral distance [229]. The COSH distance used in this work is given by

$$E^{(g)}_{\mathrm{COSH}}(\mathbf{x}) = K^{-1} \sum_{m \in [\![K]\!]} \left(\cosh \left(\ln \frac{\left| H^{(g)}(k_m) \right|}{\check{P}(\mathbf{x}; k_m)} \right) - 1 \right), \tag{4.48}$$

where $H^{(g)}(k)$ is the filter to be measured, $\check{P}(\mathbf{x}; k)$ is the pressure spectrum at $\mathbf{x}$ and $E^{(g)}_{\mathrm{COSH}}(\mathbf{x})$ is the COSH distance for all K frequencies. To evaluate the leaked spectrum, the mean COSH distance over some zone, $\mathbb{D}_{\mathrm{z}}$, of size $\mathfrak{d}_{\mathrm{z}}$ for $\mathrm{z} \in \{\mathrm{b}, \mathrm{q}\}$, is found as

$$\mathcal{E}^{(g)}_{\mathrm{COSH,z}} = \mathfrak{d}_{\mathrm{z}}^{-1} \int_{\mathbb{D}_{\mathrm{z}}} E^{(g)}_{\mathrm{COSH}}(\mathbf{x})\, d\mathbf{x}. \tag{4.49}$$

The values in Figure 4.7 and Table 4.1 given by (4.49) show that the proposed cascaded filter, $\{\mathcal{IB}, \mathrm{lp}\}$, provides a masker spectrum with the least mean distance to the spectrum of the speech in $\mathbb{D}_{\mathrm{b}}$ and leaked speech in $\mathbb{D}_{\mathrm{q}}$ with $\lambda = 0.5$ at -10.5 dB when compared to white noise ($\{\mathrm{wh}, \mathrm{lp}\}$), pink noise ($\{\mathrm{p}, \mathrm{lp}\}$), $\lambda = 0.0$ and $\lambda = 1.0$.

4.7.4 Speech Privacy Results

A descriptive comparison of the effectiveness and robustness of the methods outlined throughout this chapter is presented in this subsection. Results for instrumentally measured intelligibility and quality are given so the reader may intuitively interpret the relationships between noise masking, quality and privacy. The robustness of the

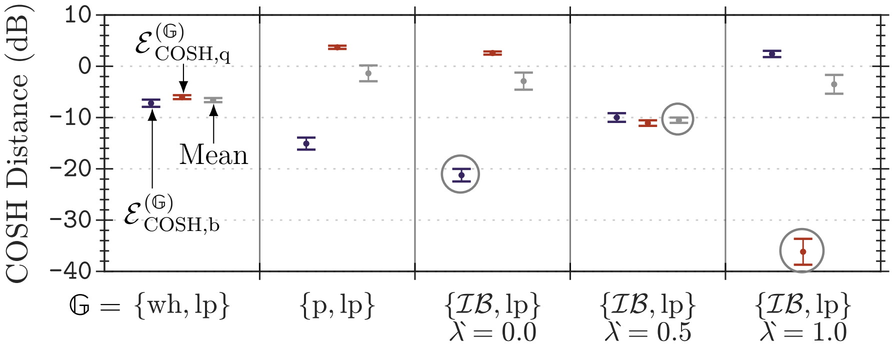

Figure 4.7: Mean COSH distances, $\mathcal{E}^{(g)}_{\mathrm{COSH,z}}$, for different noise maskers and zones are shown. The columns indicate the different noise maskers and each column contains three values which are $\mathcal{E}^{(\mathbb{G})}_{\mathrm{COSH,b}}$ (left), $\mathcal{E}^{(\mathbb{G})}_{\mathrm{COSH,q}}$ (middle) and the mean of both zones (right). COSH distance values are given in decibels and the smallest distance for each set is circled. The 95% confidence intervals shown are calculated over the area of each zone.

methods is conveyed through consistent results when varying the target bright zone virtual source angle, the array geometry and the number of available loudspeakers. The varying effectiveness of the methods is shown via results for different masking spectra and spectrum weighting parameters.

Figure 4.8, Figure 4.9 and Figure 4.12 all show results for the semi-circular and linear array in the left and right column, respectively. The figures all include variation in θ, microphone positions and speech in the 95% confidence intervals. Variation in the spectrum weight and shaping as determined by λ is shown along the rows of Figure 4.8 with white noise in the first row for comparison. Variation in the loudspeaker count, L, is shown along the rows of Figure 4.9. A discussion on the aforementioned variables is given in the following sub-subsections.

Angle

While consistently applying spatial weighting to all, or part, of the reproduction it is still natural for the acoustical brightness contrast performance to vary depending on θ. Figure 4.8 contains the variations due to the different θ in its confidence intervals which are still considerably small and show the method's robustness to variance in θ.

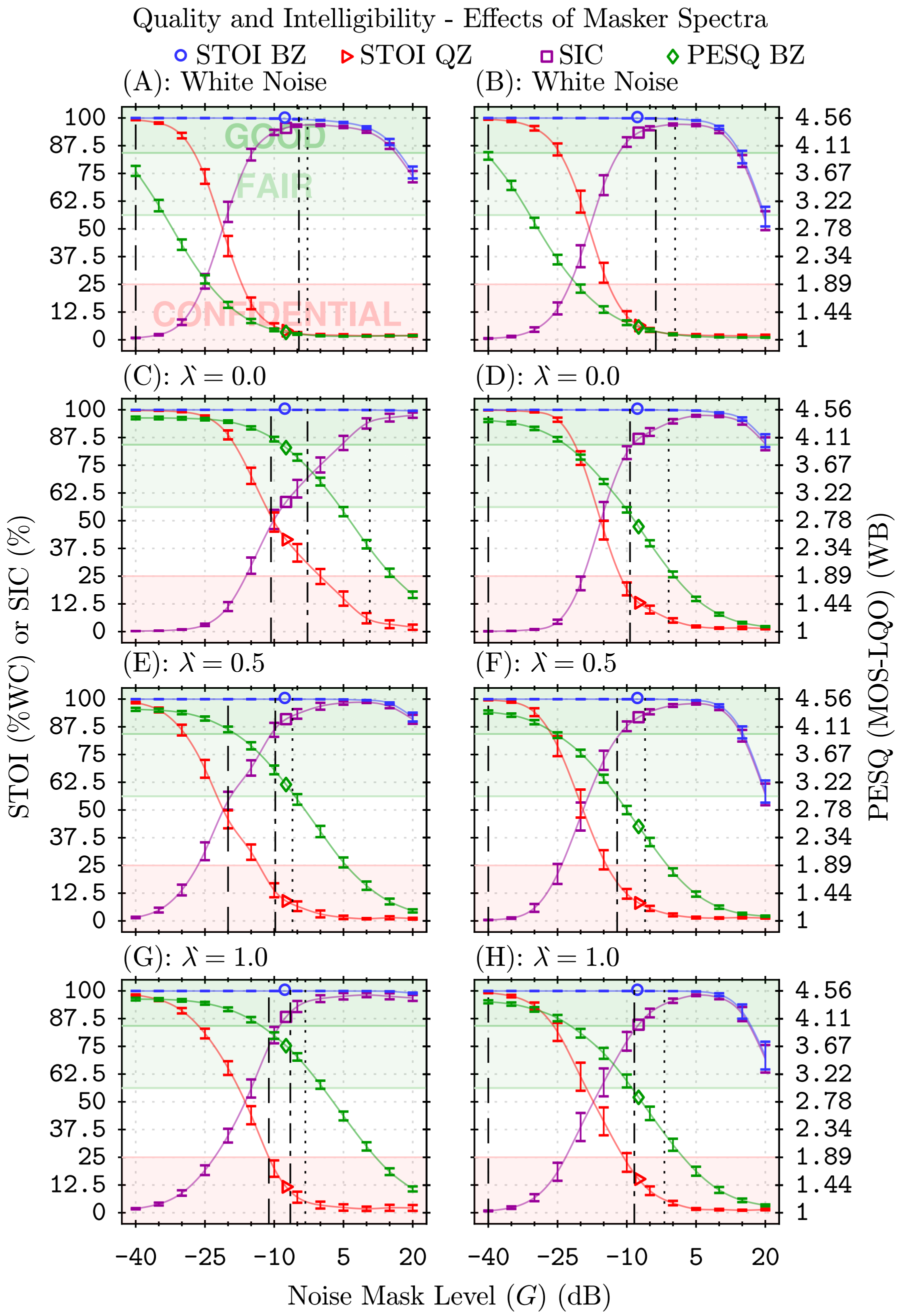

Figure 4.8: Mean STOI and PESQ are shown for different masking spectra and different array types with $L = 24$. A and B are for a white noise, C and D are $\lambda = 0.0$, E and F are $\lambda = 0.5$ and G and H are $\lambda = 1.0$. The left column is for semi-circular array reproductions and the right column is for linear array reproductions. Optimum G (dB) is indicated by the vertical black dotted lines for $\lambda = 0.33$, dash-dot lines for $\lambda = 1.0$ and dashed lines for $\lambda = 3.0$. Good and fair PESQ MOS scores [177] are labelled and shaded in green and confidential speech privacy [210] is labelled and shaded in red. BZ and QZ are the bright and quiet zone, respectively. 95% confidence intervals over θ, microphone positions and speech variation are given.

Spectrum Shape and Weighting

While Figure 4.8 (A, B) show a good separation between the two $\mathcal{I}_{\mathrm{STOI}}$ results, the wideband white masker (without the grating lobe filter, $\{\mathrm{lp}\}$) that is used still keeps the $\mathcal{B}_{\mathrm{PESQ}}$ low in the region where $\mathrm{SIC}_{\mathrm{STOI}}$ is high. To allow for both high valued $\mathcal{B}_{\mathrm{PESQ}}$ and $\mathrm{SIC}_{\mathrm{STOI}}$, the spectrum is shaped and the results in Figure 4.8 (C–H) show how $\mathcal{B}_{\mathrm{PESQ}}$ and $\mathrm{SIC}_{\mathrm{STOI}}$ can be tuned with the parameter $\grave{\lambda}$. The hypothesis that low valued $\grave{\lambda}$ improves masking performance over $\mathbb{D}_{\mathrm{q}}$ to increase $\mathrm{SIC}_{\mathrm{STOI}}$ and high valued $\grave{\lambda}$ reduces masking effects over $\mathbb{D}_{\mathrm{b}}$ to increase $\mathcal{B}_{\mathrm{PESQ}}$ is confirmed in Figure 4.8. The case where $\grave{\lambda} = 0.5$ gives on average the best separation between the two $\mathcal{I}_{\mathrm{STOI}}$ results whilst maintaining a high valued $\mathcal{B}_{\mathrm{PESQ}}$. For cases where $\mathrm{SIC}_{\mathrm{STOI}}$ is required to be high and $\mathcal{B}_{\mathrm{PESQ}}$ is of less importance, $\grave{\lambda} = 1.0$ may sometimes provide slightly better results than $\grave{\lambda} = 0.5$, as can be seen in Figure 4.8 (G).

Array Geometry

The two different array geometries evaluated are the semi-circular array and linear array where results are shown in the first and second column, respectively, in Figure 4.8 and Figure 4.9. The main observable difference is that the linear array provides slightly less contrast between the two $\mathcal{I}_{\mathrm{STOI}}$ and therefore a slightly smaller range of high valued $\mathrm{SIC}_{\mathrm{STOI}}$. This difference in contrast has more influence on the resulting $\mathcal{B}_{\mathrm{PESQ}}$ in Figure 4.8, however, it is still possible to obtain high valued $\mathcal{B}_{\mathrm{PESQ}}$ and high valued $\mathrm{SIC}_{\mathrm{STOI}}$. It should be noted that the loudspeaker spacing, ΔD_{L}, is constant for all results in Figure 4.8. To investigate the effect of differing L, the loudspeaker spacing is varied for results in Figure 4.9 which show that better performance is acquired for smaller values of ΔD_{L} and for a larger number of loudspeakers, L. The semi-circular array performs better than the linear array with the same ΔD_{L} which is caused by the fact that the semi-circular array has a higher low-frequency acoustical brightness contrast between $\mathbb{D}_{\mathrm{b}}$ and $\mathbb{D}_{\mathrm{q}}$. The higher contrast here is a result of the apparent angular window of the array to the multiple zones. However, the linear array does have a slightly higher k_{u} compared to the semi-circular array when L

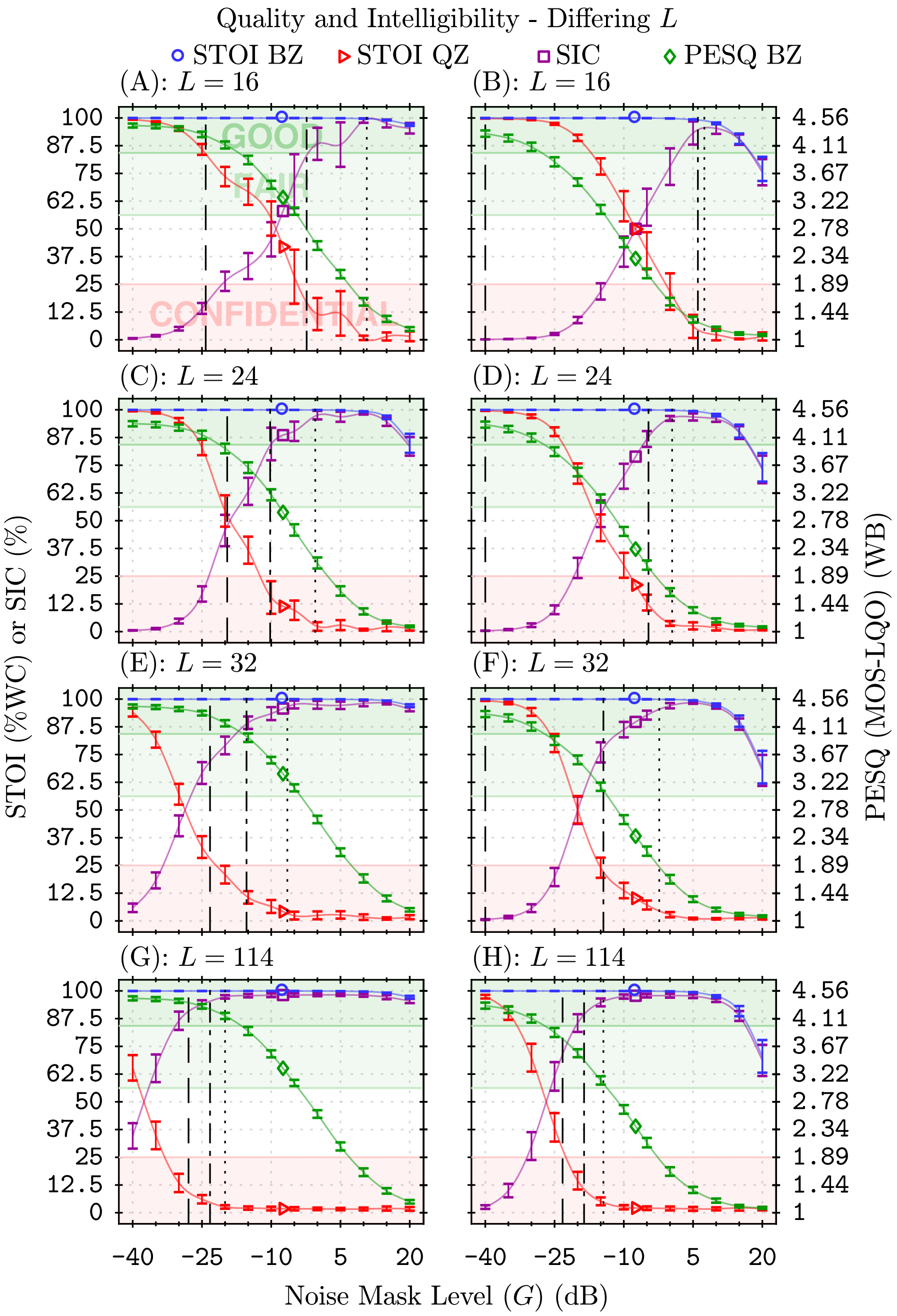

Figure 4.9: Mean STOI and PESQ are shown for different L and different array types with $\lambda = 0.5$. Each loudspeaker count is presented in a row where A and B are $L = 16$, C and D are $L = 24$, E and F are $L = 32$ and G and H are $L = 114$. The left column is for semi-circular array reproductions and the right column is for linear array reproductions where $D_\mathrm{L} = \phi_\mathrm{c} R_\mathrm{c}$. Optimum G (dB) is indicated by the vertical black dotted lines for $\lambda = 0.33$, dash-dot lines for $\lambda = 1.0$ and dashed lines for $\lambda = 3.0$. Good and fair PESQ MOS scores [177] are labelled and shaded in green and confidential speech privacy [210] is labelled and shaded in red. BZ and QZ are the bright and quiet zone, respectively. 95% confidence intervals over θ, microphone positions and speech variation are given.

and ΔD_{L} are the same between array geometries. The linear arrays higher k_{u} does not provide a better $\mathrm{SIC}_{\mathrm{STOI}}$ or $\mathcal{B}_{\mathrm{PESQ}}$, though, because the loss in low frequency contrast for the linear array reduces these values more so, primarily due to the high energy speech content at low frequencies and the only slightly larger k_{u}.

Loudspeaker Count

The loudspeaker count, L, and, more specifically, the loudspeaker spacing, ΔD_{L}, have a large influence on both the performance and the practical feasibility of the system. The influence on performance is shown in Figure 4.9 where as the loudspeaker count increases for a semi-circular array (and hence the speaker spacing decreases) the separation between the two $\mathcal{I}_{\mathrm{STOI}}$ results increases and, for the same optimised values of G, $\mathcal{B}_{\mathrm{PESQ}}$ also increases. The minimum number of loudspeakers in the semi-circular array which still attains good $\mathcal{B}_{\mathrm{PESQ}}$ and $\mathrm{SIC}_{\mathrm{STOI}}$ is the case where $L = 24$, which is good motivation for the number of real-world loudspeakers to use. As the linear array may either use a differing number of loudspeakers with a fixed ΔD_{L} or with a fixed D_{L}, in this work Figure 4.9 presents results for a fixed $D_{\mathrm{L}} = \phi_{\mathrm{c}} R_{\mathrm{c}}$ as this maintains a constant valued k_{u}, consistent with the semi-circular array for direct comparison. Results related to a potentially more practical scenario, where ΔD_{L} is fixed, and proportional to the dimensions of a smaller real-world loudspeaker, the reader is referred to Figure 4.8. Simulations for varying L with the linear array follow the same trend as those for the semi-circular array where, as L increases, both $\mathcal{B}_{\mathrm{PESQ}}$ and $\mathrm{SIC}_{\mathrm{STOI}}$ also increase.

4.8 Real-World Implementation

To compliment simulations, a practical real-world implementation has been evaluated in anechoic conditions. This section provides details of the hardware, calibration and recorded results.

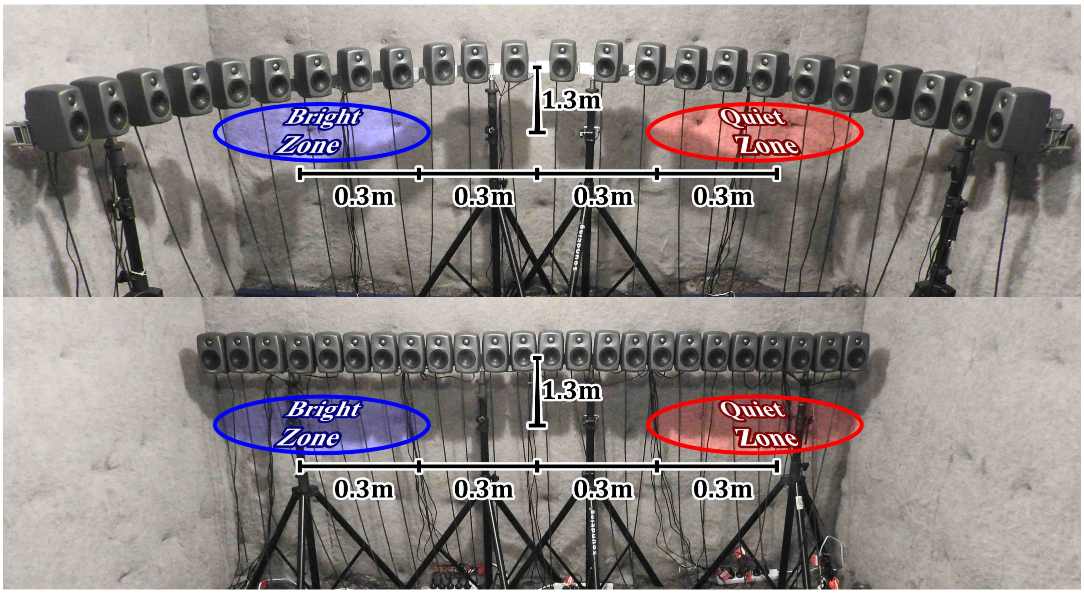

Figure 4.10: The two real-world multizone implementations are pictured. The semi-circular and linear array are shown on the top and bottom, respectively. The bright zone (blue) on the left and the quiet zone (red) on the right are separated by 1.2 m and each have a radius of 0.3 m. The centre of the reproduction region is midway between both zones and is 1.3 m from the centre of the loudspeaker array.

4.8.1 Hardware Setup

The multizone audio reproduction systems described in section 4.7.1 were implemented in a flat-walled multilayered anechoic chamber measuring 4.8 m × 3.3 m × 2.4 m. The systems consisted of 24 loudspeakers evenly spaced on a semi-circle of radius 1.3 m and a line of length 2.8 m as shown in Figure 4.10. Recordings of the reproduced speech were received using 4 × Behringer ECM8000 measurement microphones in each zone, positioned equidistant along a 0.3 m diameter circle (concentric with the zone). The loudspeaker models were all Genelec 8010A studio monitors with a free field frequency response of 74 Hz to 20 kHz (±2.5 dB). The loudspeakers and microphones were driven by 3 × Behringer ADA8200 8-channel input/output audio interfaces connected to a computer via an RME HDSPe RayDAT 36-channel input/output soundcard. The software used to generate, playback and record the multizone soundfield was Mathworks' MATLAB R2017a.

4.8.2 System Calibration and Response

In order to ensure a flat magnitude response and correct phase response for all loudspeakers, a calibration procedure is performed. The calibration is the application of system equalisation filters computed from inverse system transfer functions found by using an exponential sine sweep (ESS) method[5] [230]. Prior to applying the inverse filters, the loudspeaker signals, ($q'_l(n)$ from $Q'_l(a,k)$) are upsampled by interpolating with a factor of 3 from 16 kHz to 48 kHz, to match that of the reproduction system due to the sampling frequency mismatch. The band-pass inverse filters are then convolved with the upsampled loudspeaker signals. Soundfield recordings are performed using the upsampled calibrated loudspeaker signals and in order to compare with simulated recordings, the 48 kHz sampled recordings are downsampled to 16 kHz by a factor of 3 with decimation.

4.8.3 Simulated and Real-World Comparison

To confirm that the calibration procedure allows for a flat magnitude response in the target bright zone, within the accuracy of the loudspeakers (i.e. ±2.5 dB), the response over $\mathbb{D}_\mathrm{b}$ and $\mathbb{D}_\mathrm{q}$ is measured by reproducing and recording a multizone weighted ESS. Afterwards, the $\mathrm{SIC}_{\mathrm{STOI}}$ and $\mathcal{B}_{\mathrm{PESQ}}$ are computed and compared with simulated results using speech samples and measured ATFs.

Sound Pressure Levels

The SPL is found for $\theta = 24.8°$ and results do not vary significantly for different values of θ (as explained in 4.7.4). Figure 4.11 shows that the real-world multizone magnitude response over $\mathbb{D}_\mathrm{b}$ is flat and lies within ±2.5 dB, even after the signal has been processed and other system noises have been included. The real-world SPL over

[5] The ESS is generated as a 10 s sweep from 100 Hz to 10 kHz with a 1 s buffer of silence before and after. The system is set to a sampling frequency of 48 kHz after which the ESS is reproduced one loudspeaker at a time and recorded from the centre of $\mathbb{D}$. The calibration filters are computed from the recordings with a length of 0.5 s and are regularised so that the maximum pass-band gain is 60 dB and stop-band gain is -6 dB.

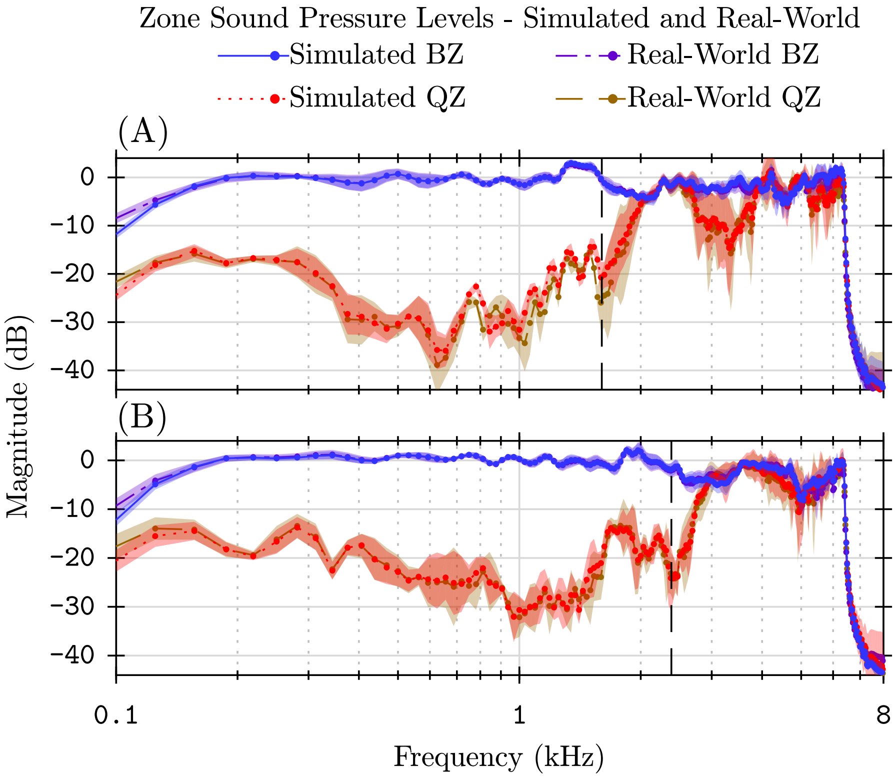

Figure 4.11: Mean SPLs are shown for the simulated and real-world cases with $L = 24$ for a semi-circular array (A) and linear array (B) where $\theta = 24.8°$. 95 % confidence intervals over the microphone positions in each zone are shaded and the vertical black dashed line is k_u. BZ and QZ are the bright and quiet zone, respectively.

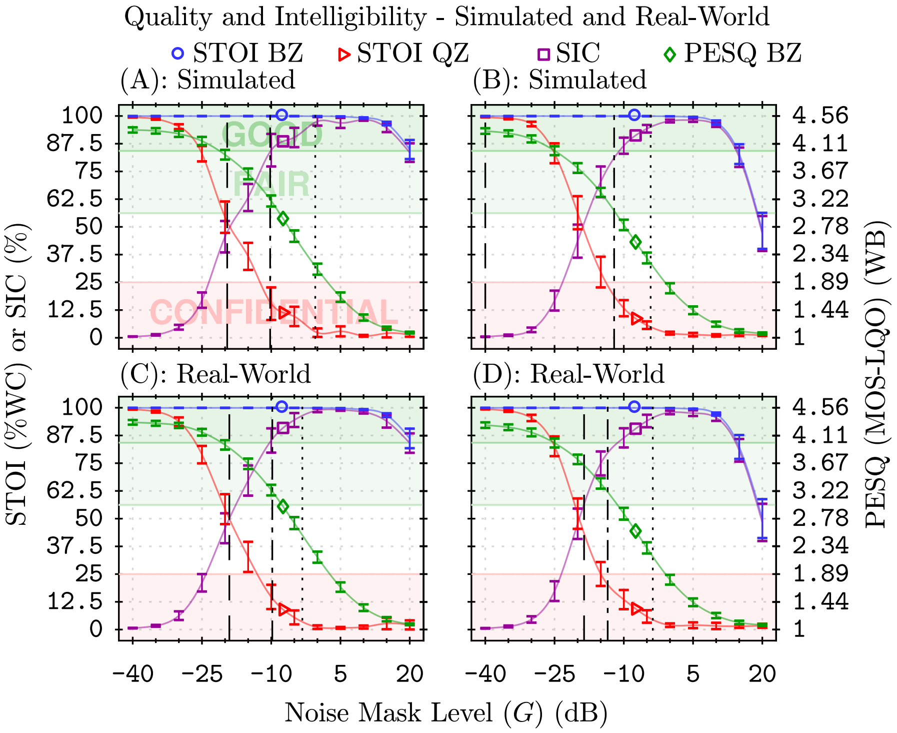

Figure 4.12: Mean STOI and PESQ are shown for the simulated (A–B) and real-world (C–D) anechoic environment with $\theta = 24.8°$, $\lambda = 0.5$. The left column is for semi-circular array reproductions and the right column is for linear array reproductions where $D_{\mathrm{L}} = \phi_{\mathrm{c}} R_{\mathrm{c}}$. Optimum G (dB) is indicated by the vertical black dotted lines for $\lambda = 0.33$, dash-dot lines for $\lambda = 1.0$ and dashed lines for $\lambda = 3.0$. Good and fair PESQ MOS scores [177] are labelled and shaded in green and confidential speech privacy [210] is labelled and shaded in red. BZ and QZ are the bright and quiet zone, respectively. 95% confidence intervals over θ, microphone positions and speech variation are given.

$\mathbb{D}_{\mathrm{q}}$ also agrees with simulated SPL over $\mathbb{D}_{\mathrm{q}}$ with only slight variations when using measured ATFs as shown in Figure 4.11. The average SPL up to $\min(\hat{k}_{\mathrm{u}}', \overline{k}_{\mathrm{u}})$ over $\mathbb{D}_{\mathrm{q}}$ for the real-world scenario is considerably low at -25.5 dB for the semi-circular array and -24.9 dB for the linear array. The equivalent acoustic brightness contrast, following [85], [86], between $\mathbb{D}_{\mathrm{b}}$ and $\mathbb{D}_{\mathrm{q}}$ for the real-world scenario is 25.6 dB for the semi-circular array and 25.0 dB for the linear array.

Speech Intelligibility Contrast and Quality

The $\mathcal{I}_{\mathrm{STOI}}$ and $\mathcal{B}_{\mathrm{PESQ}}$ in Figure 4.12 are seen to be almost identical between the real-world and simulated results. Figure 4.11 suggests this would likely be the case.

For the real-world case using a semi-circular array, $\lambda = 0.5$ and $\lambda = 0.33$ gives optimal $G = -3.26\,\text{dB}$, the results obtained are $\text{SIC}_{\text{STOI}} = 96.4\,\%$ and $\mathcal{B}_{\text{PESQ}} = 2.52\,\text{MOS}$ indicating confidential speech privacy and better than poor speech quality, respectively. $\lambda = 1.0$ gives $G = -9.77\,\text{dB}$, $\text{SIC}_{\text{STOI}} = 85.9\,\%$ and $\mathcal{B}_{\text{PESQ}} = 3.22\,\text{MOS}$ indicating confidential privacy and better than fair quality, respectively, and $\lambda = 3.0$ gives $G = -19.1\,\text{dB}$, $\text{SIC}_{\text{STOI}} = 50.0\,\%$ and $\mathcal{B}_{\text{PESQ}} = 3.92\,\text{MOS}$ indicating normal privacy and better than fair quality (close to good quality), respectively. The results show that λ successfully controls the trade-off between speech privacy and speech quality where a lower value λ emphasises privacy and higher valued λ emphasises quality. In this chapter, results are obtained with as few as 16 loudspeakers, significantly less than most modern WFS systems, and with the use of noisy real-world equipment.

For the real-world case using a linear array, $\grave{\lambda} = 0.5$ and $\lambda = 0.33$ gives optimal $G = -3.72\,\text{dB}$, the results obtained are $\text{SIC}_{\text{STOI}} = 95.5\,\%$ and $\mathcal{B}_{\text{PESQ}} = 2.17\,\text{MOS}$ indicating confidential privacy and better than poor quality, respectively. $\lambda = 1.0$ gives $G = -13.5\,\text{dB}$, $\text{SIC}_{\text{STOI}} = 79.7\,\%$ and $\mathcal{B}_{\text{PESQ}} = 3.21\,\text{MOS}$ indicating confidential privacy and better than fair quality, respectively, and when $\lambda = 3.0$ gives $G = -18.6\,\text{dB}$, $\text{SIC}_{\text{STOI}} = 56.6\,\%$ and $\mathcal{B}_{\text{PESQ}} = 3.64\,\text{MOS}$ indicating normal privacy and better than fair quality, respectively. These results show that the real-world linear array performs just as well as the real-world semi-circular array and that λ still successfully controls the trade-off between speech privacy and speech quality. This is fortuitous as a linear array is a more practical implementation for box-shaped rooms.

4.9 Conclusion and Contributions

We proposed a method for improving the speech privacy and quality in multizone soundfield reproductions by using robust spatial and temporal frequency domain filters on masking signals. Practical implementations are facilitated by the proposed methods; masking filters are analytically derived in order to avoid spatial aliasing artefacts and secondary leakage is accounted for using weighting parameters on *a priori* estimates of multizone spectral leakage. The practical benefits include robustness to variations in the reproduced speech, virtual source location and array geometry, and a significantly reduced number of the required loudspeakers.

Results have shown that it is necessary to account for multizone leakage when performing masking or when high quality reproductions are required. It is also shown that estimating the aliasing frequency is of importance when the loudspeaker count and geometry can vary. A more robust estimation of the aliasing frequency has also been shown to provide more reliable results. System performance is also dependent on the acoustic contrast between the zones which may vary depending on the reproduction technique used and the real-world equipment setup and calibration. The results presented verify the benefits of the proposed method for practical implementations. The analytically derived filters and optimal gains are shown capable of providing good and fair MOS ratings for speech quality whilst providing normal and confidential privacy, respectively, via measured SIC values in simulated environments. The real-world implementation, and the results thereof, confirm the practical feasibility of the proposed methods by also showing that good and fair speech quality, with respective normal and confidential speech privacy, can be reproduced amongst personal sound zones.

Future work could include investigations on the perceived annoyance of different sound maskers and their influence on cognitive performance. Evaluations of simultaneous reproductions of speech in multiple zones and the effect of joint optimisations using temporal and spatial filters are also potential topics for future work.

Chapter 5

The Active Control of Speech Sound Field Interference

Overview: *In this chapter, we investigate the effects of compensating for wave-domain filtering delay in an active speech control system and we compare the performance of two active dereverberation techniques using a planar array of microphones and loudspeakers. An active speech control system utilising wave-domain processed basis functions is evaluated for a linear array of dipole secondary sources. The target control soundfield is matched in a least squares sense using orthogonal wavefields to a predicted future target soundfield. Filtering is implemented using a block-based short-time signal processing approach which induces an inherent delay. We present an autoregressive method for predictively compensating for the filter delay. An approach to block-length choice that maximises the soundfield control is proposed for a trade-off between soundfield reproduction accuracy and prediction accuracy. The two dereverberation techniques are based on a solution to the Kirchhoff-Helmholtz Integral Equation (KHIE). We adapt a Wave Field Synthesis (WFS) based method to the application of real-time 3D dereverberation by using a low-latency pre-filter design. The use of First-Order Differential (FOD) models is also proposed as an alternative method to the use of monopoles with WFS and which does not assume knowledge of the room geometry or primary sources. The two dereverberation methods* *are compared by observing the suppression of a single active wall over the volume of a room in the time and temporal-frequency domain. Results show that block-length choice has a significant effect on the active suppression of speech. The FOD approach to dereverberation provides better suppression of reflections than the WFS based method but at the expense of using higher order models. The equivalent absorption coefficients are comparable to passive fibre panel absorbers. The methods proposed in this chapter indicate that significant active suppression of soundfield interference is feasible.*

5.1 Introduction

Spatial regions of controlled sound can be created using loudspeaker arrays and superposition of soundwaves can be used to actively control sound over space [39], [231]. Active Noise Control (ANC) is a technique that allows secondary sources in electro-acoustic systems to reproduce destructive soundfields thus reducing energy levels of primary soundfields. The resultant suppressed soundfields have been successfully employed in several applications, including noise-cancelling headphones [232] and ANC in vehicle cabins [35], [233], [234]. Offices, libraries, teleconferencing rooms, restaurants and cafes may also benefit from ANC over broad spatial areas where physical partitions could be replaced with an active loudspeaker array.

ANC systems typically comprise a reference signal and/or error signal which are either fed forward and/or backward, respectively, to an algorithm for generating loudspeaker signals [39], [231]. Hybrid systems exist that incorporate both feedforward and feedback techniques [235], [236]. Least Mean Squares (LMS) and Filtered-x LMS (FxLMS) control methods work by adaptively minimising the error signal in a least squares sense [237], [238]. Multichannel systems with numerous microphones inside, or near, the control space often use adaptive algorithms to minimise the error over the region [238], [239].

More recent techniques have been shown to be more accurate by measuring acoustic pressures on boundaries and using the Kirchhoff-Helmholtz integral to determine the soundfield [240]–[242]. Sampling the boundary that encloses the space, with microphones, allows the target soundfield to be estimated in the wave-domain. This extends the multipoint method by synthesising the entire spatial area and minimising the error over large spaces [241], [242].

In order to perform wave-domain analysis it is necessary to transform received signals into the (temporal) frequency domain where basis functions are a function of the wavenumber and spatial locations [114], [240]. This transformation induces a delay where numerous samples are required to analyse the signal with high resolution in the frequency domain. Adaptive algorithms overcome this issue by automatically compensating for any errors received at the error microphones [237], [241], [242]. In scenarios where microphones are not placed inside the control region, it is necessary to account for delay by other means. Linear prediction with pitch repetition has been shown to be viable for active speech cancellation with short predictions, up to 2 ms, and at discrete points in a space [243]. However, the predictions do not predict a regular length speech frame of around 16 ms and cancellation occurs only in the vicinity of the control points.

The active control of sound over a linear array has been envisioned [244] using interconnected control units consisting of a microphone, directional loudspeaker and processing modules. However, the interconnection and modules do not model the received signals on the boundary in the wave-domain and perform only a phase inversion, which is less robust to soundfield variation. Linear arrays [61] have also been investigated for improvement of noise barriers [245], [246] which aim to reduce diffraction of sound over a physical barrier by minimising the pressure at points in space, usually modelled in two spatial dimensions with the linear array normal to the plane. The use of linear arrays, without a physical barrier, for control over large spatial areas using recently advanced wave-domain processing is the first part of work explored in this chapter.

The active control of acoustic sound fields is a useful process for suppressing undesirable sound over large spaces. Acoustic reflections, or echoes, inside listening rooms are a common source of undesirable sound field contributions, notably, in the degradation of sound field reproductions using WFS [135] or HOA [135], [219] as discussed in chapter 2. There exist room equalisation and dereverberation techniques [51], [247] to reduce the influence of reflections on system performance as well as active techniques to produce desired subjective experiences by adding more reflections to rooms [248].

While the majority of dereverberation techniques focus on post-processing the recorded signals [249]–[251] there has been research into the active suppression of reflected sound fields [51], [252]–[254]. The suppression of any sound field requires a desired sound field to be synthesised, for which the process is commonly called Sound Field Synthesis (SFS) [135] or soundfield reproduction as has been described in chapter 2. Some SFS methods use higher-order loudspeakers and/or microphones to reduce error or loudspeaker counts [222], [255], [256]. While there are numerous techniques to perform SFS, the state-of-the-art methods generally rely on a solution to the wave equation [2], often through the use of the Kirchhoff-Helmholtz Integral Equation (KHIE) [61], [135], [219], which is described in detail in chapter 2.

Dedicated calibration processes often used to compensate for reverberation in listening rooms [257], [258] require knowledge of the room, or the room itself, and provide compensation tailored to the particular room. Other techniques employ pre-filtering of single loudspeaker channels by reshaping Room Impulse Responses (RIRs) [249], [250]. Further approaches rely on feedback from microphones within the cancellation region to adapt filters using Wave-Domain Adaptive Filtering (WDAF) or modal decompositions [51], [252], [253]. There have also been techniques proposed that use FOD sources in circular arrays to cancel 2D exterior fields [259].

ANC systems generally rely on a feedforward or feedback system which require error microphones to adaptively weight the system and reduce errors from the previous state of the system [39], [40]. While the adaptive nature of ANC systems generally ensure convergence to an optimal solution, the convergence rate may be slow and any abrupt changes in the environment may degrade performance [39]. These systems often require modelling of secondary paths between the secondary sources and the error microphones. Improvement of the erroneous secondary path models is a topic of ongoing research. There exist ANC techniques which do not require secondary path modelling but their convergence rate is lower than state-of-the-art ANC techniques, such as the Filtered-x Least Mean Square (FxLMS) algorithm [260]–[263]. Other methods make use of reflections to aid cancellation [264].

While the majority of ANC algorithms rely on single or multi-point approaches some applications rely on ANC over larger areas, such as the cancellation of vehicle cabin noise [37], [234]. The recording and reproduction of a sound field over a large space, termed Wave Field Reconstruction (WFR), has been thoroughly researched [60], [265] and real-time systems have been realised [266]. The inherent latency, when using current filter designs, of real-time WFR systems deems them unusable for applications of non-adaptive ANC. Low-latency, or zero-latency, WFR filters are highly beneficial for adaptive and non-adaptive low-latency ANC.

For the active speech control method, we analyse the delay caused by transforming reference ANC signals to the wave-domain using a block-based signal processing approach. We propose an autoregressive transform-delay compensator in conjunction with an inverse filter that together produce a virtual source soundfield used in wavefield decomposition to minimise energy residual of a control soundfield. Through analysis of the soundfield suppression we show that an optimal block-length can be chosen for active speech control using wave-domain filtering without error microphones in the control region. The optimal block-length is used in a simulated acoustic environment with dipole secondary sources in a linear array. Acting as an active wall, we show that the optimal block-length, along with the dipole sources, provide significant cancellation of traversing speech waves with minimal reproduction towards the primary source.

To enhance the performance of the reproduction systems we further propose methods to dereverberation in closed rooms that absorb reflections using an active wall. We look at two possible methods; the first is using monopole models with a

WFS-based method and the second is using differential (pressure gradient) models as a direct solution to the KHIE. For the first technique we provide a novel contribution by repurposing the WFR method for 3D boundary cancellations and reducing the need for adaptive filters. We propose the use of a Weighted Least-Squares (WLS) pre-filter for low-latency reproduction and cancellation. For the second method we propose the use of FOD (pressure gradient) models as implicit solutions to the KHIE or WFS/WFR pre-filter problem.

A description of the error minimised control soundfield synthesis using basis wavefields for active speech control is given in section 5.2. An explanation of dipole modelled soundfield reproduction using synthesised loudspeaker weights is given in section 5.3. The short-time block-based signal processing approach with autoregressive and geometric delay compensation is presented in section 5.4. For a description of the KHIE used in the dereverberation techniques, the reader is referred to chapter 2 and the WFR derivation is described in section 5.5. The proposed WLS pre-filter design is described in section 5.6 and the FOD models method is given in section 5.7. Results, discussion and conclusions are given in sections 5.8 and 5.9.

5.1.1 Notations and Definitions

In this chapter, we assume 3D Cartesian coordinate space with no specific origin. The volume enclosed by the room is denoted as Ω with the room boundary of interest, $\mathcal{C} \equiv \partial\Omega$, and observation points are $\mathbf{x} \in \Omega$. Loudspeaker locations are $\mathbf{l}$ and microphone locations are $\mathbf{z}$. The normal to $\mathcal{C}$ is $\mathbf{n}$ and the tangential plane, $\mathbf{t}$, is perpendicular to $\mathbf{n}$. The wavenumber is $k = \omega/c$ where ω is the angular frequency and $c = 343\,\mathrm{m\,s^{-1}}$ is the speed of sound in air. The unit imaginary number is $i = \sqrt{-1}$. The image source notation in Fig. 5.1 is given as $\iota^{(\bar{n})}_{x,y,z}$ where $\bar{n}$ is the order of the image source and (x, y, z) are the coordinates of the imaged room relative to the primary room.

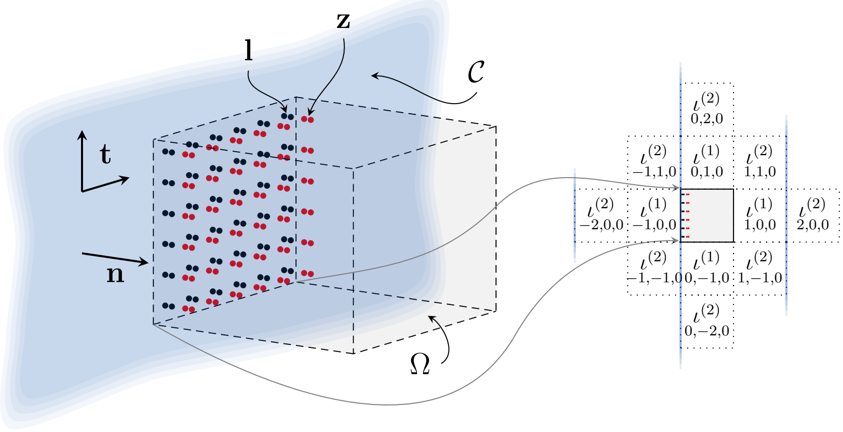

Figure 5.1: An active dereverberation scenario is shown. Left: Active dipole wall (black loudspeakers and red microphones) and spatial 3D geometry. Right: Equivalent image source layout for the evaluation.

5.2 Wave-Domain Soundfield Suppression

This section derives an expression for loudspeaker weights which reproduce a soundfield that minimises the residual energy over a control region, $\mathbb{D}_{\rm c}$. The active control layout and wave-domain solution to minimise residual energy are described.

5.2.1 Active Control Layout and Definitions

The proposed system using a 2D linear dipole array, with propagation described by the cylindrical Hankel function, is shown in Fig. 5.2 where the loudspeakers form an active wall between a talker and target quiet zone. The reproduction region for the soundfield, $\mathbb{D}$, with spatial sampling points $\mathbf{x} \in \mathbb{D}$, has a radius of $R_{\mathbb{D}}$ and contains a control subregion, $\mathbb{D}_{\rm c} \subseteq \mathbb{D}$, of radius $r_{\rm c}$. The centre of the loudspeaker array is located at angle $\bar{\phi}$ and distance $R_{\rm c}$. The length of the loudspeaker array is $D_{\rm L}$ and is designed to reproduce a soundfield for a virtual point source located at $\mathbf{v}$. In this work, we refer to the external source that is to be controlled as the talker with location $\mathbf{t} \equiv \mathbf{v} \equiv (r_{\rm t}, \theta_{\rm t})$ and with a 2D soundfield also described by the cylindrical Hankel function. We assume $\mathbf{t}$ is known, or can be reliably estimated with

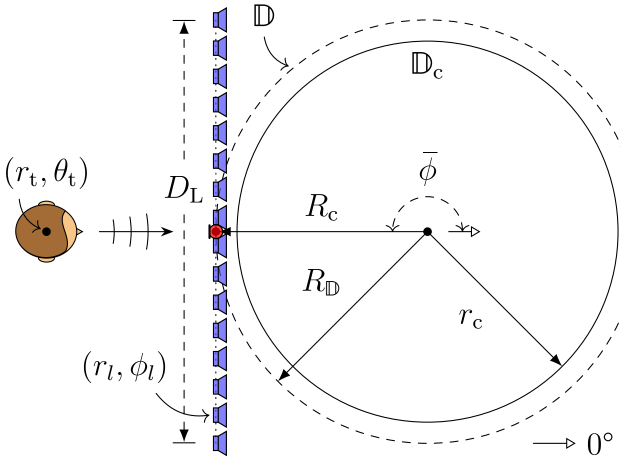

Figure 5.2: Active control layout for a linear dipole array (blue) directed to the right. The microphone (red) is used to predict the unwanted speech source crossing the array.

multiple microphones, thus a single reference microphone suffices and is placed at the centre of the loudspeaker array with location $\mathbf{z} \equiv \left(R_\mathrm{c}, \bar{\phi}\right)$. Loudspeaker locations are $\mathbf{l}_l \equiv (r_l, \phi_l)$ for $l \in [\![\bar{L}]\!]$ where $\bar{L}$ is the number of loudspeakers, $k = 2\pi f/c$ is the wavenumber and c is the speed of sound in air. The Euclidean norm is denoted using $\|\cdot\|$, $i = \sqrt{-1}$ and sets of indices are $[\![A]\!] \triangleq \{x : x \in \mathbb{N}_0, x < A\}$.

5.2.2 Soundfield Control Technique

The goal is to find coefficients for a set of basis functions that minimise the residual energy of the sum of a control soundfield, $S^\mathrm{c}(\mathbf{x}; k)$, and an arbitrary talker soundfield, $S^\mathrm{t}(\mathbf{x}; k)$. A simple solution is to perform an orthogonalisation on a set of plane-wave basis functions that produces a well-conditioned triangular matrix and a set of orthogonal basis functions. Expansion coefficients for the orthogonal basis functions can be easily solved with an inner product.

Any arbitrary soundfield can be completely defined by an orthogonal set of solutions of the *Helmholtz* equation [2]. We start by defining an arbitrary 2D control soundfield function, $S^\mathrm{c}(\mathbf{x}; k) : \mathbb{D} \times \mathbb{R} \to \mathbb{C}$, as an actual soundfield from (2.81),

$$S^\mathrm{c}(\mathbf{x}; k) = \sum_{j \in [\![J]\!]} E_j(k) F_j(\mathbf{x}; k), \tag{5.1}$$

where $\{F_j\}_{j\in[\![J]\!]}$ is the set of orthogonal basis functions, $m \in [\![N]\!]$ are N frequency indices, the expansion coefficients for a particular frequency are E_j and J is the number of basis functions [112].

Solving the inner product $E_j = \langle S^{\mathrm{t}}(\mathbf{x};k), F_j(\mathbf{x};k)\rangle$ yields the E_j that minimise

$$\min_{E_{j\in[\![J]\!],m\in[\![N]\!]}} \|\sum_j E_j(k)F_j(\mathbf{x};k) + S^{\mathrm{t}}(\mathbf{x};k)\|^2, \tag{5.2}$$

where $\|X\|^2 = \langle X, X\rangle$. The set of orthogonal basis functions, $\{F_j\}_{j\in[\![J]\!]}$, can be found by implementing an orthogonalisation on a set of planewaves, $P_h(\mathbf{x};k) = \exp(ik\mathbf{x}\cdot\boldsymbol{\rho}_h)$, where $\boldsymbol{\rho}_h \equiv (1,\rho_h)$, $\rho_h = (h-1)\Delta\rho$ and $\Delta\rho = 2\pi/J$. A Gram-Schmidt process gives the orthogonalised basis functions, which results in [112]

$$F_j(\mathbf{x};k) = \sum_{h\in[\![J]\!]} \mathbf{R}_{hj,m}P_h(\mathbf{x};k), \tag{5.3}$$

such that $\langle F_i(\mathbf{x};k), F_j(\mathbf{x};k)\rangle = \delta_{ij}$, where $\mathbf{R}_{hj}$ is the (h,j)th element of the lower triangular matrix, $\mathbf{R}$. Substituting (5.3) in (5.1), yields

$$S^{\mathrm{c}}(\mathbf{x};k) = \sum_{h\in[\![J]\!]} \mathcal{W}_{m,h}P_h(\mathbf{x};k), \tag{5.4}$$

where $\mathcal{W}_{h,m} = \sum_{j\in[\![J]\!]} E_j\mathbf{R}_{hj,m}$ are the plane-wave coefficients used to construct an approximation of the control soundfield.

5.3 Loudspeaker Weights

In this section, the loudspeaker signals needed for soundfield reproduction with monopole and dipole sources are described.

5.3.1 Monopole Secondary Source Weights

To reproduce $S^{\mathrm{c}}(\mathbf{x};k)$ with minimal error to $S^{\mathrm{t}}(\mathbf{x};k)$, frequency domain loudspeaker weights are found using [109], [143]

$$W_l(k) = \frac{2\Delta\phi_{\mathrm{s}}}{i\pi} \sum_{\bar{m}=-\overline{M}}^{\overline{M}} \sum_{h\in[\![J]\!]} \frac{i^{\bar{m}}\exp(i\bar{m}(\phi_l - \rho_h))}{\mathscr{H}_{\bar{m}}^{(1)}(r_l k)} \mathcal{W}_{h,m}, \tag{5.5}$$

where $\Delta\phi_{\mathrm{s}} = 2\tan^{-1}(D_{\mathrm{L}}/2R_{\mathrm{c}})/\bar{L}$ approximates angular spacing of $\mathbf{l}_l$ for a linear array, $\mathscr{H}_{\nu}^{(1)}(\cdot)$ is a νth-order Hankel function of the first kind and $\overline{M} = \lceil kR_{\mathbb{D}} \rceil$ is the modal truncation length [109]. However, monopole sources produce acoustic energy in all directions which may be undesirable as it would present an artificial echo towards $\mathbf{t}$.

5.3.2 Dipole Secondary Source Weights

To reproduce a soundfield with reduced acoustic energy presented towards the talker, dipole sources are modelled with cardioid radiation patterns to reproduce predominantly over $\mathbb{D}$. In this work, we refer to the monopole source pairs as dipole sources while their radiation pattern is designed to be that of a cardioid. The loudspeakers at $\mathbf{l}_l$ with weights $W_l(k)$ are split into two point sources at $\mathbf{l}_{l,s}$ for $s \in [\![2]\!]$ with weights $Q_{l,s}(k)$. The dipole source pair locations are given by

$$\mathbf{l}_{l,s} = \mathbf{l}_l + (\ddot{d}/2, \bar{\phi} - s\pi), \tag{5.6}$$

where $\ddot{d}$ is the distance between the dipole point sources. The objective of each dipole source pair is to reproduce a wave which constructs in the direction $(1, \bar{\phi} - \pi)$ from $\mathbf{l}_l$ and de-constructs in the direction $(1, \bar{\phi})$ from $\mathbf{l}_l$ whilst maintaining the same amplitude and phase as a monopole source in the constructive direction. This can be accomplished by phase shifting and amplitude panning the monopole loudspeaker weights with the following [2], [5]

$$Q_{l,s}(k) \triangleq W_l(k)\frac{\exp\Big(i(-1)^s(k\ddot{d}-\pi)/2\Big)}{2k\ddot{d}}, \tag{5.7}$$

where as $\ddot{d}$ becomes small, $\mathbf{l}_{l,s}$ approach ideal dipole sources.

5.4 Short-time Signal Processing

In order to reproduce a control soundfield, a time-domain control signal is filtered using $Q_{l,s}(k)$ in the (temporal) frequency domain and inverse transformed back to the time-domain to yield the set of loudspeaker signals. Here, a block based approach is used. This section investigates the inherent time delay that is induced during the filtering process due to the wave-domain transformation used to compute the loudspeaker weights of (5.7).

5.4.1 Block Processing

An input signal, $v(n)$, broken into blocks (frames) using an analysis windowing function, $w(n)$, of length M, results in an ath windowed frame:

$$\tilde{v}_a(n) \triangleq v(n+aR)w(n), \tag{5.8}$$

where $n \in \mathbb{Z}$ is the sample number in time, $a \in \mathbb{Z}$ is the frame index and $R \leq M$ is the step size in samples. The ath frame is transformed to the frequency domain to give the ath spectral frame as

$$\tilde{V}_a(k_m) = \sum_{n \in [\![N]\!]} \tilde{v}_a(n)\exp\Big(-icnk_m/2\dot{f}\Big), \tag{5.9}$$

where $k_m \triangleq 2\pi\dot{f}m/cN$ and the frame is oversampled with $N \geq M+L-1$ for a filter length L.

Each spectral frame is filtered using $Q_{l,s}(k)$ from (5.7) up to the maximum frequency, $\dot{f}$, and inverse transformed to the time-domain

$$\tilde{q}_{a,l,s}(n) = \mathrm{Re}\left\{ \frac{1}{N} \sum_{m \in [\![N]\!]} Q_{l,s}(k_m) \tilde{V}_a(k_m) \exp\Big(icnk_m / 2\dot{f} \Big) \right\}, \tag{5.10}$$

$\forall n \in [\![N]\!]$, where $\mathrm{Re}\{\cdot\}$ returns the real part of its argument, after which a synthesis window, $w(n)$, equivalent to the analysis window, is applied to yield the weighted output

$$q^{w}_{a,l,s}(n) = \tilde{q}_{a,l,s}(n - aR) w(n - aR). \tag{5.11}$$

The weighted output, $q^{w}_{a,l,s}(n)$, is added to the accumulated output signal, $q_{l,s}(n)$, for each dipole source. The analysis and synthesis windows are chosen so that

$$\sum_{a \in \mathbb{Z}} w(n - aR)^2 = 1, \quad \forall n \in \mathbb{Z}. \tag{5.12}$$

5.4.2 Autoregression Parameter Estimation

The soundfield filtering process induces a delay of M samples to build the current ath frame, $\tilde{v}_a(n)$, from (5.8), essential for accurate reproduction. To perform active control, it is necessary to find R future samples of the accumulated $q_{l,s}(n)$ that estimate $v(n)$.

Forecasting the input signal's future values can be accomplished using an autoregressive (AR) linear predictive filter. Assuming the signal is unknown after the current time, n, the AR parameters, $\hat{a}_j$, are estimated using $B > \mathcal{P}$ known past samples with

$$\epsilon(n + \grave{b} + 1) = v(n + \grave{b} + 1) + \sum_{j \in [\![\mathcal{P}]\!]} \hat{a}_j v(n + \grave{b} - j), \tag{5.13}$$

$\forall \grave{b} \in \mathcal{B}$, where $\mathcal{B} = \{-B, \ldots, \mathcal{P} - 1\}$, $\{\epsilon(n + \grave{b} + 1)\}_{\grave{b} \in \mathcal{B}}$ are prediction errors, the predictor order is $\mathcal{P}$ and $j \in [\![\mathcal{P}]\!]$ are the coefficient indices. Stable AR coefficients, $\hat{a}_j$, can be estimated using the *autocorrelation method* [267], [268] (equivalent to the *Yule-Walker method*) by approximating the minimisation of the expectation of

$|\epsilon(n+\grave{b}+1)|^2, \forall\grave{b}\in\mathbb{Z}$, i.e.

$$\arg\min_{\widehat{a}_j}\sum_{\grave{b}\in\mathcal{B}}\left|\epsilon(n+\grave{b}+1)\right|^2, \tag{5.14}$$

where, prior to minimisation, $v(n+\grave{b}+1)$ is windowed with $\bar{w}(\grave{b})$, assuming

$$\{\bar{w}(\grave{b})\}_{\grave{b}\notin\{-B,\ldots,-1\}}=0, \tag{5.15}$$

to give $\bar{v}(\grave{b})$. Multiplying (5.13) by $v(n+\grave{b}-\breve{b}), \breve{b}\in[\![\mathcal{P}]\!]$ and taking the expectation gives the Yule-Walker (YW) equations,

$$\sum_{j\in[\![\mathcal{P}]\!]} r_{\breve{b}-j}\widehat{a}_j=-r_{\breve{b}}. \tag{5.16}$$

We estimate the jth autocorrelation, r_j, as

$$\widehat{r}_j\triangleq B^{-1}\sum_{\grave{b}=j}^{-1}\bar{v}(\grave{b})\bar{v}(\grave{b}-j). \tag{5.17}$$

The YW equations can be written in matrix form as

$$\widehat{\mathbf{R}}\widehat{\mathbf{a}}=-\widehat{\mathbf{r}}, \tag{5.18}$$

where

$$\widehat{\mathbf{a}}=[\widehat{a}_0,\ldots,\widehat{a}_{\mathcal{P}-1}]^{\mathsf{T}}, \tag{5.19}$$

$$\widehat{\mathbf{r}}=[\widehat{r}_0,\ldots,\widehat{r}_{\mathcal{P}-1}]^{\mathsf{T}} \tag{5.20}$$

and the estimated autocorrelation matrix, $\widehat{\mathbf{R}}$, has a Toeplitz structure allowing for an efficient solution using Levinson-Durbin recursion [268].

An example of an input speech signal forecasted into the future is shown in Figure 5.3. The example prediction is performed using the autocorrelation procedure outlined in this section. Figure 5.3 shows that for a finite time into the future the

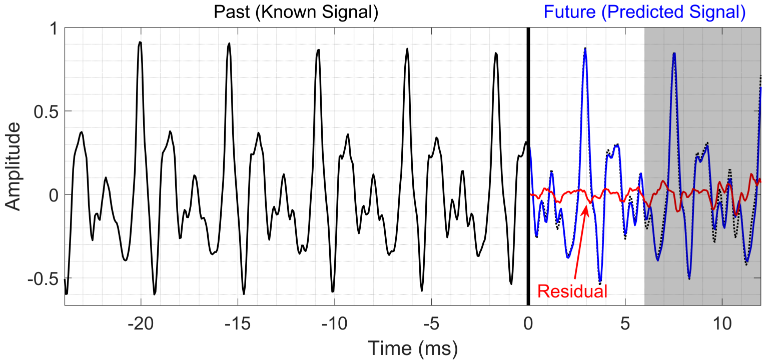

Figure 5.3: An example of the autocorrelation method of autoregression predicting a segment of a speech signal 12 ms into the future. The solid black line represents the signal that has already past in time, the dotted black line indicates the true future signal, the solid blue line shows the predicted future signal and the solid red line depicts the residual, which is the difference between the true future signal and the predicted future signal.

AR prediction works well with less than −20 dB of residual on average. It is also apparent that the further into the future the signal is forecasted the less accurate the prediction becomes. The prediction accuracy is dependent on the signal and its autocorrelation properties. For example, AR methods will have difficulty predicting any transients in the signal.

5.4.3 Filter-Delay Compensation

Once the $\hat{a}_j$ are estimated following section 5.4.2, $v(n)$ can be extrapolated by

$$v(n+\acute{b}+1) = -\sum_{j\in[\![\mathcal{P}]\!]} \hat{a}_j v(n+\acute{b}-j), \quad \forall \acute{b} \in [\![\widehat{M}]\!] \tag{5.21}$$

where $\{v(n+\acute{b}+1)\}_{\acute{b}\in[\![\widehat{M}]\!]}$ are $\widehat{M}$ future estimates of $v(n)$. From (5.8), $\tilde{v}_a(n)$ is an estimated future windowed frame when $\widehat{M} \geq M$. The estimated $\tilde{v}_a(n)$ and partially estimated $\{\tilde{v}_{a-\grave{a}-1}(n)\}_{\grave{a}\in[\![\frac{M}{R}-1]\!]}$ are transformed, filtered, inverse transformed and windowed through (5.10) and (5.11). Adding $q^{w}_{a,l,s}(n)$ to the previous frames obtains R future estimated samples for the output loudspeaker signals, $q_{l,s}(n)$. The procedures of section 5.4.2 and section 5.4.3 are repeated every R samples, including the estimation of $\hat{a}_j$.

5.4.4 Geometric-Delay Compensation

The control soundfield modelling requires a virtual source location and signal. In this work, the reference microphone recording, $z(n)$, located at $\mathbf{z}$, is an attenuated and time delayed version of $v(n)$. Under the assumption of free-space and that the talker location, $\mathbf{t}$, is known, or can be reliably estimated, the talker signal is found by

$$v(n) = \mathrm{Re}\left\{ \frac{1}{N} \sum_{m \in [\![N]\!]} \frac{4\left\{ \sum_{n \in [\![N]\!]} z(n) \exp\left(-icnk_m/2\dot{f}\right) \right\}}{i\mathcal{H}_0^{(1)}(k_m \|\mathbf{v} - \mathbf{z}\|)} \exp\left(icnk_m/2\dot{f}\right) \right\}, \tag{5.22}$$

where $z(n)$ is inverse filtered in the frequency domain with N sufficiently large compared to the time-delay. For the purpose of soundfield control, $\mathbf{t} \equiv \mathbf{v}$ and $v(n)$ is also the virtual source signal.

5.4.5 Loudspeaker Signals and Speech Suppression Reproduction

Upon receiving the reference signal, $z(n)$, the final cardioid loudspeaker signals, $q_{l,s}(n)$, are produced by firstly compensating for the geometric-delay with (5.22) to obtain $v(n)$. The virtual source signal is then extrapolated by $\widehat{M}$ future estimates computed with (5.21). The estimated $v(n)$ is transformed to the frequency domain after (5.8). The cardioid loudspeaker weights, $Q_{l,s}(k)$, are computed with (5.7) through (5.5) after $\mathcal{W}_{h,m}$ is found via (5.2) and (5.3).

For the reproduction, $Q_{l,s}(k)$ are used as filters via (5.10) to obtain $q_{l,s}(n)$. The actual reproduced control soundfield is given by

$$\mathcal{S}^{\mathrm{c}}(\mathbf{x}; k) = \sum_{l \in [\![L]\!], s \in [\![2]\!], n \in \mathbb{Z}} q_{l,s}(n) \exp\left(-icnk/2\dot{f}\right) T(\mathbf{x}, \mathbf{l}_{l,s}; k), \tag{5.23}$$

$\forall \mathbf{x} \in \mathbb{D}_{\mathrm{c}}$, where the 2D acoustic transfer function for each source is $T(\mathbf{x}, \mathbf{l}; k) = \frac{i}{4}\mathcal{H}_0^{(1)}(k\|\mathbf{l} - \mathbf{x}\|)$. Note, $\mathcal{S}^{\mathrm{c}}(\mathbf{x}; k)$ depends on $v(n)$.

5.5 Wave Field Reconstruction (WFR)

Previous work has shown that the WFS method can be used to accurately reproduce sound fields from sound field recordings [60]. Recently, the WFR filtering method has looked at efficiently transforming recorded signals into driving signals [265], [266]. In this section we propose a method for the design and use of a WFR filter for low-latency real-time dereverberation. We extend the formulations from two spatial dimensions, for active speech cancellation over a boundary, to three spatial dimensions for facilitating the dereverberation of entire rooms.

5.5.1 Receiving

We start by defining a desired sound field, $S^{\mathrm{d}}(\mathbf{x};\omega)$, reflected by a boundary wall and which is to be cancelled. A planar monopole microphone and loudspeaker array are placed at the boundary. The planar microphone array and secondary source loudspeaker array are both modelled as continuously distributed arrays.

The sound pressure gradient at the microphone array is used to find the reflected sound field back in to the room. The reflections are the half-space sound field of the loudspeaker wall.

Rayleigh's first integral for a plane from (2.23) gives the desired 3D spatio-temporal sound field [265]

$$S^{\mathrm{d}}(\mathbf{x};\omega) = -2 \iint\limits_{\mathcal{C}} \frac{\partial S^{\mathrm{d}}(\mathbf{z};\omega)}{\partial \mathbf{n}} G(\mathbf{x},\mathbf{z};\omega)\, d\mathcal{C},\ \forall \mathbf{z} \in \mathcal{C}, \tag{5.24}$$

where $\partial/\partial\mathbf{n}$ is the pressure gradient at $\mathcal{C}$, the noiseless desired sound pressure at the microphones, $\mathbf{z} \equiv \mathbf{x}_0$, is $S^{\mathrm{d}}(\mathbf{z};k)$ and, for half-space and small $\|\mathbf{z}-\mathbf{l}\|$, we assume the free space Greens function [2],

$$G(\mathbf{x},\mathbf{x}';\omega) = \frac{\exp\!\left(i\frac{\omega}{c}\|\mathbf{x}-\mathbf{x}'\|\right)}{4\pi\|\mathbf{x}-\mathbf{x}'\|}. \tag{5.25}$$

The goal now is to find the relationship between the microphone signals and the desired loudspeaker signals by using $S^{\mathrm{d}}(\mathbf{x};\omega)$.

5.5.2 Reproduction

The actually reproduced sound field, $S^{\mathrm{a}}(\mathbf{x};\omega)$, of the planar loudspeaker array is given by [2],

$$S^{\mathrm{a}}(\mathbf{x};\omega) = \iint\limits_{\mathcal{C}} Q_{\mathrm{WFS}}(\mathbf{l};\omega) G(\mathbf{x},\mathbf{l};\omega)\, d\mathcal{C},\ \forall \mathbf{l} \in \mathcal{C}, \tag{5.26}$$

where $Q_{\mathrm{WFS}}(\mathbf{l};\omega)$ is the WFS loudspeaker driving signal [60]. The reproduced sound field, $S^{\mathrm{a}}(\mathbf{x};\omega)$, must match that of the inverted reflected sound field, $-S^{\mathrm{d}}(\mathbf{x};\omega)$, so that $S^{\mathrm{a}}(\mathbf{x};\omega) = -S^{\mathrm{d}}(\mathbf{x};\omega)$. The loudspeaker array and microphone array share the boundary, $\mathcal{C}$, where $\mathbf{l} \equiv \mathbf{z}$, and so (2.23) and (5.24) give

$$Q_{\mathrm{WFS}}(\mathbf{l};\omega) = 2\frac{\partial S^{\mathrm{d}}(\mathbf{z};\omega)}{\partial \mathbf{n}}, \tag{5.27}$$

where the sound pressure gradient at $\mathbf{z}$ is found using Euler's equation as (a tilde indicating the spatial frequency domain) [265]

$$\begin{aligned} \frac{\partial S^{\mathrm{d}}(\mathbf{z};\omega)}{\partial \mathbf{n}} &= \frac{\partial}{\partial \mathbf{n}}\left(\frac{1}{4\pi^2}\iint\limits_{\mathcal{C}} \widetilde{S}^{\mathrm{d}}(k_{\mathbf{t}}, \mathbf{x}\cdot\mathbf{n}, \omega)\exp(-ik\mathbf{x})\, dk_{\mathbf{t}}\right) && (5.28)\\ &= \frac{1}{4\pi^2}\iint\limits_{\mathcal{C}} -ik_{\mathbf{n}}\widetilde{S}^{\mathrm{d}}(k_{\mathbf{t}}, \mathbf{x}\cdot\mathbf{n}, \omega)\exp(-ik_{\mathbf{t}}\mathbf{t})\, dk_{\mathbf{t}} && (5.29)\\ &= -ik_{\mathbf{n}} S^{\mathrm{d}}(\mathbf{z};\omega) && (5.30) \end{aligned}$$

and $k_{\mathbf{n}} = \sqrt{k^2 - k_{\mathbf{t}}^2}$. The loudspeaker driving signal is then

$$Q_{\mathrm{WFS}}(\mathbf{l};\omega) = F(\exp(ik)) S^{\mathrm{d}}(\mathbf{z};\omega), \tag{5.31}$$

$$F(\exp(ik)) = -2ik_{\mathbf{n}}. \tag{5.32}$$

The desired loudspeaker signals are given by the microphone signals with the parameter-independent multiplier operator, $F(\exp(ik))$.

5.6 Planar Array WFS/SDM Pre-Filter Design

The relationship between sound pressure and particle velocity gives rise to a +6 dB/oct magnitude gain with a constant 90° phase shift. This section describes the design of a filter required to compensate for $F(\exp(ik))$ so that the reproduced sound field is of the correct amplitude and phase for cancellation to occur.

5.6.1 Weighted Least Squares (WLS) Method

While it is simple to create a linear-phase Finite Impulse Response (FIR) filter directly from $F(\exp(ik))$ which provides +6 dB/oct. gain and 90° phase shift, it is not as simple to design a minimum-phase equivalent. Linear-phase is suitable for applications which do not require low-latency filtering, such as for the reproduction of a pre-recorded sound field. However, for sound field cancellation, low-latency and filter accuracy is important. The WLS method can approximate the desired response while the weighting relieves constraint on the minimisation for frequency bands that are of less importance.

The WLS compensation filter

$$H(z) \triangleq \frac{B(z)}{A(z)} \triangleq \frac{\sum_{m=0}^{N_b} b_m z^{-m}}{\sum_{m=0}^{N_a} a_m z^{-m}}, \tag{5.33}$$

can be found by minimising the error

$$\min_{b_m, a_m} \sum_{m=0}^{\widehat{N}} \left|(A(\exp(ik_m))F(\exp(ik_m)) - B(\exp(ik_m)))W(k_m)\right|^2, \tag{5.34}$$

where $\widehat{N} = N - 1$, discrete Fourier transform (DFT) length is N and $W(k_m)$ is a bandpass weighting for $F(\exp(ik))$.

The WLS approach is implemented with

$$\begin{bmatrix}\mathbf{B}\\ \mathbf{A}\end{bmatrix} = (\mathbf{D}^H\mathbf{W}\mathbf{D})^{-1}\mathbf{D}^H\mathbf{W}\mathbf{f}, \tag{5.35}$$

where $\{\cdot\}^H$ denotes a Hermitian transpose,

$$\mathbf{D} = \begin{bmatrix} -1 & \cdots & -1 \\ -\exp\Big(-i\pi k_0/\hat{k}\Big) & \cdots & -\exp\Big(-i\pi k_{\widehat{N}}/\hat{k}\Big) \\ \vdots & \ddots & \vdots \\ -\exp\Big(-N_b i\pi k_0/\hat{k}\Big) & \cdots & -\exp\Big(-N_b i\pi k_{\widehat{N}}/\hat{k}\Big) \\ F(\exp(ik_0))\exp\Big(-i\pi k_0/\hat{k}\Big) & \cdots & F\Big(\exp\Big(ik_{\widehat{N}}\Big)\Big)\exp\Big(-i\pi k_{\widehat{N}}/\hat{k}\Big) \\ \vdots & \ddots & \vdots \\ F(\exp(ik_0))\exp\Big(-N_a i\pi k_0/\hat{k}\Big) & \cdots & F\Big(\exp\Big(ik_{\widehat{N}}\Big)\Big)\exp\Big(-N_a i\pi k_{\widehat{N}}/\hat{k}\Big) \end{bmatrix}^{\mathsf{T}}, \tag{5.36}$$

$$\mathbf{W} = \mathrm{diag}\Big(\Big[W(k_0) \quad \cdots \quad W(k_{\widehat{N}})\Big]\Big), \tag{5.37}$$

$$\mathbf{f} = -\Big[F(\exp(ik_0)) \quad \cdots \quad F\Big(\exp\Big(ik_{\widehat{N}}\Big)\Big)\Big]^{\mathsf{T}} \tag{5.38}$$

and $\hat{k} = 2\pi f_{\mathrm{s}}/c$ with sampling frequency, f_{s}. The WLS solution gives the coefficients

$$\mathbf{B} = \Big[b_0 \quad \cdots \quad b_{N_b}\Big]^{\mathsf{T}}, \qquad \mathbf{A} = \Big[a_0 \quad \cdots \quad a_{N_a}\Big]^{\mathsf{T}}, \tag{5.39}$$

which are used to construct the desired filter, $H(z)$, using (5.33). The weight can then be designed to relieve the constraint on the least squares optimisation as described in the following section.

5.6.2 Weight Design

Due to the discretised loudspeaker and microphone array, there is an aliasing frequency, k_u, where accuracy degrades at higher frequencies. In practice it is unnecessary to constrain the filter design above k_u.

The value of k_u is dependent on the finite spacing between array elements. Its lowest value is used to find the least squares weighting

$$W(k) = \begin{cases} 1, & k \leq k_\mathrm{u} \\ 0, & k > k_\mathrm{u} \end{cases}, \qquad k_\mathrm{u} = \frac{\pi}{\Delta D_\mathrm{L}} = \frac{\pi}{\Delta \mathbf{z}}, \tag{5.40}$$

where ΔD_L and $\Delta \mathbf{z}$ are the spacing between adjacent loudspeakers and microphones, respectively, and $W(k)$ weights the importance of the minimisation above and below k_u. Other weights may give low-latency low-pass filters thus reducing the influence of aliasing.

5.7 Half-space recording and reproduction

In practice, it is important for the microphone wall to record only the signal coming from the half-space within the room and, similarly, for the loudspeaker wall to only reproduce into the half-space that is the room. While omnidirectional monopole models simplify the analysis of the problem, their implementation in practice is less desirable than FOD models which can be less dependent on feedback loops. In this section, an overview of the FOD model used in this work is given.

5.7.1 First-Order Differential (FOD) Source/Receiver Model

As can be seen from (5.32), the multiplier operator has most influence along the normal, $\mathbf{n}$. However, the filter designed using (5.33) and (5.35) is spatially independent and, therefore, does not approximate the response of $F(\exp(ik))$ along the plane, $\mathbf{t}$.

This results in inaccurate cancellation for sound components propagating parallel to $\mathbf{t}$.

FOD receivers and sources are better suited to the KHIE as they are, themselves, a combination of monopole and dipole responses. Measuring the pressure and particle velocity on $\partial\Omega$ allows for the driving signals to be directly obtained. Using (2.15) from chapter 2 we derive

$$S^{\mathrm{a}}(\mathbf{x};\omega) = \iint\limits_{\partial\Omega} G(\mathbf{x},\mathbf{l};\omega)\dot{Q}(\mathbf{l};\omega) - \ddot{Q}(\mathbf{l};\omega)\frac{\partial G(\mathbf{x},\mathbf{l};\omega)}{\partial\mathbf{n}}\,ds, \tag{5.41}$$

$$\dot{Q}(\mathbf{l};\omega) = -\frac{\partial S^{\mathrm{d}}(\mathbf{z};\omega)}{\partial\mathbf{n}}, \qquad \ddot{Q}(\mathbf{l};\omega) = -S^{\mathrm{d}}(\mathbf{z};\omega), \tag{5.42}$$

where the monopole and dipole driving signals are $\dot{Q}(\mathbf{x}_0;\omega)$ and $\ddot{Q}(\mathbf{x}_0;\omega)$, respectively. This results in the monopole and dipole driving signals being directly obtained from the dipole and monopole microphone signals, respectively. The ratio of $\left|\dot{Q}(\mathbf{l};\omega)\right|$ to $\left|\ddot{Q}(\mathbf{l};\omega)\right|$ gives the *time delay ratio* which can be used to determine the radiation pattern of the FOD model for small dipole separation distances.

5.8 Results and Discussion

In this section we describe the experimental setup and discuss the results obtained from the methods for the proposed active speech cancellation and dereverberation techniques.

5.8.1 Active Speech Control Setup

For the active speech control evaluation, the layout of Fig. 5.2 is used with $R_{\mathbb{D}} = R_{\mathrm{c}} = 1\,\mathrm{m}$, $r_{\mathrm{c}} = 0.9\,\mathrm{m}$, $\bar{\phi} = \pi$ and $D_{\mathrm{L}} = 2.1\,\mathrm{m}$. There are $\bar{L} = 18$ dipole speaker pairs with $\ddot{d} \ll 1/k_{\max} = 2.73\,\mathrm{cm}$ spacing [2], [5], where $k_{\max} = 2\pi(2\,\mathrm{kHz})/c$ and $c = 343\,\mathrm{m\,s^{-1}}$. Spatial aliasing in the soundfield reproduction begins to occur near 2 kHz which reduces the control capability. All signals are sampled at a rate of 16 kHz with a frame

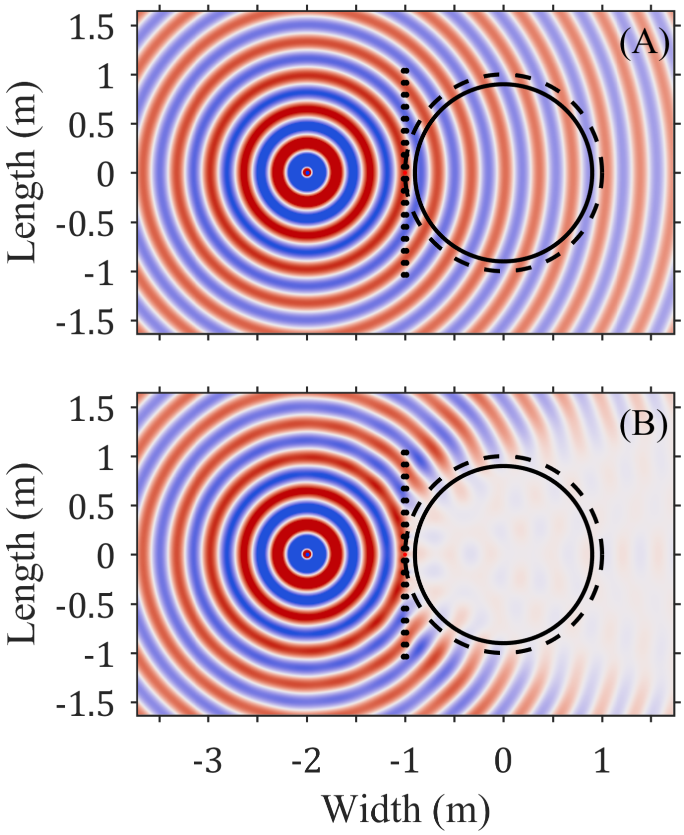

Figure 5.4: The pressure field for an ideal periodic cancellation at 1kHz when the linear dipole array is inactive (A) and active (B).

step of $R = 0.5M$ for 50% overlapping and $M = \{64, 128, 192, 256, 320, 384, 448, 512\}$ are window lengths in samples. A prediction of $\widehat{M} = M$ future samples is made using $B = 2M$ past samples with an order of $\mathcal{P} = M$. The window, $w(n)$, is a square root Hann window. The location of the talker is $\mathbf{t} = (2\,\mathrm{m}, \pi)$ and speech samples used to evaluate the performance were obtained from the TIMIT corpus [214]. Twenty speech segments, approximately 3 s each, were randomly chosen such that the selection was constrained to have a final male to female speaker ratio of $1:1$.

5.8.2 Soundfield Suppression

In order to evaluate the suppression of the control system, 32 virtual microphones are placed in random locations throughout $\mathbb{D}_\mathrm{c}$. The actual control and talker soundfields, $\mathcal{S}^\mathrm{c}(\mathbf{x}; k)$ and $\mathcal{S}^\mathrm{t}(\mathbf{x}; k)$, respectively, are approximated over $\mathbb{D}_\mathrm{c}$ using the

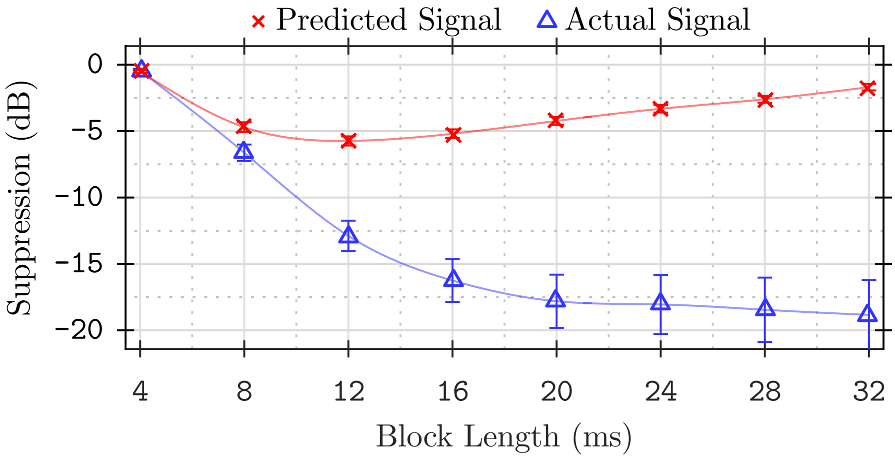

Figure 5.5: The mean suppression, ζ, computed using 1/6th octave band means from 156 Hz to 2 kHz over 2.54 m^2 for an actual future block in blue and predicted in red. 95% confidence intervals are shown.

32 virtual recordings. To gauge the performance of the system, the normalised acoustic suppression between $\mathcal{S}^{\mathrm{c}}(\mathbf{x};k)$ and $\mathcal{S}^{\mathrm{t}}(\mathbf{x};k)$ is defined as

$$\zeta(k) \triangleq \frac{\int_{\mathbb{D}_{\mathrm{c}}} |\mathcal{S}^{\mathrm{t}}(\mathbf{x};k) + \mathcal{S}^{\mathrm{c}}(\mathbf{x};k)| \, d\mathbf{x}}{\int_{\mathbb{D}_{\mathrm{c}}} |\mathcal{S}^{\mathrm{t}}(\mathbf{x};k)| \, d\mathbf{x}}, \tag{5.43}$$

where $\mathcal{S}^{\mathrm{c}}(\mathbf{x};k)$ is from (5.23) and, in this work, for simplicity,

$$\mathcal{S}^{\mathrm{t}}(\mathbf{x};k) = \sum_{n\in\mathbb{Z}} v(n) \exp\left(-icnk/2\dot{f}\right) \frac{i}{4} \mathcal{H}_0^{(1)}(k\|\mathbf{v}-\mathbf{x}\|). \tag{5.44}$$

$\zeta(k)$ is found from (5.43) for a range of frequencies from 100 Hz to 8 kHz. The real part of $\mathcal{S}^{\mathrm{t}}(\mathbf{x};k)$ is shown in Figure 5.4 at 1 kHz for when $\mathcal{S}^{\mathrm{c}}(\mathbf{x};k)$ is active and inactive, as an example. Figure 5.4 clearly shows significant suppression on only one side of the linear dipole array providing a large quiet zone across the wall of loudspeakers. It is also apparent that by not strictly sampling the entire boundary of the control region for the Kirchhoff-Helmholtz integral, the loudspeaker array does not restrict the movement of a listener in and out of $\mathbb{D}$.

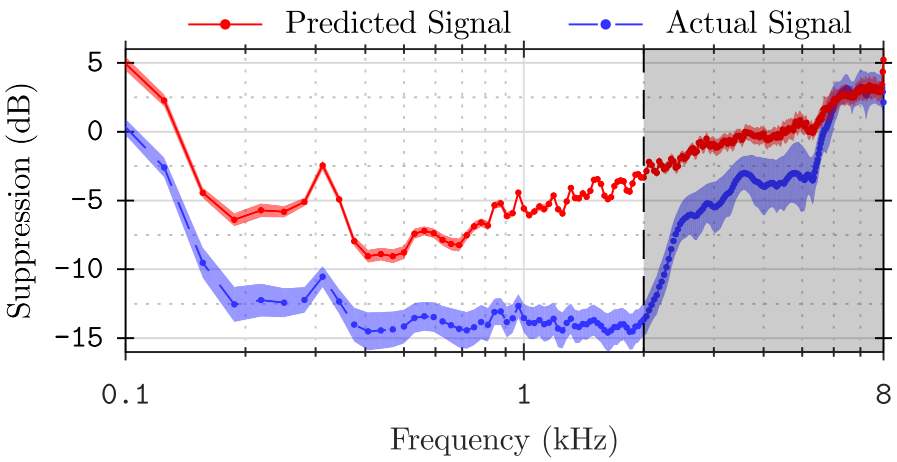

Figure 5.6: The suppression, $\zeta(k)$, for a 12 ms block length from 100 Hz to 8 kHz over $2.54\,\mathrm{m}^2$. 95% confidence intervals are shaded red and blue. The bandwidth where spatial aliasing occurs is shaded grey.

5.8.3 Synthesis and Prediction Accuracy Trade-off

A trade-off between soundfield reproduction accuracy and prediction accuracy is apparent in Figure 5.5 which shows mean suppression from 156 Hz to 2 kHz. Assuming the signal is known (equivalent to a perfect prediction), as shown in blue in Figure 5.5, the longer block length provides better control whereas a longer (and presumably therefore less accurate) prediction is required. A smaller block length is expected to perform worse as it results in fewer analysis frequencies in the wave domain and, hence, is filtered with less accuracy. Using a larger block length overcomes this issue and, assuming perfect prediction, is capable of −18.8 dB of suppression on average over $\mathbb{D}_c$ with a 32 ms block length. However, with the necessary prediction to overcome the filtering delay, as shown in red in Figure 5.5, the longer prediction results in less suppression. The peak suppression occurs with a 12 ms block length and −5.74 dB of suppression on average.

Choosing the block length which attains maximum suppression from Figure 5.5 has the potential to provide the best suppression for wave-domain processed soundfield control. The optimal block length in this case is 12 ms and the suppression for this block length is shown per frequency in Figure 5.6. The downward trend in Figure 5.6 as frequency decreases from 2 kHz suggests that the control from the predicted block

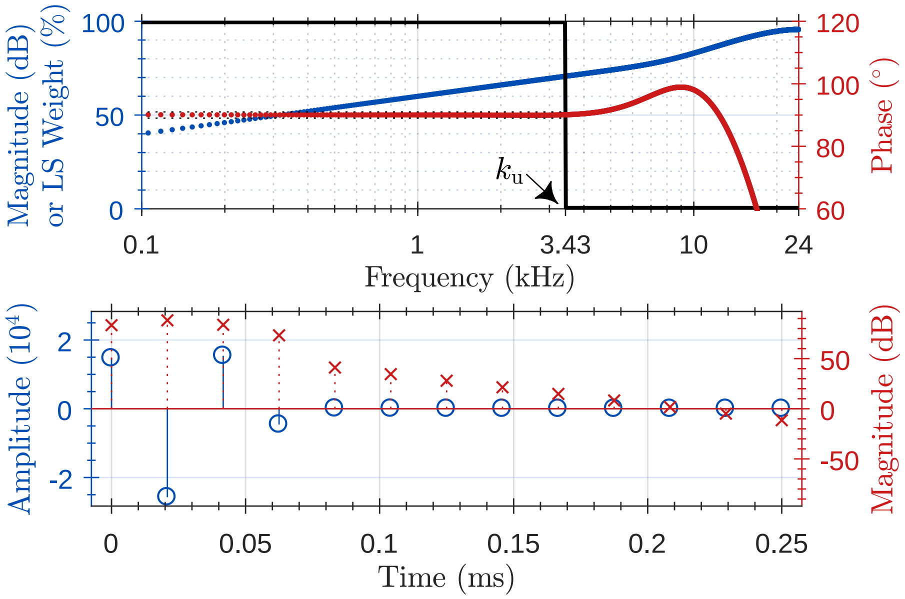

Figure 5.7: Low-latency WFR WLS filter frequency response (top) and impulse response (bottom) are shown. The LS weight shown in black.

performs best for lower frequencies. The increase below 156 Hz and peak near 300 Hz is due to the finite length filter causing a loss of reproduction accuracy. It can be seen from Figure 5.6 that the mean suppression reaches a peak of -9.1 dB near 400 Hz and maintains mean suppression below -7.5 dB from 365 Hz to 730 Hz. Future active speech control work could include investigating the control above the spatial Nyquist frequency by either increasing the loudspeaker density or using hybrid loudspeaker and ANC systems [269]. We discuss methods for reproducing soundfields above the spatial Nyquist frequency in chapter 6. The active speech control system can then be used in conjunction with the dereverberation system, described in the next section, to significantly reduce soundfield interference in reproductions.

5.8.4 Dereverberation Setup

For the dereverberation evaluations, a cube shaped room is used with 3 m length sides and a single wall consists of a planar microphone and loudspeaker array as depicted in Figure 5.1. Both microphone and loudspeaker arrays consist of a 60×60 grid of receivers and sources, respectively. The microphone and loudspeaker spacings

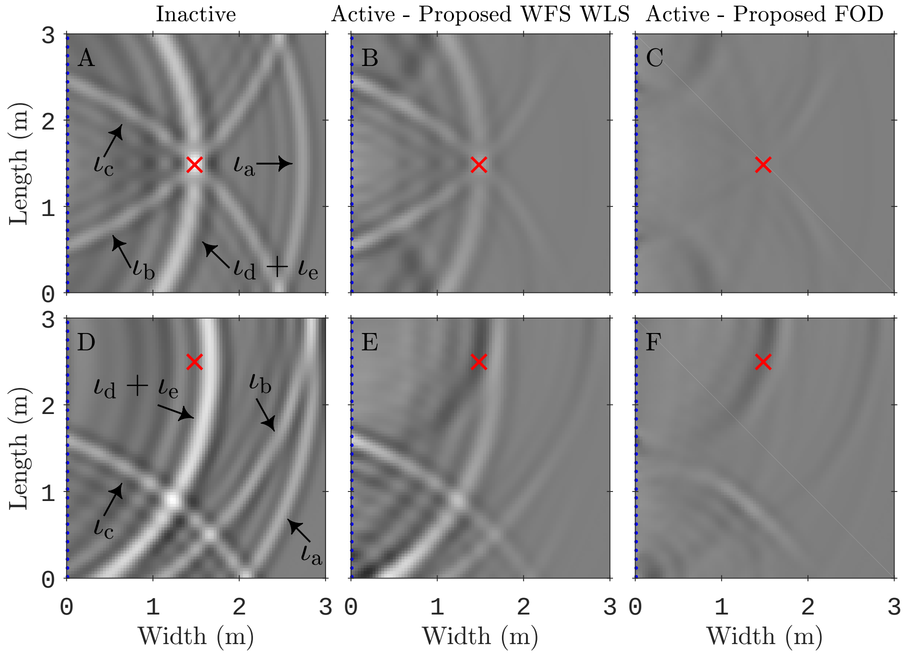

Figure 5.8: The time-domain suppression of first and second order reflections that rebound from $\mathcal{C}$ are shown. The red cross marks the location of the primary source (Top row: room centre. Bottom row: $(1.5\,\mathrm{m}, 2.5\,\mathrm{m}, 1.5\,\mathrm{m})$) and amplitudes are grey-scale normalised.

are $\Delta\mathbf{z} = \Delta D_\mathrm{L} = 5\,\mathrm{cm}$ and the aliasing frequency is $k_\mathrm{u} = 2\pi(3.43\,\mathrm{kHz})/(343\,\mathrm{m\,s^{-1}})$. The sampling frequency is $f_\mathrm{s} = 48\,\mathrm{kHz}$ and DFT length $N = 4096$ with $N_b = 4$ and $N_a = 1$. The order of reflections is set to $\bar{n} = 2$ for an initial investigation of the spatial disparities, however, the formulations are independent of the order of reflections and should behave consistently for increasing reflection order. The image source method of acoustic room reflection modelling is used for evaluation [125], [270].

The WLS frequency response and impulse response can be seen in Figure 5.7. The magnitude and phase response are within ±1 dB and ±1° of the desired, respectively. The filter latency is considered neglible at less than 100 µs and is desirable for real-time cancellation.

5.8.5 Time-Domain Suppression Comparison

The time-domain suppression of a band-limited (150 Hz to 1500 Hz $\ll k_{\mathrm{u}}$) impulse response over a slice of the room ($(x, y, 1.5\,\mathrm{m})$) is shown in Fig. 5.8 for a primary point source located in the centre of the room (top row) and at $(1.5\,\mathrm{m}, 2.5\,\mathrm{m}, 1.5\,\mathrm{m})$ (bottom row). The labels in (A) and (D) of Fig. 5.8 are simplified from Fig. 5.1 with $\iota_{\mathrm{a}} = \iota^{(1)}_{-1,0,0}$, $\iota_{\mathrm{b}} = \iota^{(2)}_{-1,1,0}$, $\iota_{\mathrm{c}} = \iota^{(2)}_{-1,-1,0}$, $\iota_{\mathrm{d}} = \iota^{(2)}_{-1,0,1}$, $\iota_{\mathrm{e}} = \iota^{(2)}_{-1,0,-1}$. Only the reflections that can be suppressed are shown. It is clear from Fig. 5.8 (B) that suppression of ι_{a} is greatest due to $H(\exp(ik))$ being a better approximation to $F(\exp(ik))$ for propagation parallel to $\mathbf{n}$. Fig. 5.8 (C) shows significant improvement for reflections arriving closer to perpendicular to $\mathbf{n}$ which is a direct result of using higher order models to determine the gradient of the sound field at $\mathcal{C}$.

After moving the primary source and observing Fig. 5.8 (E) it is clear that suppression using the WLS pre-filter works best in the direction of $\mathbf{n}$. Using the FOD models, again, provides a better suppression of reflections arriving from angles off the normal direction, $\mathbf{n}$. The small errors that can be seen in Fig. 5.8 (C) and (F) are due to the finite length of the arrays and the finite spacings between array elements which cause diffraction at the edges and time-aliased artefacts, respectively, in the recording and reproduction.

5.8.6 Frequency-Domain Suppression Comparison

The mean frequency suppression and confidence intervals shown in Fig. 5.9 are computed over 200 randomly positioned primary point sources and observation points. The degradation in performance due to spatial aliasing artefacts above k_{u} can be seen in Fig. 5.9 above 3.43 kHz. A cascaded low-latency low-pass filter could be used to mitigate the effect of the spatial aliasing artefacts. While the spatial aliasing artefacts are a limitation of the separation between microphones and loudspeakers, the performance below k_{u} is significantly better than an inactive system, however, low frequency performance is limited by the finite size of the array. Absorption coefficients [271] are found from reflection coefficients which are equivalent to the

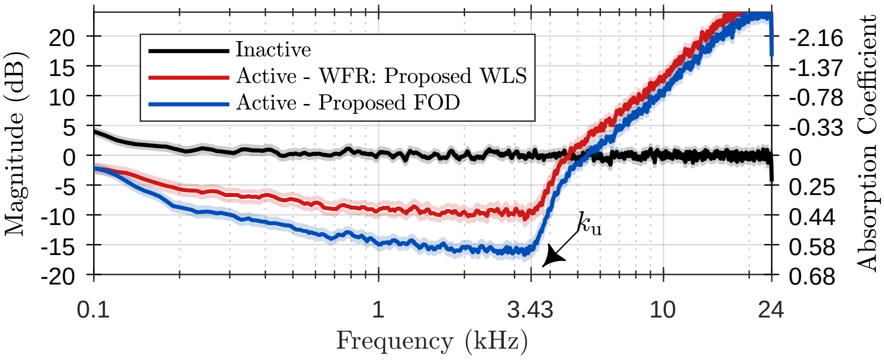

Figure 5.9: The time-domain suppression of first and second order reflections that rebound from $\mathcal{C}$ are shown. The red cross marks the location of the primary source (Top row: room centre. Bottom row: $(1.5\,\mathrm{m}, 2.5\,\mathrm{m}, 1.5\,\mathrm{m})$) and amplitudes are grey-scale normalised.

suppression [270]. The mean suppression below k_u is 9.2 dB for the WFR WLS method, equivalent to a mean absorption coefficient of 0.41. Further improvements in suppression are observed when using the FOD method with a mean suppression of approximately 14.8 dB below k_u, equivalent to a mean absorption coefficient of 0.57.

5.9 Conclusions and Contributions

In this chapter, we have investigated several techniques for actively controlling propagating speech fields and reflected soundfield components for reducing soundfield interference in shared environments. We have investigated the effects of autoregressive delay compensation on active speech control when using wave-domain processing to improve active control over large spatial regions. A system has been proposed using a linear array of secondary dipole sources which uses autoregressive prediction with wavefield decompositions used to minimise residual soundfield energy. We have further considered two active sound field dereverberation techniques for suppressing reflections in closed rooms. We have shown that WFS and WFR systems can be extended to allow real-time low-latency active room compensation using the proposed WLS pre-filter. A system comprised of FOD models has been proposed as an alternative to using the WLS pre-filter method and does not assume knowledge of room geometry or primary sources. The performance of both the active speech control and dereverberation techniques have been evaluated. The proposed active speech control system is capable of a significant mean speech suppression of $-18.8\,\mathrm{dB}$ with an ideally predicted $32\,\mathrm{ms}$ block over a large $2.54\,\mathrm{m}^2$ area. Through analysis of the proposed speech control system, a trade-off between reproduction accuracy and prediction accuracy has been shown to exist. A predicted block with an optimal length of $12\,\mathrm{ms}$ has shown to provide a mean suppression of $-5.74\,\mathrm{dB}$ over a $2.54\,\mathrm{m}^2$ area. Comparison of the two proposed dereverberation methods shows that the relative active absorption performance with the WLS pre-filter method provides a mean suppression of $9.2\,\mathrm{dB}$ (0.41 absorption coefficient) and the FOD model method provides $14.8\,\mathrm{dB}$ of suppression (0.57 absorption coefficient).

In the next chapter, we will discuss techniques for improving the reproduction accuracy of soundfields that are above the spatial Nyquist frequency, as determined by the finite and discrete number of loudspeakers that are necessary for reproductions. The methods in the next chapter can be used to enhance performance of the active interference control methods that have been proposed in this chapter.

Chapter 6

Electrodynamic and Parametric Loudspeaker Hybrid Acoustic Contrast Enhancement

Overview: *This chapter proposes a hybrid approach to personal sound zones utilising multizone soundfield reproduction techniques and parametric loudspeakers. Crossover filters are designed, to switch between reproduction methods, through analytical analysis of aliasing artifacts in multizone reproductions. By realising the designed crossover filters, wideband acoustic contrast between zones is significantly improved. The trade-off between acoustic contrast and the bandwidth of the reproduced soundfield is investigated. Results show that by incorporating the proposed hybrid model the whole wideband bandwidth is spatial-aliasing free with a mean acoustic contrast consistently above 54.2dB, an improvement of up to 24.2dB from a non-hybrid approach, with as few as 16 dynamic loudspeakers and one parametric loudspeaker.*

6.1 Introduction

As was discussed in previous chapters, a high contrast and high quality multizone soundfield reproduction has many useful real-world applications, such as, vehicle cabin entertainment/communication systems, cinema surround sound systems, multi-participant teleconferencing and personal audio in restaurant/cafés. One large limitation of these reproductions is the loss of acoustic contrast and soundfield reproduction accuracy when spatial aliasing occurs. In this chapter, we investigate the use of an alternative loudspeaker design, known as a parametric loudspeaker (PL) [157], which is capable of reproducing audible sounds at high frequencies with significantly less spatial aliasing artefacts. In chapter 4, we described how to model the spatial aliasing frequency for multizone soundfield reproduction scenarios so that the aliasing artefacts could be suppressed using filters. Rather than suppressing the frequencies where aliasing occurs, in this chapter we consider the use of PLs to reproduce soundfields above the spatial aliasing frequency.

Scenarios where a finite number of loudspeakers are used as secondary sources for soundfield reproduction, are limited to accurate reproduction below a (spatial aliasing) frequency [135]. A fundamental issue with MSR using discrete secondary sources is that the spatial aliasing induces so-called *grating lobes* which can interfere across zones [221] and which we have shown can be accurately modelled. Recent research [106], [109] suggests a full circle array of ≈ 300 loudspeakers are required to reproduce audio up to 8 kHz with high acoustic contrast.

PLs, on the otherhand, are capable of providing high directivity at high frequencies [272] and were first theorised in 1963 [273]. PLs have gained interest due to their high directivity with a relatively small physical size when compared to dynamic (conventional) loudspeakers. Audio is generated from a parametric loudspeaker when the ultrasonic carrier frequency reacts non-linearly in air [274]. The non-linear interaction demodulates an audio signal from the envelope of the modulated carrier wave. Practical implementations have shown PLs can provide immersive spatial audio [275], [276], however, neither of the hybrid approaches use MSR with dynamic loudspeakers or consider spatial aliasing. When comparing PLs to MSR from dynamic loudspeakers, PLs lack directivity at low frequencies [272], contain higher Total Harmonic Distortion (THD) [157], [277] and can have potential health risks due to the high Sound Pressure Level (SPL) of the ultrasonic carrier frequency [157].

A hybrid system utilising the better aspects of both MSRs and PLs would allow for high acoustic contrast at low and high frequencies. Reproduction of speech soundfields would require low carrier SPL in PLs due to the low energy of high frequency components in speech [225], thus reducing related health risks. Further, frequency dependent PL distortions are less of a problem at higher frequencies [277].

In this chapter, novel contributions are made through an analytical approach to a hybrid MSR and PL system with application to personal sound zones. A zone dependent crossover filter is designed to shift the loudspeaker signals between the MSR and PL in the frequency domain. A wideband acoustic contrast is presented for the hybrid system and the trade-off between the acoustic contrast, crossover frequency and reproduced bandwidth is discussed.

Beginning this chapter, in Section 6.2, is an explanation of the MSR layout and soundfield reproduction aliasing. Section 6.3 gives a brief overview of the PL directivity model used in this work. In Section 6.4 a hybrid method is formulated for MSR and PL reproduction of personal sound zones with results and discussion in Section 6.5 and conclusions in Section 6.6.

6.2 Multizone Soundfield Reproduction (MSR)

In this section a general MSR layout is described along with a description of a recent MSR technique. The aliasing which occurs from reproductions with spatial discretisation artifacts is also explained for later use in the hybrid model.

In this work, the acoustical brightness contrast between two zones, $\mathbb{D}_b$ and $\mathbb{D}_q$, is defined as

$$\zeta_{\mathcal{R}}(k) = \frac{\mathfrak{d}_q \int_{\mathbb{D}_b} \left|S^a_{\mathcal{R}}(\mathbf{x},k)\right|^2 d\mathbf{x}}{\mathfrak{d}_b \int_{\mathbb{D}_q} \left|S^a_{\mathcal{R}}(\mathbf{x},k)\right|^2 d\mathbf{x}}, \tag{6.1}$$

where $\mathfrak{d}_b$ and $\mathfrak{d}_q$ are the areas (sizes) of $\mathbb{D}_b$ and $\mathbb{D}_q$, respectively. The mean square error (MSE) between the desired soundfield, $S^d(\mathbf{x},k)$, and the actual reproduced

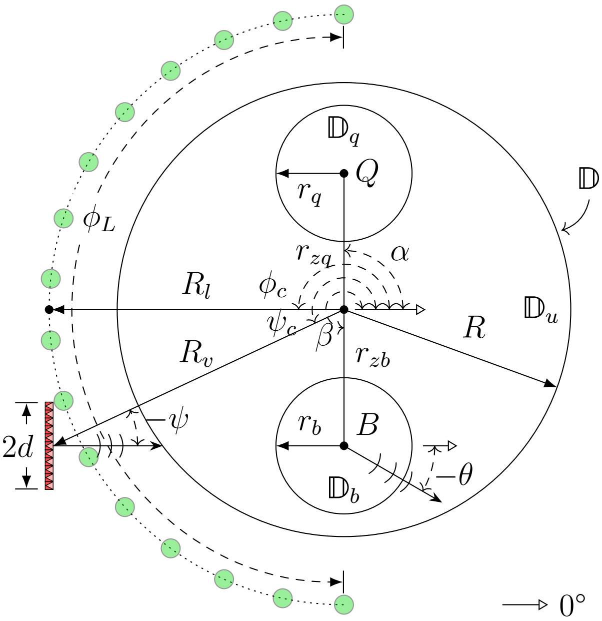

Figure 6.1: MSR layout for a circular loudspeaker array (green) with a companion PL (red) for hybrid soundfield reproduction in $\mathbb{D}_b$.

soundfield, $S^a_{\mathcal{R}}(\mathbf{x}, k)$, is [106], [112]

$$\epsilon_{\mathcal{R}}(k) = \frac{\int_{\mathbb{D}_b} \left| S^d(\mathbf{x}, k) - S^a_{\mathcal{R}}(\mathbf{x}, k) \right|^2 d\mathbf{x}}{\int_{\mathbb{D}_b} \left| S^d(\mathbf{x}, k) \right|^2 d\mathbf{x}}, \tag{6.2}$$

which is used to measure reproduction accuracy. These measures can be used for any actual soundfield, $S^a_{\mathcal{R}}(\mathbf{x}, k)$, created with any reproduction technique, $\mathcal{R}$, such as MSR, PL or any combination thereof.

6.2.1 MSR Layout

The geometry of a generic MSR layout is depicted in Fig. 6.1 for a circular array with a companion PL. An MSR reproduction region, $\mathbb{D}$, of radius R is shown and contains three sub-regions called the bright, quiet and unattended zone, labelled $\mathbb{D}_b$, $\mathbb{D}_q$ and $\mathbb{D}_u = \mathbb{D} \setminus (\mathbb{D}_b \cup \mathbb{D}_q)$, respectively. The centre of $\mathbb{D}$ is the origin from which other geometrical locations are related. The centres of $\mathbb{D}_b$ and $\mathbb{D}_q$ have radius and angle pair polar coordinates (r_{zb}, β) and (r_{zq}, α), respectively. The radius of $\mathbb{D}_b$ and $\mathbb{D}_q$ is r_b and r_q, respectively, and the direction of the soundfield within the regions is θ and

ϑ, respectively. The MSR loudspeaker arc containing L loudspeakers has a centre located at (R_l, ϕ_c) and subtends an angle of ϕ_L. The directional PL has a centre located at (R_v, ψ_c) and is directed at an angle of ψ clockwise from the origin. In practice, the PL is a circular array of transducers, with effective radius d, protruding normal to the reproduction plane. In this work, the imaginary unit is $i = \sqrt{-1}$ and the Euclidean norm is denoted with $\|\cdot\|$. The wavenumber $k = 2\pi f/c$ is interchanged with frequency, f, under the assumption that the speed of sound, c, is constant.

6.2.2 MSR Technique

An infinite set of planewaves arriving from every angle is capable of entirely describing any arbitrary desired soundfield [2]. A soundfield fulfilling the wave equation, in this work, is denoted by the function $S(\mathbf{x}, k)$, where $\mathbf{x} \in \mathbb{D}$ is an arbitrary spatial sampling point. As shown in the orthogonal basis expansion approach [109], [112] to MSR, an additional spatial weighting function, $w(\mathbf{x})$, can be used to set relative importance between zones. The weighted MSR soundfield function used in this work can be written as

$$S(\mathbf{x}, k) = \sum_j P_j(k) F_j(\mathbf{x}, k), \tag{6.3}$$

where the orthogonal wavefields, $F_j(\mathbf{x}, k)$, have coefficients, $P_j(k)$, for a given weighting function and desired soundfield, $S^d(\mathbf{x}, k)$; and $j \in \{1, \dots, J\}$ where J is the number of basis planewaves [109].

The complex loudspeaker weights used to reproduce the soundfield in the (temporal) frequency domain are [109], [143]

$$U_l(k) = \sum_{\bar{m}=-M}^{M} \frac{2\exp(i\bar{m}\phi_l)\Delta\phi_s \sum_j \left(P_j(k) i^{\bar{m}} \exp(-i\bar{m}\rho_j)\right)}{i\pi \mathscr{H}_{\bar{m}}^{(1)}(kR_l)}, \tag{6.4}$$

where $\rho_j = (j-1)\Delta\rho$ are the wavefield angles, $\Delta\rho = 2\pi/J$, ϕ_l is the angle of the l^{th} dynamic loudspeaker from 0°, $\Delta\phi_s$ is the angular spacing of the loudspeakers, $\mathscr{H}_{\nu}^{(1)}(\cdot)$ is a ν^{th}-order Hankel function of the first kind and the modal truncation length [109] is

$$M = \lceil kR \rceil. \tag{6.5}$$

Here, P_j is chosen to minimise the difference between the desired soundfield and the actual soundfield [109].

The actual soundfield from MSR is the result from superposition of all individual loudspeaker responses

$$S^a_{\text{MSR}}(\mathbf{x}, k) = G_{\text{MSR}}(k) \sum_l U_l(k) T(\mathbf{x}, \mathbf{l}_l, k), \tag{6.6}$$

where $G_{\text{MSR}}(k)$ is introduced as an arbitrary weighting for hybrid soundfields (described later in 6.4.1), the loudspeaker's 2-D acoustic transfer function (ATF) is

$$T(\mathbf{x}, \mathbf{l}_l, k) = \frac{i}{4} \mathscr{H}_0^{(1)}(k \|\mathbf{x} - \mathbf{l}_l\|), \tag{6.7}$$

and $\mathbf{l}_l$ is the position of the l^{th} dynamic loudspeaker. Setting $G_{\text{MSR}}(k) = 1$ in (6.6) will render the multizone soundfield.

6.2.3 Soundfield Reproduction Aliasing

A fundamental issue with reproducing soundfields using a limited number of loudspeakers is spatial aliasing which gives rise to grating lobes which may impede the quiet zone at higher frequencies [221]. Due to this phenomenon, the bandwidth of reproducible soundfields with high acoustic contrast (which may be lost above the aliasing frequency) is reduced. For a part-circle array, the minimum number of dynamic loudspeakers to use before aliasing problems begin to occur is given by [106], [109]

$$L \geq \left\lceil \frac{\phi_L (2M + 1)}{2\pi} \right\rceil + 1. \tag{6.8}$$

Substituting (6.5) into (6.8) and rearranging to find an approximation for upper frequency limit $k = k_{\mathrm{u}}$, gives

$$k_{\mathrm{u}} = \frac{2\pi(L-1) - \phi_L}{2R'\phi_L}, \tag{6.9}$$

where, instead of R, R' is used which is the radius of the smallest circle concentric with $\mathbb{D}$ encompassing all zones. The upper frequency from (6.9) agrees with [221] and is dependent on the number of loudspeakers, the reproduction radius and the angle subtending the loudspeaker arc.

6.3 Parametric Loudspeaker (PL)

A few PL directivity models are reviewed in this section as well as common disadvantages of PLs. The disadvantages are discussed in regards to speech soundfields, further motivating the use of a hybrid model for such applications.

6.3.1 Directivity Models

The literature provides a handful of directivity models for PLs which are algorithmic approximations of the demodulated acoustic pressure at different angles. Earlier models include Westervelt's directivity (WD) [273] and product directivity (PD) [274], [278], though, these models do not accurately match measured directivity from a PL. Recently a convolutional directivity (CD) model, used in this work, was proposed [272], [279] utilising both WD and PD which has better correlation to measured directivity.

The actual soundfield reproduced by the PL, where the PL is located at $\mathbf{p}$, is defined in this work as

$$S^{a}_{\mathrm{PL}}(\mathbf{x}, k) = G_{\mathrm{PL}}(k) E(k) \mathcal{D}(\mathbf{x}, k) \frac{\exp(ik\|\mathbf{x} - \mathbf{p}\|)}{4\pi\|\mathbf{x} - \mathbf{p}\|}, \tag{6.10}$$

where $G_{\mathrm{PL}}(k)$ is introduced as an arbitrary weighting for hybrid soundfields (described later in 6.4.1), $\mathcal{D}(\mathbf{x},k)$ is the CD and the directivity coefficient is

$$E(k) = \frac{\tilde{\beta}k^2}{\tilde{\alpha}_s\tilde{\rho}_0 c^2}, \tag{6.11}$$

where $\tilde{\beta}$ is the coefficient of non-linearity, $\tilde{\alpha}_s$ is the sum of the absorption coefficients for both primary frequencies and $\tilde{\rho}_0$ is the density of the medium. Here, we assume sound waves propagate through the medium obeying the free-field Green's function.

The CD is defined as the convolution between the PD and WD with the linear convolution operator, $*$, as [272], [279]

$$\mathcal{D}(\mathbf{x},k) = [\mathcal{D}_{\mathrm{G}}(\mathbf{x},k_{\mathrm{c}})\mathcal{D}_{\mathrm{G}}(\mathbf{x},k_{\mathrm{c}}+k)] * \mathcal{D}_{\mathrm{W}}(\mathbf{x},k), \tag{6.12}$$

where k_{c} is the ultrasonic carrier frequency, $\mathcal{D}_{\mathrm{G}}\left(\mathbf{x},\hat{k}\right)$ is the Gaussian directivity [278]

$$\mathcal{D}_{\mathrm{G}}\left(\mathbf{x},\hat{k}\right) = \exp\left(\left(\frac{i}{2}d\hat{k}\tan\left(\rho_{\mathbf{x}}+\varPsi\right)\right)^2\right), \tag{6.13}$$

where $\rho_{\mathbf{x}}$ is the angle of vector $\mathbf{x}-\mathbf{p}$ from 0°, $\varPsi = (\psi + \psi_c - \pi)$ and WD is [279]

$$\mathcal{D}_{\mathrm{W}}(\mathbf{x},k) = \frac{\tilde{\alpha}_s}{\sqrt{\tilde{\alpha}_s^2 + k^2\tan^4\left(\rho_{\mathbf{x}}+\varPsi\right)}}. \tag{6.14}$$

The far-field PL soundfield can then be found using (6.11) and (6.12) in (6.10) with $G_{\mathrm{PL}}(k) = 1$. However, as k decreases $S^a_{\mathrm{PL}}(\mathbf{x},k)$ approaches that of a point source and $\zeta_{\mathrm{PL}}(k)$ is consequently reduced. It is assumed in this work that the PL is designed such that grating lobes are negligible [280] and for different virtual source locations, multiple steerable PL arrays can be used [276], [280].

6.3.2 PLs for Speech Soundfields

While PLs have been studied extensively over the years there are still some drawbacks when it comes to reproducing loud and clear audible sound. Audible reproductions from PLs are known to require a large carrier SPL (>110 dB) for typical speech conversation levels of ≈ 60 dBA, which has potential inadvertent health risks [157].

Fortunately, for applications of speech soundfields, high SPLs from the PL are not necessary for high frequency ($\gtrapprox 2\,\mathrm{kHz}$) components of speech [225], further, harmonic distortions are lower above this frequency [277]. Taking into account the PL location so that the far-field demodulated audio [281] overlays $\mathbb{D}_b$ and under the assumption that high SPL from the PL is not required over $\mathbb{D}_b$, health risks from the PLs could be argued to be negligible.

6.4 Hybrid Multizone Soundfield Reproduction and Parametric Loudspeaker System

A hybrid MSR and PL system is presented in this section for use in personal sound zone applications. A crossover filter is designed to switch target audio in the (temporal) frequency domain to each of the constituent reproduction techniques.

6.4.1 Crossover Filter Design

Ideally the combination of low and high frequency acoustic contrast from $S^{a}_{\mathrm{MSR}}(\mathbf{x}, k)$ and $S^{a}_{\mathrm{PL}}(\mathbf{x}, k)$, respectively, is desired for personal sound zones. The weightings, $G_{\mathrm{MSR}}(k)$ and $G_{\mathrm{PL}}(k)$, are introduced in (6.6) and (6.10), respectively, in order to facilitate a hybrid soundfield, $S^{a}_{\mathcal{H}}(\mathbf{x}, k)$. When composing a hybrid soundfield it is natural to limit spectral distortion of the reproduction at the crossover frequency, for this, we propose the use of Linkwitz-Riley (LR) filters. Here, a low-pass $\hat{n}^{\mathrm{th}}$ order LR filter with a roll-off of $6\hat{n}$ dB/octave is a cascaded Butterworth filter

$$H^{\urcorner}_{\mathrm{LR}}(k) = \mathcal{B}_{\frac{\hat{n}}{2}}(k/k_{\mathrm{u}})^{-2}, \tag{6.15}$$

where $\mathcal{B}_{\frac{\hat{n}}{2}}$ are Butterworth polynomials of order $\frac{\hat{n}}{2}$ and k_{u} from (6.9) is suggested as the crossover frequency. The matching LR high-pass is

$$H^{\ulcorner}_{\mathrm{LR}}(k) = \mathcal{B}_{\frac{\hat{n}}{2}}(k_{\mathrm{u}}/k)^{-2} \tag{6.16}$$

and together the crossover magnitude response is

$$\left| H^{\urcorner}_{\mathrm{LR}}(k) + H^{\ulcorner}_{\mathrm{LR}}(k) \right| = 1. \tag{6.17}$$

For further definitions and examples of Butterworth filters the reader is referred to [282], [283]. In this work, the arbitrary MSR weighting is set to

$$G_{\mathrm{MSR}}(k) = H^{\urcorner}_{\mathrm{LR}}(k), \tag{6.18}$$

and the arbitrary PL weighting is

$$G_{\mathrm{PL}}(k) = H^{\ulcorner}_{\mathrm{LR}}(k). \tag{6.19}$$

Using the new weights from (6.18) and (6.19) in (6.6) and (6.10), respectively, a hybrid, $\mathcal{H}$, soundfield is defined as the superposition of a set of reproduction methods, $\mathscr{R}$ (in this work the cardinality of $\mathscr{R}$ is 2), as

$$S^a_{\mathcal{H}}(\mathbf{x}, k) = \sum_{\mathcal{R} \in \mathscr{R}} \frac{\mathfrak{d}_b |G_{\mathcal{R}}(k)| S^a_{\mathcal{R}}(\mathbf{x}, k)}{\int_{\mathbb{D}_b} |S^a_{\mathcal{R}}(\mathbf{x}, k)| \, d\mathbf{x}}, \tag{6.20}$$

where each component soundfield is normalised to the mean amplitude over $\mathbb{D}_b$. $\zeta_{\mathcal{R}}(k)$ and $\epsilon_{\mathcal{R}}(k)$ can be evaluated using $S^a_{\mathcal{H}}(\mathbf{x}, k)$ in place of $S^a_{\mathcal{R}}(\mathbf{x}, k)$ in (6.1) and (6.2), respectively.

6.4.2 Loudspeaker Signals

The time domain loudspeaker signals (unmodulated for a PL) are defined in general in this section for the reproduction of speech input signals, $y(n)$. The discrete Fourier transform of the g^{th} overlapping windowed frame of $y(n)$ is $\tilde{Y}_g(k)$. The overlapping

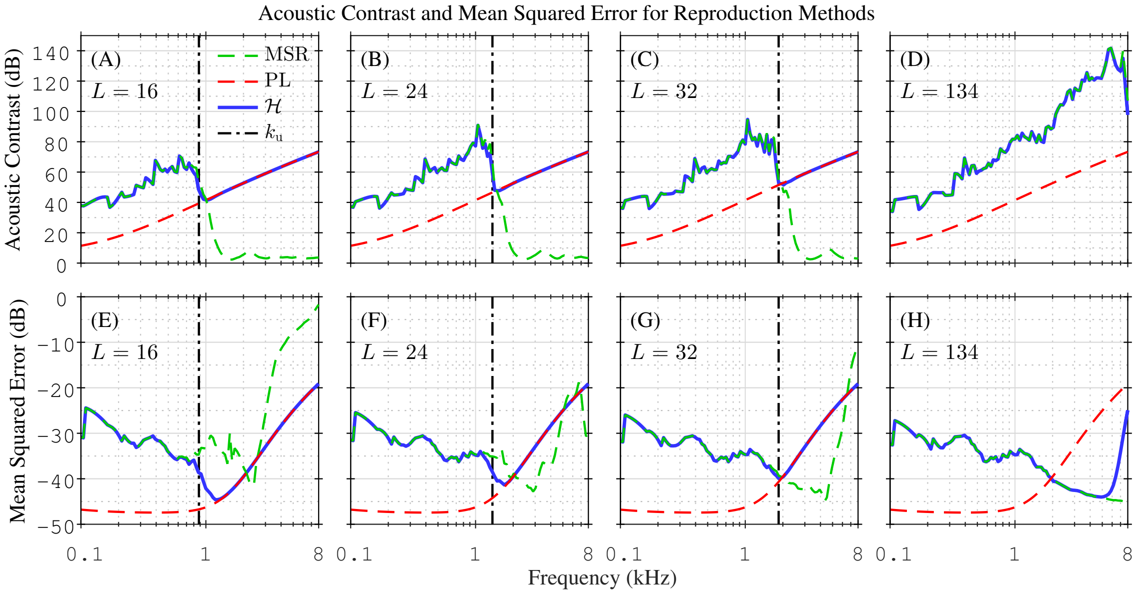

Figure 6.2: Results are shown for three reproduction methods and four values of L. Acoustic contrast results (ζ_{MSR}, ζ_{PL} and $\zeta_{\mathcal{H}}$) are shown in (A)–(D). Mean squared error results (ϵ_{MSR}, ϵ_{PL} and $\epsilon_{\mathcal{H}}$) are shown in (E)–(H). The case where $L = 134$ is alias free up to 8 kHz.

windowed frame of each loudspeaker signal is

$$\tilde{Q}_{\mathcal{R}lg}(k) = \tilde{Y}_g(k)\, G_{\mathcal{R}}(k)\, U_l(k), \tag{6.21}$$

$$\tilde{q}_{\mathcal{R}lg}(n) = \frac{1}{K}\sum_{m=0}^{K-1} \tilde{Q}_{\mathcal{R}lg}(k_m \hat{f})\exp(icnk_m), \tag{6.22}$$

where $k_m \triangleq 2\pi m/cK$, the number of frequencies is K, the maximum frequency is $\hat{f}$ and each loudspeaker signal, $q_{\mathcal{R}l}(n)$, for a particular $\mathcal{R}$, is reconstructed by performing overlap-add reconstruction with the synthesis window on $\tilde{q}_{\mathcal{R}lg}(n)$. For the case where there is a single loudspeaker, $l = \{1\}$, for a given $\mathcal{R}$, such as for the PL in this work, $U_l(k) = 1$ is used.

6.5 Results and Discussion

6.5.1 Experimental Setup

Simulations were carried out using the geometry shown in Figure 6.1 with $r_{zb} = r_{zq} = 0.6\,\mathrm{m}$, $r_b = r_q = 0.3\,\mathrm{m}$, $R = 1.0\,\mathrm{m}$ and $\alpha = \beta/3 = 90°$. The desired soundfield angle was $\theta = 0°$ and in this work $w(\mathbf{x})$ was set to one in $\mathbb{D}_b$, 100 in $\mathbb{D}_q$ and 0.05 in $\mathbb{D}_u$ based on [109], [112]. The target soundfield in $\mathbb{D}_b$ was a virtual point source located at the centre of the PL and $\mathbb{D}_q$ was set to be quiet. The loudspeakers had $R_l = R_v = 1.3\,\mathrm{m}$, $\phi_L = 180°$, $\phi_c = 180°$ and $\psi = \psi_c - 180° = 27.5°$. The speed of sound in air was $c = 343\,\mathrm{m\,s^{-1}}$.

The PL was designed with $k_\mathrm{c} = 2\pi(40\,\mathrm{kHz})/c$, $\tilde{\beta} = 1.2$, $\tilde{\alpha}_s = 2.328\,\mathrm{m^{-1}}$, $\tilde{\rho}_0 = 1.225\,\mathrm{kg\,m^{-3}}$ and $d = 6.18\,\mathrm{cm}$. In this work, it was assumed that the PL had ultrasonic transducers spaced less than 4.3 cm [280] between each other, thus avoiding spatial aliasing.

The LR filters used to reproduce $S^a_{\mathrm{MSR}}(\mathbf{x}, k)$ and $S^a_{\mathrm{PL}}(\mathbf{x}, k)$ had order $\hat{n} = 12$. The number of MSR loudspeakers used was $L = \{16, 24, 32, 134\}$ where k_u was found from (6.9). To compare with MSR, $L = 134$ was chosen to reproduce the speech with no spatial aliasing. The hybrid reproduction method used $\mathscr{R} = \{\mathrm{MSR}, \mathrm{PL}\}$ to find $S^a_{\mathcal{H}}(\mathbf{x}, k)$ using (6.20).

6.5.2 Wideband Spatial Error Reduction

Figure 6.2 shows $\epsilon_{\mathrm{MSR}}(k)$, $\epsilon_{\mathrm{PL}}(k)$ and $\epsilon_{\mathcal{H}}(k)$ computed from (6.2) in (E)–(H) as dashed green, dashed red and solid blue lines, respectively. The crossover frequencies are the vertical dash-dot black lines. Comparing the proposed hybrid approach, it can be seen in Fig. 6.2 that $\epsilon_{\mathcal{H}}$ was on average similar to the aliasing free MSR. Table 6.1 confirms this by showing that, on average, $\epsilon_{\mathcal{H}}$ was slightly less than ϵ_{MSR}. While this was partly due to the low MSE of ϵ_{PL} at lower frequencies, acoustic contrast was also reduced when using a PL at those lower frequencies as seen in Fig. 6.2 (A)–(D). The trade-off between MSE and acoustic contrast is shown in Table 6.1 where $\epsilon_{\mathcal{H}}$

Table 6.1: Wideband mean $\epsilon_{\mathcal{R}}$ and $\zeta_{\mathcal{R}}$ comparisons as a function of the number of dynamic loudspeakers (L) for one PL

	$\epsilon_{\mathcal{R}}$ (dB)			$\zeta_{\mathcal{R}}$ (dB)		
L	MSR	PL	$\mathcal{H}$	MSR	PL	$\mathcal{H}$
16	−27.2	−40.7	−32.5	30.0	40.4	**54.2**
24	−32.7	−40.7	−31.7	38.1	40.4	**58.1**
32	−33.7	−40.7	−31.6	43.5	40.4	**60.3**
134	−36.4	−40.7	−35.6	79.6	40.4	79.3

reduces with L.

6.5.3 Wideband Acoustic Contrast Improvement

Figure 6.2 shows $\zeta_{\text{MSR}}(k)$, $\zeta_{\text{PL}}(k)$ and $\zeta_{\mathcal{H}}(k)$, computed from (6.1), in (A)–(D) as dashed green, dashed red and solid blue lines, respectively. The crossover frequencies are the vertical dash-dot black lines which clearly indicate the point where $\zeta_{\text{MSR}}(k)$ begins to decrease due to spatial aliasing. Note that the *multizone occlusion problem* [25] (should it occur) may be difficult to overcome with one PL, however, the MSR grating lobes interfere less over $\mathbb{D}_q$ during this phenomenon. Also shown in Fig. 6.2 is the limited bandwidth with high acoustic contrast when reducing L. The mean acoustic contrast over the wideband bandwidth for all reproduction techniques is given in Table 6.1 and the mean improvement using the hybrid method can be deduced. While the MSR mean acoustic contrast decreased significantly, from 79.6 dB to 30.0 dB, due to spatial aliasing, the proposed hybrid method decreased to only 54.2 dB. For all reduced loudspeaker cases the hybrid approach outperformed both MSR and PL methods. The maximum improvement was 24.2 dB when $L = 16$ and for all cases the wideband acoustic contrast remained above 54.2 dB, despite the fundamental spatial aliasing that occurred.

6.6 Conclusions and Contributions

This chapter has proposed a hybrid approach to personal sound zones, including speech soundfields. An analytical solution to the combination of MSR and PL soundfields is presented along with a solution to a robust crossover filter. The crossover filter is analytically derived from the geometry of the soundfield layout whilst taking into account spatial aliasing artifacts. Experimental results show that a significant improvement in acoustic contrast from non-hybrid MSR and PL soundfields of 24.2 dB and 19.9 dB, respectively, is achievable. The proposed hybrid method also yields mean wideband acoustic contrast consistently above 54.2 dB with as few as 16 dynamic loudspeakers and a single PL. Some topics for future work are improving speech intelligibility contrast (SIC) and quality in private speech sound zones using hybrid techniques.

Chapter 7

Conclusions

In this thesis, we have addressed several problems that exist with current methods of providing personal sound zones. We have considered computational complexity, psychoacoustic modelling, sound masking, spatial aliasing, active cancellation, dereverberation and alternative loudspeaker designs to solve drawbacks of current methods that are used to provide personal sound. The approaches proposed throughout the thesis have each been evaluated and shown to be highly effective at solving the particular problems at hand.

A common issue with multizone soundfield reproduction is that quiet zones are often over-constrained, to the point where specification of zero energy is, in practice, perceptually unnecessary. Allowing the energy to leak into the quiet is a solution to the problem. However, providing *a priori* estimates of weighted soundfields is computationally demanding but necessary to predetermine the resulting leakage.

In chapter 3, we proposed novel approaches for reducing the computational complexity of synthesising dynamically weighted zones in multizone soundfield reproductions by interpolating sparsely sampled look-up tables. The methods were used to perform dynamic weighting with novel psychoacoustic model implementations, which used the spreading function models of the psychoacoustic frequency masking phenomena to relieve energy constraints on quiet zones. The interpolation method was shown to provide significant computational improvements whilst having little effect on the quality of synthesised soundfields and the dynamic weighting was shown to perform well with large reductions in reproduction error, especially when zones are occluded.

The leakage is, conventionally, uncontrollable above the spatial Nyquist sampling rate and can allow information to leak between spaces as well as reduce reproduction quality. We proposed multizone field metrics for use as optimisation cost functions maximising the effect of noise maskers to provide private and high quality reproduction. The cost functions were complimented with analytical derivations for spatial aliasing and secondary zone leakage, which allowed us to propose filters to control the trade-off between quality and privacy in zones. Simulations and real-world experiments were conducted and showed that the techniques proposed were practically feasible and robust. The real-world implementation provided high quality and confidential multizone soundfield reproduction using a relatively small number of loudspeakers. The implementations were evaluated for both semi-circular and linear loudspeaker arrays, which further justified the feasibility for real-world public systems.

While it is relatively straightforward to reproduce single zone soundfields, it is considerably more difficult to control the influence of external sources, for instance when a third-party talker speaks across an open space. A similar, but not identical, problem is that of wall reflections where echoes can bounce back into rooms reducing the acoustic privacy over the open space. We developed solutions to both of these problems in chapter 5 and showed that there exists a trade-off between soundfield reproduction accuracy and predication accuracy. The trade-off was shown to exist for cases when soundfield filters are not minimum phase and reference microphones are located far from the source that is to be cancelled. By choosing the optimal block length we showed that significant suppression across loudspeaker barriers is possible. We also showed that by using either weighted least square optimised WFS filters or differential source/receiver acoustic models, it is possible to design active acoustic barriers that are capable of suppressing significant sound energy, as well as some passive fibre panels, by using the complete Kirchhoff-Helmholtz integral equation.

After establishing several techniques for a solid framework for the reproduction of personal sound zones in various environmental conditions, the last contribution to the thesis is a novel approach to increase acoustic contrast above the spatial Nyquist frequency. We propose the use of a hybrid ultrasonic parametric and regular loudspeaker setup to be used in a multizone soundfield reproduction system. The ultrasonic parametric array is capable of providing highly directional audio with little to no spatial aliasing grating lobes at frequencies above the spatial Nyquist frequency (aliasing frequency). We show that by using the parametric array, it is possible to reproduce audio content in the targeted bright zone above the aliasing frequency without leaking to the quiet zone. This was shown to dramatically increase the acoustic contrast above the aliasing frequency and reproduction error was shown to be low when using the three dimensional Green's function.

Overall, we can conclude that personal sound zones are now practically feasible even though many challenges still remain in the area. We have shown that it is possible to efficiently implement real-world high quality and private personal sound zones in open and shared public environments.

7.1 Future Research

In this section, we offer suggestions for directions of future work in the area.

Real-time Dynamic Psychoacoustic Zone Weighting

We established methods to dynamically weight zones in multizone soundfield reproductions and techniques to actively cancel sound propagating over barriers. Extending the psychoacoustic weighting to the suppression of soundfields using active control is a promising direction for future work. Currently, active suppression techniques lack the psychoacoustic modelling features that we have described in this thesis. The human hearing models could be applied to the active suppression over loudspeaker barriers, for cancelling speech traversing a room, or applied as a psychoacoustic based reverberation suppression system. This could further reduce the energy required by systems that require a large number of loudspeakers when implemented as three-dimensional arrays. These systems could be further tested using subjective analysis approaches.

Cognitive Performance in 3D Private Sound Zoning Systems

With recent advances in multizone soundfield reproduction for three dimensional sound zones, there is a wealth of opportunity for performance analysis on larger scales, such as restaurants and open offices. The relative subjective comfort of these systems could be investigated, which may include cognitive performance tests to gauge the influence on task completion. The three dimensional implementation of multizone soundfields can be demanding on processing and hardware requirements. With the contributions from this thesis, loudspeaker counts and processing time could be significantly reduced, thus facilitating implementation of large scale systems.

Subjective Analysis of Active Dereverberation Walls and Scattered Soundfields

We have seen through this thesis that active dereverberation is a viable method to providing reflection free soundfields. The dereverberation techniques could be used in conjunction with multizone soundfield reproduction methods to provide personal sound. The subjective opinion of the dereverberation technique could be studied within separate zones. The sound scattering from objects within a reproduction region can also be analysed, from the point of view of the subject and in terms of the suppression performance from the active wall. Where scattering effects and reflections are reduced enough that subjects cannot perceive them, systems could be used for applications such as multilingual cinema and entertainment. Further studies in the area of active dereverberation methods could look at the reduction in reverberation measures, such as RT60, and/or the performance of using multiple active walls in arbitrarily shaped rooms. The use of autoregressive models for the prediction of propagating stationary waves across rooms is also a promising topic for future work.

Further Investigation into Alternative Loudspeaker Designs

The investigation into the performance of ultrasonic parametric loudspeakers in this thesis led to significant improvements in acoustic contrast for reproductions above the spatial aliasing frequency of multizone soundfield reproductions. The potential benefits of using such alternative loudspeaker designs for soundfield reproduction are not fully understood. The non-linearity of large amplitude ultrasound in air has been shown as a good fundamental physical candidate for reproduction where large amplitudes result in open air demodulation. Large channel counts, miniaturised and higher order loudspeaker designs are also excellent directions for future multizone soundfield reproduction studies. A significant limitation of current multizone reproduction is due to spatial aliasing artefacts, such as grating lobes. By making use of the techniques outlined in this thesis, miniaturised higher order loudspeakers (potentially using micro-electro-mechanical systems (MEMS) technology) with large channel counts could result in real-world implementations of soundfield reproduction that do not suffer from spatial aliasing artefacts. Such loudspeaker systems could provide control up to frequencies that are well over the limits of human hearing and could be compared in efficiency with the ultrasonic parametric loudspeaker approach discussed in this thesis.

Bibliography

[1] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

[2] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. Academic Press, 1999.

[3] S. A. Gelfand and H. Levitt, *Hearing: An introduction to psychological and physiological acoustics*. Marcel Dekker New York, 1998.

[4] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[5] F. Dunn, W. M. Hartmann, D. M. Campbell, N. H. Fletcher, and T. Rossing, *Springer handbook of acoustics*. Springer, 2015.

[6] A. D. Blumlein, "Improvements in and relating to sound-transmission, sound-recording and sound reproducing systems.," U.K. Patent 394 325, 1931.

[7] C. Kyriakakis, P. Tsakalides, and T. Holman, "Surrounded by sound," *IEEE Signal Process. Mag.*, vol. 16, no. 1, pp. 55–66, 1999.

[8] S. P. Lipshitz, "Stereo microphone techniques: Are the purists wrong?" *J. Audio Eng. Soc.*, vol. 34, no. 9, pp. 716–744, 1986.

[9] *Multichannel stereophonic sound system with and without accompanying picture*. Int. Telecommun. Union (ITU), ITU-R Rec. BS.775, 1992.

[10] J. Blauert, *Spatial hearing: the psychophysics of human sound localization.* MIT press, 1997.

[11] L. S. Simon, R. Mason, and F. Rumsey, "Localization curves for a regularly-spaced octagon loudspeaker array," in *Audio Eng. Soc. Conv. 127*, Audio Eng. Soc., 2009.

[12] L. S. Simon and R. Mason, "Time and level localization curves for a regularly-spaced octagon loudspeaker array," in *Audio Eng. Soc. Conv. 128*, Audio Eng. Soc., 2010.

[13] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, 2007.

[14] G. Theile and G. Plenge, "Localization of lateral phantom sources," *J. Audio Eng. Soc.*, vol. 25, no. 4, pp. 196–200, 1977.

[15] V. Pulkki, M. Karjalainen, and V. Välimäki, "Localization, coloration, and enhancement of amplitude-panned virtual sources," in *Int. Conf. Spatial Sound Reproduction*, Audio Eng. Soc., 1999.

[16] M. Frank, F. Zotter, and A. Sontacchi, "Localization experiments using different 2d ambisonics decoders," in *Tonmeistertagung - VDT Int. Conv.*, 2008.

[17] G. Martin, W. Woszczyk, J. Corey, and R. Quesnel, "Controlling phantom image focus in a multichannel reproduction system," in *Audio Eng. Soc. Conv. 107*, Audio Eng. Soc., 1999.

[18] S. Braun and M. Frank, "Localization of 3D ambisonic recordings and ambisonic virtual sources," in *Int. Conf. Spatial Audio*, 2011.

[19] G. Theile, "On the localisation in the superimposed soundfield," PhD thesis, Universität Berlin, 1980.

[20] Dolby Laboratories, Inc., "Dolby Digital," Tech. Rep., 2017. [Online]. Available: `https://www.dolby.com` (visited on 12/01/2017).

[21] Xperi Corporation, "DTS (Dedicated To Sound)," Tech. Rep., 2017. [Online]. Available: `http://dts.com/` (visited on 12/01/2017).

[22] Dolby Laboratories, Inc., "Dolby Atmos," Tech. Rep., 2017. [Online]. Available: `https://www.dolby.com/us/en/brands/dolby-atmos.html` (visited on 12/01/2017).

[23] Xperi Corporation, "DTS:X," Tech. Rep., 2017. [Online]. Available: `http://dts.com/dtsx` (visited on 12/01/2017).

[24] W. F. Druyvesteyn and J. Garas, "Personal sound," *J. Audio Eng. Soc.*, vol. 45, no. 9, pp. 685–701, 1997.

[25] T. Betlehem, W. Zhang, M. Poletti, and T. D. Abhayapala, "Personal Sound Zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 81–91, Mar. 2015, doi: `10.1109/MSP.2014.2360707`.

[26] G. W. Gardner, *3-D Audio Using Loudspeakers*. Springer, 1998.

[27] R. H. Gilkey, B. D. Simpson, S. K. Isabelle, A. J. Kordik, and J. M. Weisenberger, "Audition and the sense of presence in virtual environments," *J. Acoust. Soc. Am.*, vol. 105, no. 2, pp. 1163–1164, 1999.

[28] C. Schissler, A. Nicholls, and R. Mehra, "Efficient HRTF-based spatial audio for area and volumetric sources," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 4, pp. 1356–1366, 2016.

[29] C. Schissler, P. Stirling, and R. Mehra, "Efficient construction of the spatial room impulse response," in *Virtual Reality (VR)*, IEEE, 2017, pp. 122–130.

[30] S. J. Elliott, J. Cheer, H. Murfet, and K. R. Holland, "Minimally radiating arrays for mobile devices," in *Int. Congr. Sound and Vibration (ICSV16)*, Kraków, Poland, Jul. 2009, pp. 1–7.

[31] S. J. Elliott, J. Cheer, H. Murfet, and K. R. Holland, "Minimally radiating sources for personal audio," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 1721–1728, 2010.

[32] E. Jovanov, K. Wegner, V. Radivojevic, D. Starcevic, M. S. Quinn, and D. B. Karron, “Tactical audio and acoustic rendering in biomedical applications,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 3, no. 2, pp. 109–118, 1999.

[33] S. A. Salehin and T. D. Abhayapala, “Localizing lung sounds: Eigen basis decomposition for localizing sources within a circular array of sensors,” *J. Signal Process. Syst.*, vol. 64, no. 2, pp. 205–221, 2011.

[34] M. M. Boone and W. P. de Bruijn, “Improving speech intelligibility in teleconferencing by using wave field synthesis,” in *Audio Eng. Soc. Conv. 114*, Audio Eng. Soc., 2003.

[35] T. J. Sutton, S. J. Elliott, A. M. McDonald, and T. J. Saunders, “Active control of road noise inside vehicles,” *Noise Control Eng. J.*, vol. 42, no. 4, 1994.

[36] J. Cheer, S. J. Elliott, and M. F. S. Gálvez, “Design and implementation of a car cabin personal audio system,” *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 412–424, 2013.

[37] P. N. Samarasinghe, W. Zhang, and T. D. Abhayapala, “Recent advances in active noise control inside automobile cabins: Toward quieter cars,” *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 61–73, Nov. 2016, doi: `10.1109/MSP.2016.2601942`.

[38] A. W. Peterson and S. V. Tsynkov, “Active control of sound for composite regions,” *SIAM J. Appl. Math.*, vol. 67, no. 6, pp. 1582–1609, Jan. 2007.

[39] Y. Kajikawa, W.-S. Gan, and S. M. Kuo, “Recent advances on active noise control: Open issues and innovative applications,” *APSIPA Trans. Signal Inform. Process.*, vol. 1, pp. 1–21, 2012.

[40] C. Hansen, S. Snyder, X. Qiu, L. Brooks, and D. Moreau, *Active Control of Noise and Vibration, Second Edition.* CRC Press, 2012.

[41] H. Chen, “Theory and design of spatial active noise control systems,” PhD thesis, Australian National University, 2017.

[42] W. J. Trott, "Underwater-sound-transducer calibration from nearfield data," *J. Acoust. Soc. Am.*, vol. 36, no. 8, pp. 1557–1568, Aug. 1964.

[43] O. Kirkeby and P. A. Nelson, "Reproduction of plane wave sound fields," *J. Acoust. Soc. Am.*, vol. 94, no. 5, pp. 2992–3000, 1993.

[44] O. Kirkeby, P. A. Nelson, F. Orduna-Bustamante, and H. Hamada, "Local sound field reproduction using digital signal processing," *J. Acoust. Soc. of Am.*, vol. 100, no. 3, pp. 1584–1593, 1996.

[45] F. M. Fazi and P. A. Nelson, "A theoretical study of sound field reconstruction techniques," pp. 1–6, 2007.

[46] P.-A. Gauthier and A. Berry, "Adaptive wave field synthesis for sound field reproduction: Theory, experiments, and future perspectives," in *Audio Eng. Soc. Conv. 123*, Audio Eng. Soc., 2007.

[47] M. Poletti, "Robust two-dimensional surround sound reproduction for nonuniform loudspeaker layouts," *J. Audio Eng. Soc.*, vol. 55, no. 7/8, pp. 598–610, 2007.

[48] F. M. Fazi and P. A. Nelson, "The ill-conditioning problem in sound field reconstruction," in *Audio Eng. Soc. Conv. 123*, Audio Eng. Soc., 2007.

[49] M. Kolundzija, C. Faller, and M. Vetterli, "Sound field reconstruction: An improved approach for wave field synthesis," in *Audio Eng. Soc. Conv. 126*, Audio Eng. Soc., 2009.

[50] M. Kolundzija, C. Faller, and M. Vetterli, "Reproducing Sound Fields Using MIMO Acoustic Channel Inversion," *J. Audio Eng. Soc.*, vol. 59, pp. 721–734, 2011.

[51] D. S. Talagala, W. Zhang, and T. D. Abhayapala, "Efficient multi-channel adaptive room compensation for spatial soundfield reproduction using a modal decomposition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1522–1532, Oct. 2014.

[52] A. J. Berkhout, “A holographic approach to acoustic control,” *J. Audio Eng. Soc.*, vol. 36, no. 12, pp. 977–995, 1988.

[53] A. J. Berkhout, “Wave-front synthesis: A new direction in electroacoustics,” *J. Acoust. Soc. Am.*, vol. 92, no. 4, pp. 2396–2396, 1992.

[54] A. J. Berkhout, D. de Vries, and P. Vogel, “Acoustic control by wave field synthesis,” *J. Acoust. Soc. of Am.*, vol. 93, no. 5, pp. 2764–2778, 1993.

[55] M. M. Boone, E. N. Verheijen, and P. F. Van Tol, “Spatial sound-field reproduction by wave-field synthesis,” *J. Audio Eng. Soc.*, vol. 43, no. 12, pp. 1003–1012, 1995.

[56] D. De Vries, “Sound reinforcement by wavefield synthesis: Adaptation of the synthesis operator to the loudspeaker directivity characteristics,” *J. Audio Eng. Soc.*, vol. 44, no. 12, pp. 1120–1131, 1996.

[57] D. de Vries and M. M. Boone, “Wave field synthesis and analysis using array technology,” in *Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, IEEE, 1999, pp. 15–18.

[58] S. Spors and R. Rabenstein, “Spatial aliasing artifacts produced by linear and circular loudspeaker arrays used for wave field synthesis,” in *Audio Eng. Soc. Conv. 120*, Audio Eng. Soc., 2006.

[59] S. Spors and J. Ahrens, “Comparison of higher-order ambisonics and wave field synthesis with respect to spatial aliasing artifacts,” in *Int. Congr. Acoust.*, Madrid, Spain, 2007, pp. 1–6.

[60] S. Spors, R. Rabenstein, and J. Ahrens, “The theory of wave field synthesis revisited,” in *Audio Eng. Soc. Conv. 124*, Audio Eng. Soc., 2008, pp. 17–20.

[61] J. Ahrens and S. Spors, “Sound field reproduction using planar and linear arrays of loudspeakers,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2038–2050, 2010.

[62] T. Okamoto, S. Enomoto, and R. Nishimura, “Least squares approach in wavenumber domain for sound field recording and reproduction using multiple parallel linear arrays,” *Appl. Acoust.*, vol. 86, pp. 95–103, Dec. 2014.

[63] J. Ahrens and S. Spors, “Reproduction of a plane-wave sound field using planar and linear arrays of loudspeakers,” in *Int. Symp. Commun., Control Signal Process. (ISCCSP)*, IEEE, 2008, pp. 1486–1491.

[64] S. Spors and J. Ahrens, “Analysis and improvement of pre-equalization in 2.5-dimensional wave field synthesis,” in *Audio Eng. Soc. Conv. 128*, Audio Eng. Soc., 2010.

[65] M. A. Gerzon, “Periphony: With-height sound reproduction,” *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.

[66] M. A. Gerzon, “Practical periphony: The reproduction of full-sphere sound,” in *Audio Eng. Soc. Conv. 65*, Audio Eng. Soc., 1980.

[67] J. S. Bamford, “An analysis of ambisonic sound systems of first and second order,” PhD thesis, University of Waterloo, 1995.

[68] D. B. Ward and T. D. Abhayapala, “Reproduction of a plane-wave sound field using an array of loudspeakers,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 697–707, 2001, doi: `10.1109/89.943347`.

[69] J. Daniel, S. Moreau, and R. Nicol, “Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging,” in *Audio Eng. Soc. Conv. 114*, Audio Eng. Soc., 2003.

[70] J. Daniel and S. Moreau, “Further study of sound field coding with higher order ambisonics,” in *Audio Eng. Soc. Conv. 116*, Audio Eng. Soc., 2004.

[71] M. A. Poletti, “Three-dimensional surround sound systems based on spherical harmonics,” *J. Audio Eng. Soc.*, vol. 53, no. 11, pp. 1004–1025, Nov. 2005.

[72] S. Moreau, J. Daniel, and S. Bertet, “3D sound field recording with higher order ambisonics–objective measurements and validation of a 4th order spherical microphone,” in *Audio Eng. Soc. Conv. 120*, Audio Eng. Soc., 2006.

[73] R. A. Kennedy, P. Sadeghi, T. D. Abhayapala, and H. M. Jones, “Intrinsic limits of dimensionality and richness in random multipath fields,” *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2542–2556, Jun. 2007.

[74] J. Ahrens and S. Spors, “An analytical approach to sound field reproduction using circular and spherical loudspeaker distributions,” *Acta Acust. Utd. with Acust.*, vol. 94, no. 6, pp. 988–999, Nov. 2008.

[75] J. Ahrens, *Analytic Methods of Sound Field Synthesis.* Springer Science & Business Media, 2012.

[76] B. D. Van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

[77] E. Mabande and W. Kellermann, “Towards superdirective beamforming with loudspeaker arrays,” in *Int. Congr. Acoust.*, Madrid, Spain, Sep. 2007.

[78] H. Morgenstern and B. Rafaely, “Spherical loudspeaker array beamforming in enclosed sound fields by MIMO optimization,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2013, pp. 370–374.

[79] I. Tashev, “Personal Audio Space,” Tech. Rep., 2017. [Online]. Available: `https://www.microsoft.com/en-us/research/people/ivantash/` (visited on 12/01/2017).

[80] Microsoft Research, “Personal audio space: The headphones experience sans headphones,” Tech. Rep., 2007. [Online]. Available: `https://www.microsoft.com/en-us/research/blog/personal-audio-space-headphones-experience-sans-headphones/` (visited on 12/01/2017).

[81] K. Helwani, S. Spors, and H. Buchner, “Spatio-temporal signal preprocessing for multichannel acoustic echo cancellation,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2011, pp. 93–96.

[82] T. Okamoto, “Generation of multiple sound zones by spatial filtering in wavenumber domain using a linear array of loudspeakers,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2014, pp. 4733–4737.

[83] T. Okamoto, “Near-field sound propagation based on a circular and linear array combination,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2015, pp. 624–628.

[84] T. Okamoto and A. Sakaguchi, “Experimental validation of spatial fourier transform-based multiple sound zone generation with a linear loudspeaker array,” *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. 1769–1780, Mar. 2017.

[85] J.-W. Choi and Y.-H. Kim, “Generation of an acoustically bright zone with an illuminated region using multiple sources,” *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1695–1700, Apr. 2002, doi: `10.1121/1.1456926`.

[86] J.-H. Chang, C.-H. Lee, J.-Y. Park, and Y.-H. Kim, “A realization of sound focused personal audio system using acoustic contrast control,” *J. Acoust. Soc. Am.*, vol. 125, no. 4, pp. 2091–2097, Apr. 2009, doi: `10.1121/1.3082114`.

[87] M. Shin *et al.*, “Maximization of acoustic energy difference between two spaces,” *J. Acoust. Soc. of Am.*, vol. 128, no. 1, p. 121, Jul. 2010, doi: `10.1121/1.3438479`.

[88] S. J. Elliott, J. Cheer, J.-W. Choi, and Y. Kim, “Robustness and regularization of personal audio systems,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 2123–2133, Sep. 2012.

[89] N. Radmanesh and I. S. Burnett, “Reproduction of independent narrowband soundfields in a multizone surround system and its extension to speech signal sources,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2011, pp. 461–464.

[90] N. Radmanesh and I. S. Burnett, “Wideband sound reproduction in a 2d multizone system using a combined two-stage lasso-LS algorithm,” in *Sensor Array and Multichannel Signal Process. Workshop (SAM)*, IEEE, 2012, pp. 453–456.

[91] N. Radmanesh and I. S. Burnett, “Effectiveness of horizontal personal sound systems for listeners of variable heights,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2013, pp. 316–320.

[92] N. Radmanesh and I. S. Burnett, "Generation of isolated wideband sound fields using a combined two-stage lasso-LS algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 378–387, Feb. 2013, doi: `10.1109/TASL.2012.2227736`.

[93] N. Radmanesh, "Multizone wideband sound field reproduction," PhD thesis, Royal Melbourne Institute of Technology, 2013.

[94] N. Radmanesh, I. S. Burnett, and B. D. Rao, "A lasso-LS optimization with a frequency variable dictionary in a multizone sound system," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 583–593, Mar. 2016.

[95] P. J. Jackson, F. Jacobsen, P. Coleman, and J. Abildgaard Pedersen, "Sound field planarity characterized by superdirective beamforming," in *Proc. Meetings Acoust.*, vol. 19, Acoust. Soc. Am., 2013, p. 055 056.

[96] P. Coleman, P. Jackson, M. Olik, and J. A. Pedersen, "Optimizing the planarity of sound zones," in *Int. Conf. Sound Field Control*, Audio Eng. Soc., Sep. 2013.

[97] P. Coleman, P. J. Jackson, M. Olik, and J. A. Pedersen, "Numerical optimization of loudspeaker configuration for sound zone reproduction," in *Int. Congr. Sound and Vibration*, IIAV, 2014, pp. 1–8.

[98] P. Coleman, P. J. Jackson, M. Olik, M. Møller, M. Olsen, and J. Abildgaard Pedersen, "Acoustic contrast, planarity and robustness of sound zone methods using a circular loudspeaker array," *J. Acoust. Soc. Am.*, vol. 135, no. 4, pp. 1929–1940, Apr. 2014, doi: `10.1121/1.4866442`.

[99] P. Coleman, P. Jackson, M. Olik, and J. A. Pedersen, "Personal audio with a planar bright zone," *J. Acoust. Soc. Am.*, vol. 136, no. 4, pp. 1725–1735, Oct. 2014, doi: `10.1121/1.4893909`.

[100] K. Baykaner *et al.*, "The relationship between target quality and interference in sound zone," *J. Audio Eng. Soc.*, vol. 63, no. 1/2, pp. 78–89, Jan. 2015, doi: `10.17743/jaes.2015.0007`.

[101] P. Coleman, "Loudspeaker array processing for personal sound zone reproduction," PhD thesis, University of Surrey, 2014.

[102] T. Abhayapala and Y. J. Wu, "Spatial soundfield reproduction with zones of quiet," in *Audio Eng. Soc. Conv. 127*, Audio Eng. Soc., Oct. 2009.

[103] Y. J. Wu and T. D. Abhayapala, "Spatial multizone soundfield reproduction," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2009, pp. 93–96.

[104] Y. J. Wu and T. D. Abhayapala, "Multizone 2d soundfield reproduction via spatial band stop filters," in *Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, IEEE, 2009, pp. 309–312.

[105] Y. J. Wu and T. D. Abhayapala, "Simultaneous soundfield reproduction at multiple spatial regions," in *Audio Eng. Soc. Conv. 128*, Audio Eng. Soc., 2010.

[106] Y. J. Wu and T. D. Abhayapala, "Spatial multizone soundfield reproduction: Theory and design," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1711–1720, Aug. 2011, doi: 10.1109/TASL.2010.2097249.

[107] H. Chen, T. D. Abhayapala, and W. Zhang, "Enhanced sound field reproduction within prioritized control region," in *INTER-NOISE and NOISE-CON Congr. and Conf. Proc.*, vol. 249, Inst. of Noise Control Eng., 2014, pp. 4055–4064.

[108] T. Okamoto, "Angular spectrum decomposition-based 2.5d higher-order spherical harmonic sound field synthesis with a linear loudspeaker array," in *Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New York, USA: IEEE, Oct. 2017, pp. 180–184.

[109] W. Jin, W. B. Kleijn, and D. Virette, "Multizone soundfield reproduction using orthogonal basis expansion," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2013, pp. 311–315.

[110] W. Jin and W. B. Kleijn, "Multizone soundfield reproduction in reverberant rooms using compressed sensing techniques," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2014, pp. 4728–4732.

[111] "Audio rendering system," U.S. Patent WO2014082683 A1, Jun. 2014.

[112] W. Jin and W. B. Kleijn, "Theory and design of multizone soundfield reproduction using sparse methods," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2343–2355, Dec. 2015, doi: `10.1109/TASLP.2015.2479037`.

[113] W. Jin, "Spatial multizone soundfield reproduction design," PhD thesis, Victoria University of Wellington, 2015.

[114] W. Jin, "Adaptive reverberation cancelation for multizone soundfield reproduction using sparse methods," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2016, pp. 509–513.

[115] J. Donley and C. Ritz, "An efficient approach to dynamically weighted multizone wideband reproduction of speech soundfields," in *China Summit Int. Conf. Signal Inform. Process. (ChinaSIP)*, IEEE, Jul. 2015, pp. 60–64.

[116] J. Donley and C. Ritz, "Multizone reproduction of speech soundfields: A perceptually weighted approach," in *Asia-Pacific Signal Inform. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, IEEE, 2015, pp. 342–345.

[117] J. Donley, C. Ritz, and W. B. Kleijn, "Improving speech privacy in personal sound zones," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2016, pp. 311–315.

[118] J. Donley, C. Ritz, and W. B. Kleijn, "Multizone soundfield reproduction with privacy- and quality-based speech masking filters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1041–1055, 2018.

[119] J. Donley, C. Ritz, and W. B. Kleijn, "Active speech control using wave-domain processing with a linear wall of dipole secondary sources," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2017, pp. 1–5.

[120] J. Donley, C. Ritz, and W. B. Kleijn, “On the comparison of two room compensation / dereverberation methods employing active acoustic boundary absorption,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, Apr. 2018, pp. 221–225.

[121] J. Donley, C. Ritz, and W. B. Kleijn, “Reproducing personal sound zones using a hybrid synthesis of dynamic and parametric loudspeakers,” in *Asia-Pacific Signal & Inform. Process. Assoc. Annu. Summit and Conf. (APSIPA ASC)*, IEEE, Dec. 2016, pp. 1–5.

[122] P. M. Morse and K. U. Ingard, *Theoretical acoustics.* Princeton university press, 1968.

[123] J. W. S. B. Rayleigh, *The theory of sound.* Macmillan, 1896, vol. 2.

[124] E. W. Start, “Direct sound enhancement by wave field synthesis,” PhD thesis, Delft University of Technology, 1997.

[125] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. of Am.*, vol. 65, pp. 943–950, 1979.

[126] D. M. Leakey, “Some measurements on the effects of interchannel intensity and time differences in two channel sound systems,” *J. Acoust. Soc. Am.*, vol. 31, no. 7, pp. 977–986, 1959.

[127] J. Vanderkooy and S. P. Lipshitz, “Anomalies of wavefront reconstruction in stereo and surround-sound reproduction,” in *Audio Eng. Soc. Conv. 83*, Audio Eng. Soc., 1987.

[128] V. Pulkki, “Virtual sound source positioning using vector base amplitude panning,” *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.

[129] G. W. Stewart, “On the perturbation of pseudo-inverses, projections and linear least squares problems,” *SIAM Review*, vol. 19, no. 4, pp. 634–662, 1977.

[130] G. Deschamps and H. Cabayan, "Antenna synthesis and solution of inverse problems by regularization methods," *IEEE Trans. Antennas Propag.*, vol. 20, no. 3, pp. 268–274, 1972.

[131] D. Colton and R. Kress, *Inverse acoustic and electromagnetic scattering theory.* Springer Science & Business Media, 2012, vol. 93.

[132] G. H. Golub, P. C. Hansen, and D. P. O'Leary, "Tikhonov regularization and total least squares," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 1, pp. 185–194, 1999.

[133] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of acoustics.* Wiley, 1999.

[134] D. Colton and R. Kress, *Integral equation methods in scattering theory.* SIAM, 2013.

[135] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: A review of the current state," *Proc. IEEE*, vol. 101, no. 9, pp. 1920–1938, 2013.

[136] M. M. Boone, E. N. Verheijen, and P. F. Van Tol, "The wave-field synthesis concept applied to sound reproduction," in *Audio Eng. Soc. Conv. 96*, Audio Eng. Soc., 1994.

[137] E. W. Stuart, "Application of curved arrays in wave field synthesis," in *Audio Eng. Soc. Conv. 100*, Audio Eng. Soc., 1996.

[138] E. N. G. Verheijen, "Sound reproduction by wave field synthesis," PhD thesis, Delft University of Technology, 1998.

[139] M. M. Boone, "Multi-actuator panels (MAPs) as loudspeaker arrays for wave field synthesis," *J. Audio Eng. Soc.*, vol. 52, no. 7, pp. 712–723, 2004.

[140] E. Corteel, "Equalization in an extended area using multichannel inversion and wave field synthesis," *J. Audio Eng. Soc.*, vol. 54, no. 12, pp. 1140–1161, 2006.

[141] S. Spors and J. Ahrens, “Spatial sampling artifacts of wave field synthesis for the reproduction of virtual point sources,” in *Audio Eng. Soc. Conv. 126*, Audio Eng. Soc., 2009.

[142] S. Spors and J. Ahrens, “Reproduction of focused sources by the spectral division method,” in *Int. Symp. Commun., Control Signal Process. (ISCCSP)*, IEEE, 2010, pp. 1–5.

[143] Y. J. Wu and T. D. Abhayapala, “Theory and design of soundfield reproduction using continuous loudspeaker concept,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 107–116, Jan. 2009, doi: `10.1109/TASL.2008.2005340`.

[144] J. Ahrens and S. Spors, “Analytical driving functions for higher order ambisonics,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2008, pp. 373–376.

[145] J. Francombe *et al.*, “Perceptually optimized loudspeaker selection for the creation of personal sound zones,” in *Int. Conf. Sound Field Control*, Audio Eng. Soc., 2013.

[146] Y. Cai, M. Wu, and J. Yang, “Sound reproduction in personal audio systems using the least-squares approach with acoustic contrast control constraint,” *J. Acoust. Soc. Am.*, vol. 135, no. 2, pp. 734–741, Feb. 2014.

[147] M. A. Poletti and F. M. Fazi, “An approach to generating two zones of silence with application to personal sound systems,” *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. 598–605, 2015.

[148] M. F. Simon Galvez, S. J. Elliott, and J. Cheer, “Time domain optimization of filters used in a loudspeaker array for personal audio,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 11, pp. 1869–1878, Nov. 2015.

[149] M. A. Poletti and F. M. Fazi, “Generation of half-space sound fields with application to personal sound systems,” *J. Acoust. Soc. Am.*, vol. 139, no. 3, pp. 1294–1302, 2016.

[150] X. Ma, P. J. Hegarty, J. A. Pedersen, L. G. Johansen, and J. J. Larsen, "Personal sound zones: The significance of loudspeaker driver nonlinear distortion," in *Int. Conf. Sound Field Control*, Audio Eng. Soc., 2016.

[151] X. Ma, P. J. Hegarty, J. A. Pedersen, L. G. Johansen, and J. J. Larsen, "Assessing the influence of loudspeaker driver nonlinear distortion on personal sound zones," in *Audio Eng. Soc. Conv. 142*, Audio Eng. Soc., 2017.

[152] J. Rämö, L. Christensen, S. Bech, and S. Jensen, "Validating a perceptual distraction model using a personal two-zone sound system," in *Proc. Meetings Acoust.*, vol. 30, 2017, p. 050 003.

[153] S.-M. Kerr, C. Gibson, and N. Klocker, "Parenting and neighbouring in the consolidating city: The emotional geographies of sound in apartments," *Emotion, Space and Society*, vol. 26, pp. 1–8, Feb. 2018.

[154] E. Manor, W. Martens, A. Marui, and D. Cabrera, "Nearfield crosstalk increases listener preferences for headphone-reproduced stereophonic imagery," *J. Audio Eng. Soc.*, vol. 63, no. 5, pp. 324–335, 2015.

[155] W.-H. Cho, M. Boone, J.-G. Ih, and T. Toi, "Control of the beamwidth of a beamformer with a fixed array configuration," in *Audio Eng. Soc. Conv. 130*, Audio Eng. Soc., 2011.

[156] F. J. Pompei, "The use of airborne ultrasonics for generating audible sound beams," *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 726–731, Sep. 1999.

[157] W.-S. Gan, J. Yang, and T. Kamakura, "A review of parametric acoustic array in air," *Appl. Acoust.*, vol. 73, no. 12, pp. 1211–1219, Dec. 2012.

[158] M. Poletti, "An investigation of 2-D multizone surround sound systems," in *Audio Eng. Soc. Conv. 125*, Audio Eng. Soc., 2008.

[159] M.-f. Zha, C.-c. Bao, M.-s. Jia, and B. Bu, "3D multizone soundfield reproduction using spherical harmonic analysis," in *China Summit Int. Conf. Signal Inform. Process. (ChinaSIP)*, IEEE, Jul. 2015, pp. 625–629.

[160] *Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3: Audio.* Int. Standard ISO/IEC 11172-3, 1993.

[161] *Generic coding of moving pictures and associated audio information – Part 3: Audio.* Int. Standard ISO/IEC 13818-3, 1998.

[162] *Digital compression and coding of continuous-tone still images - Requirements and guidelines.* Int. Telecommun. Union (ITU), ITU-T Rec. T.81, 1992.

[163] *Digital compression and coding of continuous-tone still images: Requirements and guidelines.* Int. Standard ISO/IEC 10918-1, 1994.

[164] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards.* Springer, 2003.

[165] J. B. Allen, "Nonlinear cochlear signal processing and masking in speech perception," in *Springer Handbook of Speech Processing*, Springer, 2008, pp. 27–60.

[166] *Acoustics – Normal Equal-Loudness-Level Contours.* Int. Standard ISO 226, 2003.

[167] H. Fletcher, *Speech and hearing in communication*, 2nd. 1953.

[168] B. Kollmeier, T. Brand, and B. Meyer, "Perception of speech and sound," in *Springer handbook of speech processing*, Springer, 2008, pp. 61–82.

[169] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[170] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Am.*, vol. 66, no. 6, pp. 1647–1652, Dec. 1979.

[171] M. R. Schroeder, "Recognition of complex acoustic signals," *Life Sciences Research Report*, vol. 5, no. 324, p. 130, 1977.

[172] E. Zwicker, "Über die lautheit von ungedrosselten und gedrosselten schallen," *Acustica*, vol. 13, no. 3, pp. 194–211, 1963.

[173] V. Grancharov and W. B. Kleijn, "Speech quality assessment," in *Springer Handbook of Speech Processing*, Springer, 2008, pp. 83–100.

[174] S. R. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality.* Prentice Hall, 1988.

[175] N. S. Jayant and P. Noll, *Digital coding of waveforms: principles and applications to speech and video.* Prentice Hall, 1984.

[176] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 2, pp. 242–248, Feb. 1988.

[177] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2001, pp. 749–752.

[178] *Perceptual evaluation of speech quality (PESQ).* Int. Telecommun. Union (ITU), ITU-T Rec. P.862, 2003.

[179] *Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs.* Int. Telecommun. Union (ITU), ITU-T Rec. P.862.2, 2003.

[180] *Perceptual objective listening quality assessment (POLQA).* Int. Telecommun. Union (ITU), ITU-T Rec. P.863, 2014.

[181] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, Sep. 1953.

[182] J. D. Johnston and Y. H. V. Lam, "Perceptual soundfield reconstruction," in *Audio Eng. Soc. Conv. 109*, Audio Eng. Soc., 2000.

[183] E. De Sena, H. Hacıhabiboğlu, and Z. Cvetković, "Analysis and design of multichannel systems for perceptual sound field reconstruction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 8, pp. 1653–1665, 2013.

[184] J. Hannemann, C. A. Leedy, K. D. Donohue, S. Spors, and A. Raake, "A comparative study of perceptional quality between wavefield synthesis and multipole-matched rendering for spatial audio," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2008, pp. 397–400.

[185] A. J. Tucker, W. L. Martens, G. Dickens, and M. P. Hollier, "Perception of reconstructed sound-fields: The dirty little secret," in *Int. Conf. Sound Field Control*, Audio Eng. Soc., 2013.

[186] H. Wierstorf, A. Raake, M. Geier, and S. Spors, "Perception of focused sources in wave field synthesis," *J. Audio Eng. Soc.*, vol. 61, no. 1, pp. 5–16, 2013.

[187] M. Schoeffler, A. Silzle, and J. Herre, "Evaluation of spatial/3D audio: Basic audio quality versus quality of experience," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 75–88, Feb. 2017.

[188] W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. Petkov, B. Sauert, and P. Vary, "Optimizing speech intelligibility in a noisy environment: A unified view," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 43–54, Mar. 2015, doi: `10.1109/MSP.2014.2365594`.

[189] P. N. Petkov, G. E. Henter, and W. B. Kleijn, "Maximizing phoneme recognition accuracy for enhanced speech intelligibility in noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 1035–1045, 2013.

[190] M. Zhang, P. N. Petkov, and W. B. Kleijn, "Rephrasing-based speech intelligibility enhancement.," in *Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2013, pp. 3587–3591.

[191] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective"signal processing in the auditory system. i. model structure," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615–3622, 1996.

[192] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A low-complexity spectro-temporal distortion measure for audio processing applications," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1553–1564, 2012.

[193] N. R. French and J. C. Steinberg, “Factors governing the intelligibility of speech sounds,” *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[194] K. D. Kryter, “Methods for the calculation and use of the articulation index,” *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1689–1697, 1962.

[195] G. A. Studebaker, C. V. Pavlovic, and R. L. Sherbecoe, “A frequency importance function for continuous discourse,” *J. Acoust. Soc. Am.*, vol. 81, no. 4, pp. 1130–1138, 1987.

[196] J. B. Allen, “How do humans process and recognize speech?” *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 567–577, 1994.

[197] *Methods for Calculation of the Speech Intelligibility Index.* Amer. Nat. Standards Inst. S3.5, 1997.

[198] K. S. Rhebergen and N. J. Versfeld, “A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners,” *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[199] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011, doi: `10.1109/TASL.2011.2114881`.

[200] J. Jensen and C. H. Taal, “An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[201] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, “An evaluation of intrusive instrumental intelligibility metrics,” *arXiv*, 2017.

[202] S. V. Kuyk, W. B. Kleijn, and R. C. Hendriks, “An instrumental intelligibility metric based on information theory,” *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, Jan. 2018.

[203] J. D. Griffiths, "Optimum linear filter for speech transmission," *J. Acoust. Soc. Am.*, vol. 43, no. 1, pp. 81–86, 1968.

[204] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 277–282, Aug. 1976.

[205] J. B. Crespo and R. C. Hendriks, "Multizone speech reinforcement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 54–66, Jan. 2014.

[206] B. N. Gover and J. S. Bradley, "Measures for assessing architectural speech security (privacy) of closed offices and meeting rooms," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3480–3490, 2004.

[207] J. S. Bradley, M. Apfel, and B. N. Gover, "Some spatial and temporal effects on the speech privacy of meeting rooms," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3038–3051, 2009.

[208] J. S. Bradley and B. N. Gover, "A new system of speech privacy criteria in terms of Speech Privacy Class (SPC) values," pp. 1–5, 2010.

[209] B. N. Gover and J. S. Bradley, "ASTM metrics for rating speech privacy of closed rooms and open plan spaces," *Canadian Acoust.*, vol. 39, pp. 50–51, 2011.

[210] *Standard test method for objective measurement of speech privacy in open plan spaces using articulation index.* ASTM Int. E1130-08, 2008.

[211] *Standard test method for objective measurement of the speech privacy provided by a closed room.* ASTM Int. E2638-10, 2010.

[212] M. Unoki, Y. Kashihara, M. Kobayashi, and M. Akagi, "Study on method for protecting speech privacy by actively controlling speech transmission index in simulated room," in *Asia-Pacific Signal Inform. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, IEEE, 2017, pp. 1–6.

[213] T. Betlehem and P. D. Teal, “A constrained optimization approach for multi-zone surround sound,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2011, pp. 437–440.

[214] J. Garofolo *et al.*, “TIMIT acoustic-phonetic continuous speech corpus,” *Linguistic Data Consortium*, 1993.

[215] *Mapping function for transforming P. 862 raw result scores to MOS-LQO.* Int. Telecommun. Union (ITU), ITU-T Rec. P.862.1, 2003.

[216] H. Chen, P. Samarasinghe, and T. D. Abhayapala, “In-car noise field analysis and multi-zone noise cancellation quality estimation,” in *Asia-Pacific Signal Inform. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, IEEE, 2015, pp. 773–778.

[217] L. Ward, B. G. Shirley, Y. Tang, and W. J. Davies, “The effect of situation-specific non-speech acoustic cues on the intelligibility of speech in noise,” in *Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2017.

[218] W. Zhang, T. D. Abhayapala, T. Betlehem, and F. M. Fazi, “Analysis and control of multi-zone sound field reproduction using modal-domain approach,” *J. Acoust. Soc. Am.*, vol. 140, no. 3, pp. 2134–2144, Sep. 2016, doi: `10.1121/1.4963084`.

[219] W. Zhang, P. Samarasinghe, H. Chen, and T. Abhayapala, “Surround by Sound: A Review of Spatial Audio Recording and Reproduction,” *Appl. Sciences*, vol. 7, no. 6, p. 532, May 2017, doi: `10.3390/app7050532`.

[220] J. Francombe, R. Mason, M. Dewhirst, and S. Bech, “Determining the threshold of acceptability for an interfering audio programme,” in *Audio Eng. Soc. Conv. 132*, Audio Eng. Soc., 2012.

[221] F. Winter, J. Ahrens, and S. Spors, “On Analytic Methods for 2.5-D Local Sound Field Synthesis Using Circular Distributions of Secondary Sources,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 5, pp. 914–926, May 2016, doi: `10.1109/TASLP.2016.2531902`.

[222] P. N. Samarasinghe, M. A. Poletti, S. A. Salehin, T. D. Abhayapala, and F. M. Fazi, "3D soundfield reproduction using higher order loudspeakers," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2013, pp. 306–310.

[223] *Sound system equipment-Part 16: Objective rating of speech intelligibility by speech transmission index.* IEC 60268-16, 2003.

[224] D. Byrne *et al.*, "An international comparison of long-term average speech spectra," *J. Acoust. Soc. Am.*, vol. 96, no. 4, pp. 2108–2120, Oct. 1994, doi: 10.1121/1.410152.

[225] *Artificial voices.* Int. Telecommun. Union (ITU), ITU-T Rec. P.50, 1999.

[226] T. W. Parks and C. S. Burrus, *Digital Filter Design.* New York, NY, USA: John Wiley & Sons, 1987.

[227] Y.-H. Kim and J.-W. Choi, *Sound Visualization and Manipulation.* Singapore: John Wiley & Sons Singapore Pte. Ltd., Sep. 2013.

[228] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Int. Congr. Acoust.*, Tokyo, Japan, 1968, pp. 17–20.

[229] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 380–391, Oct. 1976.

[230] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Audio Eng. Soc. Conv. 122*, Audio Eng. Soc., 2007.

[231] S. J. Elliott and P. A. Nelson, "The active control of sound," *Electronics & communication engineering journal*, vol. 2, no. 4, pp. 127–136, 1990.

[232] S. M. Kuo, S. Mitra, and W.-S. Gan, "Active noise control system for headphone applications," *IEEE Trans. Control Syst. Technol.*, vol. 14, no. 2, pp. 331–335, 2006.

[233] H. Sano, T. Inoue, A. Takahashi, K. Terai, and Y. Nakamura, "Active control system for low-frequency road noise combined with an audio system," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 7, pp. 755–763, 2001.

[234] J. Cheer and S. J. Elliott, "The design and performance of feedback controllers for the attenuation of road noise in vehicles," *Int. J. Acoust. Vibration*, vol. 19, no. 3, pp. 155–164, 2014.

[235] Y. Xiao and J. Wang, "A new feedforward hybrid active noise control system," *IEEE Signal Process. Lett.*, vol. 18, no. 10, pp. 591–594, 2011.

[236] N. V. George and G. Panda, "On the development of adaptive hybrid active noise control system for effective mitigation of nonlinear noise," *Signal Process.*, vol. 92, no. 2, pp. 509–516, 2012.

[237] O. J. Tobias and R. Seara, "Leaky delayed LMS algorithm: Stochastic analysis for gaussian data and delay modeling error," *IEEE Trans. Signal Process.*, vol. 52, no. 6, pp. 1596–1606, 2004.

[238] I. T. Ardekani and W. H. Abdulla, "Adaptive signal processing algorithms for creating spatial zones of quiet," *Digital Signal Process.*, vol. 27, pp. 129–139, 2014.

[239] S. Elliott, *Signal processing for active control.* Academic press, 2000.

[240] S. Spors and H. Buchner, "Efficient massive multichannel active noise control using wave-domain adaptive filtering," in *Int. Symp. Commun., Control Signal Process. (ISCCSP)*, IEEE, 2008, pp. 1480–1485.

[241] J. Zhang, W. Zhang, and T. D. Abhayapala, "Noise cancellation over spatial regions using adaptive wave domain processing," in *Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, IEEE, 2015, pp. 1–5.

[242] J. Zhangg, T. D. Abhayapala, P. N. Samarasinghe, W. Zhang, and S. Jiang, "Sparse complex FxLMS for active noise cancellation over spatial regions," in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2016, pp. 524–528.

[243] K. Kondo and K. Nakagawa, “Speech emission control using active cancellation,” *Speech Commun.*, vol. 49, no. 9, pp. 687–696, Sep. 2007.

[244] L. Athanas, “Open air noise cancellation,” U.S. Patent 2011/0274283 A1, Nov. 2011.

[245] C. R. Hart and S.-K. Lau, “Active noise control with linear control source and sensor arrays for a noise barrier,” *Journal of Sound and Vibration*, vol. 331, no. 1, pp. 15–26, 2012.

[246] W. Chen, H. Min, and X. Qiu, “Noise reduction mechanisms of active noise barriers,” *Noise Control Eng. J.*, vol. 61, no. 2, pp. 120–126, 2013.

[247] C. Hofmann, M. Guenther, M. Buerger, and W. Kellermann, “Higher-order listening room compensation with additive compensation signals,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2016, pp. 534–538.

[248] M. A. Poletti, “Active acoustic systems for the control of room acoustics,” *Noise & Vibration Worldwide*, vol. 44, no. 4, pp. 10–26, 2013.

[249] J. O. Jungmann, R. Mazur, and A. Mertins, “Joint time-domain reshaping and frequency-domain equalization of room impulse responses,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2014, pp. 6642–6646.

[250] L. Krishnan, P. D. Teal, and T. Betlehem, “A robust sparse approach to acoustic impulse response shaping,” in *Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2015, pp. 738–742.

[251] K. Kinoshita *et al.*, “A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research,” *EURASIP J. Advances Signal Process.*, vol. 2016, no. 1, Dec. 2016.

[252] S. Spors, H. Buchner, R. Rabenstein, and W. Herbordt, “Active listening room compensation for massive multichannel sound reproduction systems using wave-domain adaptive filtering,” *J. Acoust. Soc. Am.*, vol. 122, no. 1, pp. 354–369, 2007.

[253] S. Spors and H. Buchner, "An approach to massive multichannel broadband feedforward active noise control using wave-domain adaptive filtering," in *Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, IEEE, 2007, pp. 171–174.

[254] M. A. Poletti, T. Betlehem, and T. D. Abhayapala, "Higher-order loudspeakers and active compensation for improved 2d sound field reproduction in rooms," *J. Audio Eng. Soc.*, vol. 63, no. 1, pp. 31–45, 2015.

[255] M. A. Poletti, T. Betlehem, and T. Abhayapala, "Higher order loudspeakers for improved surround sound reproduction in rooms," in *Audio Eng. Soc. Conv. 133*, Audio Eng. Soc., 2012.

[256] T. Betlehem and M. A. Poletti, "Two dimensional sound field reproduction using higher order sources to exploit room reflections," *J. Acoust. Soc. Am.*, vol. 135, no. 4, pp. 1820–1833, 2014.

[257] G. N. Lilis, D. Angelosante, and G. B. Giannakis, "Sound field reproduction using the lasso," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 1902–1912, Nov. 2010.

[258] T. Betlehem and C. Withers, "Sound field reproduction with energy constraint on loudspeaker weights," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 8, pp. 2388–2392, Oct. 2012.

[259] M. A. Poletti and T. D. Abhayapala, "Interior and exterior sound field control using general two-dimensional first-order sources," *J. Acoust. Soc. Am.*, vol. 129, no. 1, pp. 234–244, 2011.

[260] D. Zhou and V. DeBrunner, "A new active noise control algorithm that requires no secondary path identification based on the SPR property," *IEEE Trans. Signal Process.*, vol. 55, no. 5, pp. 1719–1729, May 2007.

[261] M. Wu, G. Chen, and X. Qiu, "An improved active noise control algorithm without secondary path identification based on the frequency-domain subband

architecture," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1409–1419, Nov. 2008.

[262] M. Gao, J. Lu, and X. Qiu, "A simplified subband ANC algorithm without secondary path modeling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 7, pp. 1164–1174, Jul. 2016.

[263] P. Zech, V. Lato, and S. Rinderknecht, "Direct adaptive feedforward compensation of narrowband disturbances without explicit identification of the secondary path model," *J. of Sound and Vibration*, vol. 401, pp. 282–296, Aug. 2017.

[264] J. Tao, S. Wang, X. Qiu, and J. Pan, "Performance of a multichannel active sound radiation control system near a reflecting surface," *Appl. Acoust.*, vol. 123, pp. 1–8, Aug. 2017.

[265] S. Koyama, K. Furuya, Y. Hiwasaki, and Y. Haneda, "Analytical approach to wave field reconstruction filtering in spatio-temporal frequency domain," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 685–696, Apr. 2013.

[266] S. Koyama, K. Furuya, H. Uematsu, Y. Hiwasaki, and Y. Haneda, "Real-time sound field transmission system by using wave field reconstruction filter and its evaluation," *IEICE Trans. on Fundamentals of Electron., Commun. and Comput. Sci.*, vol. 97, no. 9, pp. 1840–1848, 2014.

[267] K. K. Paliwal and W. B. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis*, Elsevier Science Inc., 1995, ch. 12, pp. 433–466.

[268] P. Stoica and R. L. Moses, *Spectral analysis of signals.* Upper Saddle River, NJ: Pearson Prentice Hall, 2005.

[269] K. Tanaka, C. Shi, and Y. Kajikawa, "Binaural active noise control using parametric array loudspeakers," *Applied Acoustics*, vol. 116, pp. 170–176, Jan. 2017, ISSN: 0003-682X.

[270] E. A. Habets, “Room impulse response generator,” *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2, p. 1, 2006.

[271] C.-N. Wang and J.-H. Torng, “Experimental study of the absorption characteristics of some porous fibrous materials,” *Appl. Acoust.*, vol. 62, no. 4, pp. 447–459, 2001.

[272] C. Shi, Y. Kajikawa, and W.-S. Gan, “An overview of directivity control methods of the parametric array loudspeaker,” *APSIPA Trans. Signal Inform. Process.*, pp. 1–30, Dec. 2014.

[273] P. J. Westervelt, “Parametric acoustic array,” *J. Acoust. Soc. Am.*, vol. 35, no. 4, pp. 535–537, 1963.

[274] H. O. Berktay and D. J. Leahy, “Farfield performance of parametric transmitters,” *J. Acoust. Soc. Am.*, vol. 55, no. 3, pp. 539–546, 1974.

[275] Y. Sugibayashi, S. Kurimoto, D. Ikefuji, M. Morise, and T. Nishiura, “Three-dimensional acoustic sound field reproduction based on hybrid combination of multiple parametric loudspeakers and electrodynamic subwoofer,” *Appl. Acoust.*, vol. 73, no. 12, pp. 1282–1288, Dec. 2012.

[276] C. Shi, E.-L. Tan, and W.-S. Gan, “Hybrid immersive three-dimensional sound reproduction system with steerable parametric loudspeakers,” in *Proc. Meetings Acoust.*, vol. 19, Acoust. Soc. Am., 2013, p. 055 003.

[277] C. Shi and Y. Kajikawa, “A comparative study of preprocessing methods in the parametric loudspeaker,” in *Asia-Pacific Signal Inform. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, IEEE, 2014, pp. 1–5.

[278] C. Shi and W.-S. Gan, “Product directivity models for parametric loudspeakers,” *J. Acoust. Soc. Am.*, vol. 131, no. 3, pp. 1938–1945, 2012.

[279] C. Shi and Y. Kajikawa, “A convolution model for computing the far-field directivity of a parametric loudspeaker array,” *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. 777–784, 2015.

[280] Chuang Shi and Woon-Seng Gan, "Grating lobe elimination in steerable parametric loudspeaker," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 58, no. 2, pp. 437–450, 2011.

[281] F. Farias and W. Abdulla, "On rayleigh distance and absorption length of parametric loudspeakers," in *Asia-Pacific Signal Inform. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, IEEE, 2015, pp. 1262–1265.

[282] S. Butterworth, "On the theory of filter amplifiers," *Wireless Engineer*, vol. 7, no. 6, pp. 536–541, 1930.

[283] M. E. Van Valkenburg, *Analog filter design.* Holt, Rinehart, and Winston, 1982.