UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Sciences

Institute of Sound and Vibration Research

Practical Audio System Design for Private Speech Reproduction

by

Daniel Wallace

ORCID ID: 0000-0003-0212-5395

Thesis for the degree of Doctor of Philosophy

September 2020

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES
INSTITUTE OF SOUND AND VIBRATION RESEARCH

Thesis for the degree of Doctor of Philosophy

**PRACTICAL AUDIO SYSTEM DESIGN
FOR PRIVATE SPEECH REPRODUCTION**

by Daniel Wallace

Multi-zone sound field control allows individuals to listen to personalised audio content whilst sharing a physical space. Applications of this technology include home entertainment, audio reproduction in public spaces such as museums, shops or exhibitions, and providing areas where the privacy of sensitive communication can be safeguarded without the need for physical barriers. The problem of transmitting a speech signal to a single listener and reducing the intelligibility of that signal elsewhere is the focus of the present thesis. The motivation behind the presented experiments and simulations is to identify the practical trade-offs that must be considered in the design of these *Speech Privacy Control* systems.

Conventional personal audio systems use loudspeaker array processing to produce a bright zone for the intended user of the system and a dark zone where silence is desired. However, established performance metrics and system optimisation techniques do not necessarily yield privacy for the target listener, as attenuated speech may remain intelligible within the dark zone. A system is proposed that focusses a synthetic masking signal into the dark zone to selectively reduce the intelligibility of the leaked speech. Privacy is ensured by adjusting the masker to meet pre-defined constraints on the speech intelligibility in each zone. This design methodology utilises information from speech intelligibility tests and subjective preference evaluations in order to improve the utility and acceptability of such systems for all nearby listeners.

In addition to the design of the masking signal, the performance of a speech privacy control system is affected by the loudspeaker array design and the location of the listening zones. These effects are explored using experimental measurements of a loudspeaker array in a room, and the results are used to select two system configurations for additional evaluation using listening tests. The perceived performance of a system is also affected by the surrounding acoustic environment, notably due to reverberation and background noise, which may change over time. The effects of room reverberation are investigated using image source simulations and acoustical measurements within a room, and the performance is evaluated in terms of the achievable level of acoustic contrast, the difference in speech intelligibility between zones, and the masking signal levels that are required to achieve privacy. A proposal is made to further enhance privacy by combining the effects of background noise and artificial masking signals. This method reduces the level of acoustic contrast that is required to achieve a given level of privacy, compared to the case where the masking is provided by the background noise alone.

Contents

List of Figures

Accompanying Materials

MATLAB implementations of the psychoacoustic metrics used in this thesis are available at doi:10.5258/SOTON/D1486.

Declaration of Authorship

I, Daniel Wallace, declare that this thesis entitled Practical Audio System Design for Private Speech Reproduction and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where this thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
    - "Optimisation of Personal Audio Systems for Intelligibility Contrast," in *Proceedings of the 144^{th} Audio Engineering Society Convention*, (Milan, Italy), May 2018.
    - "Combining Artificial and Natural Background Noise in Personal Audio Systems," in *Proceedings of the 10^{th} IEEE Sensor Array and Multichannel Signal Processing Workshop*, (Sheffield, UK), July 2018.
    - "The Design of Personal Audio Systems for Speech Transmission using Analytical and Measured Responses," in *Proceedings of the 44^{th} International Conference on Acoustics, Speech and Signal Processing*, (Brighton, UK), May 2019.
    - "Design and Evaluation of Personal Audio Systems Based on Speech Privacy Constraints," *Journal of the Acoustical Society of America* 147(4):2271-2282, 2020.
    - "A Low-Cost Loudspeaker Array for Personal Audio with Enhanced Vertical Directivity", in *Proceedings of the 49^{th} International Congress and Exposition on Noise Control Engineering*, August 2020.

Signed:

Date:

Acknowledgements

I am extremely thankful to Jordan Cheer for his many valuable inputs into the work presented in this thesis. He has been a dependable advisor, patient co-author, and has willingly shared his experience by providing guidance into both technical and subjective experiments. Thank you for keeping me motivated, despite periodic pessimism, programming problems and a pandemic.

Throughout the production of this thesis, I have been grateful for many opportunities to share my research with the public. I am grateful to Silvia Lanati and Steve Dorney from the University of Southampton Public Engagement with Research Unit, and Peter Horak from NGCM for resourcing and supporting this vital work.

I am grateful to Charlie House for many helpful conversations throughout this project, and for the reliability and flexibility of the loudspeaker array prototype designed by him and his team.

The work presented in this thesis was funded and supported by the EPSRC Centre for Doctoral Training in Next Generation Computational Modelling (Grant EP/L015382/1). I am particularly grateful to Jacqui Bonnin at NGCM for thoroughly enhancing my postgraduate experience through her organisation of outreach, training and social events, all whilst maintaining *gentle* pressure to keep on writing.

Finally, I offer my enduring thanks to my wonderful wife, Hannah. Thank you for being the first to read this thesis all the way through, and for sustaining me in every way throughout its writing.

Abbreviations

ACC	Acoustic Contrast Control
AI	Articulation Index
AITF	Audibility-Intelligibility Transfer Function
ANSI	American National Standards Institute
ASTM	American Society for Testing and Materials
CSTR VCTK	Centre for Speech Technology Research Voice Cloning Toolkit
ESTOI	Extended Short-Time Objective Intelligibility
FIR	Finite Impulse Response
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organisation for Standardisation
ISVR	Institute of Sound and Vibration Research
ITU	International Telecommunication Union
JND	Just-Noticeable Difference
KEMAR	Knowles Electronics Mannequin for Acoustic Research
MADI	Multichannel Audio Digital Interface
PA	Psychoacoustic Annoyance
PESQ	Perceptual Evaluation of Speech Quality
PI	Privacy Index
PM	Pressure Matching
POLQA	Perceptual Objective Listening Quality Analysis
POSZ	Perceptually Optimised Sound Zones
PSD	Power Spectral Density
SATF	Signal-Audibility Transfer Function
SII	Speech Intelligibility Index
SNR	Signal-to-Noise Ratio
SPC	Speech Privacy Class
SPL	Sound Pressure Level
STI	Speech Transmission Index
STOI	Short-Time Objective Intelligibility
VAST	Variable Span Trade-off

Chapter 1

Introduction

Conventional audio systems are often praised for their "room-filling sound", but this is not always desirable when multiple people share a listening space. When sound is indiscriminately radiated into spaces characterised by reverberation and acoustic scattering, the sound field tends towards spatial uniformity [1], raising the potential for annoyance and distraction [2, 3]. The personal audio technology [4] discussed in this thesis offers a solution to these problems by producing spatially separated listening zones.

When personal audio systems are used for entertainment purposes, e.g. [5–7], poor separation between the listening zones can be regarded as an inconvenience, but certain applications of personal audio technology in public spaces could have more significant consequences regarding listeners' privacy [8, 9]. For example, a personal audio system could be used to transmit sensitive conversations between staff and customers through security partitions in places such as banks, surgeries or pharmacy counters. Similarly, in-vehicle telecommunication could also be enhanced through the use of appropriately designed personal audio systems; ensuring that private phone conversations remain private. The conceptual design of such a system is the subject of a recent Jaguar-Land Rover patent [10].

These *Speech Privacy Control* systems necessarily have more complex design constraints than conventional multi-zone loudspeaker systems. As well as the standard considerations of reverberation [11, 12], background noise [13], filter design [5, 14–17], loudspeaker selection [6, 18] and zonal geometry [19, 20], speech privacy control requires an understanding of human speech perception to be integrated into the design. A successful system must ensure that speech intended for the target listener is clear and intelligible, whilst also guaranteeing that listeners outside of the target region cannot understand these messages [8].

The goal of this thesis is to identify the practical trade-offs that must be considered in the design of private personal audio systems. A variety of methods are required in order to fully achieve this goal, as the speech privacy control problem contains both physical and perceptual challenges. Experimental measurements and numerical simulations of loudspeaker arrays reveal how changes to system geometry, room reverberation and background noise each affect the technical performance of speech privacy control systems. This technical assessment is augmented by

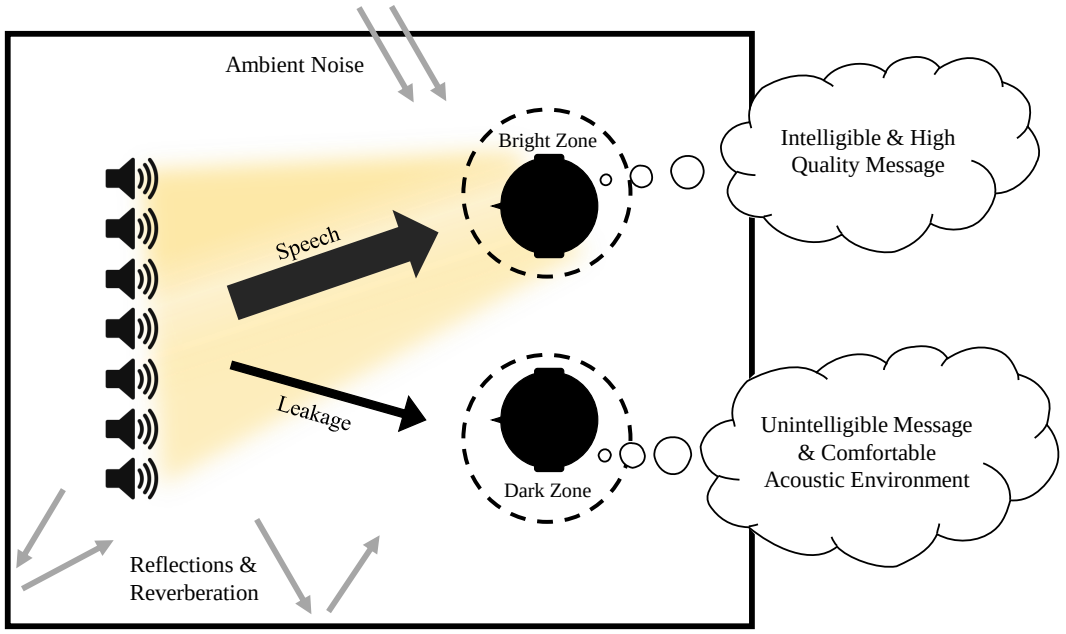


FIGURE 1.1: Schematic of a conventional personal audio system built using an array of loudspeakers. The signals output by each loudspeaker are designed to produce a region of constructive interference termed the *bright zone*, where the intended listener of the system is located, and a region of destructive interference, the *dark zone*, in which silence is desired.

objective and subjective listening tests, which respectively provide information on the intelligibility of speech and the perceived acoustical environment in each listening zone. Together, these methods are used to inform the design requirements of private personal audio systems and to generate a design methodology which can be followed in a variety of applications.

By way of introduction, Figure 1.1 shows a diagram of a conventional personal audio system, which uses an array of appropriately driven loudspeakers to produce a pair of sound zones. These zones are described as acoustically *bright* and *dark* [14], borrowing terminology from visual descriptions for areas of light and shadow. A target signal is focussed into the bright zone of a system, but commensurate with the visual case, absolute silence (or *darkness*) in the dark zone of the system is not practically possible, due to room reverberation, ambient noise and practical limitations in array directivity. A common and intuitive way to quantify the performance of a personal audio system is the *Acoustic Contrast* [14], i.e. the ratio of mean-square pressure between the zones. However, this measure does not capture the essence of what listeners expect a speech privacy control system to deliver: a significant difference in the intelligibility of a speech signal between the zones. This ratio has been described as the speech intelligibility contrast [8].

Speech privacy control systems need not rely solely on a single sound zoning process in order to satisfy their designed function. Rather, a greater difference in the speech intelligibility between zones can be achieved by focussing a secondary *masking signal* into the dark zone, using a similar beamforming technique to that used with the speech programme [8]. The aim of this signal is to mask any speech which is leaked from the bright zone into the dark zone, thereby affording privacy to the target listener. However, for the same reasons as described above, this masking signal cannot be strictly confined to the dark zone; leakage of the masker into the bright zone may

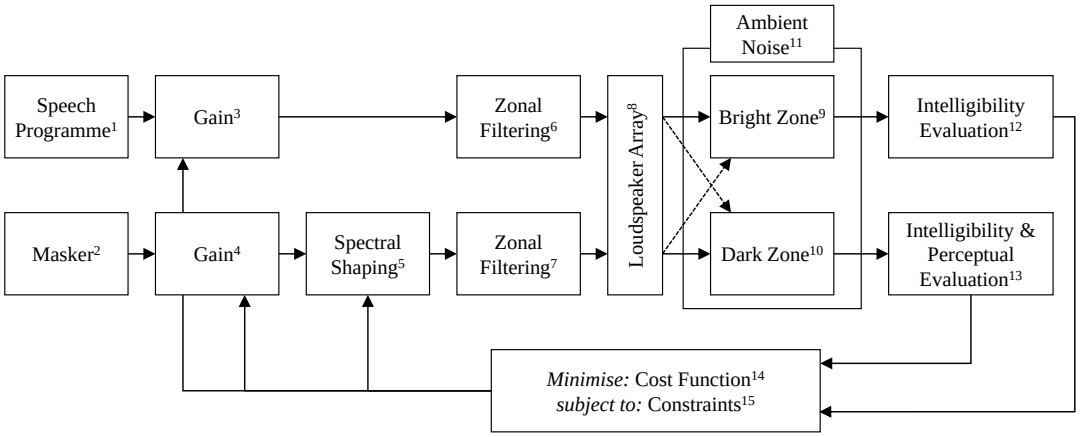


FIGURE 1.2: Block diagram of the proposed personal audio system. Superscripted numbers are referred to in the text.

undesirably affect the quality or intelligibility of the speech for the target listener, necessitating careful design.

Throughout this thesis, various configurations of speech privacy control systems are presented. Figure 1.2 shows a general block diagram of the investigated systems. The individual blocks are described in the following paragraphs and are referenced using superscripted numbers. As described above, the proposed system is driven by two signals. The programme material[1] to be delivered to the target listener is exclusively composed of speech signals. For the purposes of the experiments presented in this thesis, speech signals are reproduced from pre-recorded sentence corpora, e.g. [21–23]. This allows comparisons to be made between the original speech signal and the combined speech and noise signals that are received in each listening zone. Real-time implementations, such as those used in voice-transfer or telecommunications systems would require real-time processing of an incoming audio stream. The masker[2] may take the form of pre-recorded or synthesized samples that are appropriate for the masking of speech. The studies presented in this thesis are restricted to considerations of energetic masking, and no additional processing, such as dynamic range compression, is applied to the input speech. Accordingly, the considered systems may adjust the relative level of each input signal[3−4] and the frequency content of the masking signal may also be adjusted[5] to compensate for the frequency response of the system or to account for listener preferences. Each signal is then independently passed through zonal filterbanks[6−7] which produce signals that are output by a loudspeaker array[8] in order to focus the speech programme into the bright zone[9] and the masker into the dark zone[10]. In addition, the signals in each zone may include a contribution from the ambient noise in the reproduction environment[11].

In order to meet the intelligibility and privacy requirements described above, the signals received in each zone must be modified using the input signal controls[3−5]. It is evidently impractical to make these adjustments based on feedback from listeners during operation. Rather, proxies for intelligibility and privacy must be found that accurately predict the listening conditions in each zone - objective intelligibility metrics[12−13] are used for this purpose. Furthermore, it is not sufficient for the masking signal to be adjusted solely based on predictions of privacy for the listener in the bright zone, as high masker levels in the vicinity of the system may be regarded

as noise pollution by other listeners in the shared space. This trade-off between the suitability of the masker and its desired functionality may be resolved through the use of an optimisation procedure, represented by the feedback loop in the block diagram in Figure 1.2. The cost function[14] in the optimisation encapsulates the negative perceptual effects that the masker may have on nearby listeners. These effects are minimised whilst simultaneously providing privacy by constraining the system[15] to provide a sufficiently high level of intelligibility in the bright zone and a sufficiently low level of intelligibility in the dark zone. In this thesis, a number of perceptual metrics are applied to the signals received in the dark zone, and these are compared against listening test results to indicate which combination of metrics is most appropriate for modelling listener preferences.

1.1 Thesis Structure

Chapter 2 provides a review of literature pertaining to personal audio system design, firstly discussing a range of techniques that enable the control of zonal sound fields. Following this, several recent examples are discussed in which systems have been assessed subjectively, or the human perception of sound has been directly incorporated into the system design process. Finally, the specific constraints and challenges of speech privacy control are discussed, alongside a review of a small number of publications where this particular problem has been directly addressed.

This is followed in Chapter 3 with a review of methods for instrumentally assessing speech intelligibility and privacy. Links are drawn between physical, measurable properties of acoustic signals and the intelligibility of speech reproduced in these conditions. Similar conversions are made between speech intelligibility and the impression of privacy, following guidance from published reports and standards concerning the acoustic design of open plan offices.

Chapter 4 introduces a range of metrics that are used throughout this thesis to quantify the potential negative impacts of introducing additional masking noise to the environment. A selection of these metrics are later used to evaluate and adapt the signals emitted by a speech privacy control system in order to maximise its acceptability. Together with Chapters 2 and 3, this chapter forms the background for the work presented in this thesis.

Chapter 5 describes the loudspeaker array prototype that is used as a platform for the experimental results presented in this work. Transfer response measurements of this array, within a listening room, are described and a mathematical derivation of the Acoustic Contrast Control process is presented. The chapter closes with an investigation into how varying the location of the zones and the loudspeaker array geometry affects the frequency dependent acoustic contrast that is achievable by an array, a quantity that has a direct bearing on the level of privacy that can be achieved.

Results from a range of listening tests that have been carried out using the array prototype are presented in Chapter 6. Firstly, speech intelligibility is assessed in the dark zone of the system under a range of masking conditions and for two loudspeaker array geometries, using a matrix sentence test. Listeners' subjective preference for each of these conditions is then assessed using a paired preference format.

The results from these listening tests are used in Chapter 7 to design a method for specifying the required Signal-to-Noise Ratios (SNRs) in each listening zone of a speech privacy control system. Privacy is ensured by setting maximum and minimum speech intelligibility constraints in the bright and dark zones respectively, based on the objective speech intelligibility metrics described in Chapter 3. Subject to these constraints, and recognising that a practical system must consider the experience of all nearby listeners, the acceptability of the masker is maximised by evaluating the dark zone sound field using the subjective metrics described in Chapter 4.

Chapter 8 provides a discussion of personal audio system performance in reverberant spaces, using the method described in Chapter 7, to set speech intelligibility targets in each listening zone. Firstly, the chapter describes how analytical and measured transfer responses can be used for the optimisation of sound zoning filters, and describes how reverberation in simple rooms can be modelled using image sources. The image source model is used to quantify the degradation of acoustic contrast and speech intelligibility contrast due to reverberation, with relation to the critical distance from the source. These simulations are then compared against experimental measurements in a lightly reverberant room.

Continuing the treatment of practical issues in personal audio system design, Chapter 9 discusses the positive and negative consequences of installing a speech privacy control system in a noisy environment. The chapter introduces the concept of harnessing the masking effect of the ambient noise to assist with privacy control, and describes a trade-off between acoustic contrast requirements and the signal levels that are necessary to ensure that speech intelligibility constraints can be met. The potential problems of temporal unpredictability and spatial irregularity of the overall ambient noise in a reproduction environment are addressed by isolating the background noise component, which is relatively steady and spatially diffuse by comparison.

Chapter 10 presents a summary of the main conclusions from the thesis and provides suggestions for future work.

1.2 Contributions

The main original contributions of this thesis are:

1. An investigation into masking signal design for speech privacy control systems, which has yielded a method for choosing the level and spectral shape of a masking signal based on objective and perceptual metrics. Specifically:

   (a) Objective intelligibility metrics such as the Speech Intelligibility Index (SII) can be used to ensure that speech intended for the target listener is sufficiently intelligible, and that any speech material leaked into privacy-sensitive areas is rendered unintelligible by the masking signal.

   (b) Psychoacoustic metrics can be used to recognise and reduce the potential negative effects of introducing masking noise into the environment for listeners in the dark zone of the personal audio system.

   Analysis of these metrics has highlighted a conflict between two objectives in the masking signal design process; the masker must provide privacy whilst also minimising the potential

for noise annoyance. This trade-off may be resolved in individual system designs by setting context-appropriate constraints on the bright and dark zone intelligibility level and minimising the loudness of the masker subject to these constraints. This masking signal design method was first presented at the 144$^{\text{th}}$ Audio Engineering Society Convention [24]. A refinement and experimental validation of this approach has been published in the Journal of the Acoustical Society of America [25].

2. The effects of room reverberation on speech privacy control have been investigated using an image source simulation of a cuboid room with four different levels of surface absorption. System performance is characterised in terms of the achievable acoustic contrast, the speech intelligibility in each zone and the masking signal levels that are required to ensure privacy. A dependency on the location of the zones with respect to the loudspeaker array has been identified and linked to the critical distance in the reproduction environment.

3. The advantages and disadvantages of using measured and analytical transfer responses in the optimisation of acoustic contrast control filters have been investigated. The results indicate that when measured transfer responses are used, lower masking signal levels are required compared to when analytical transfer responses are used. However, this advantage is traded off against the additional flexibility and implementational simplicity of using analytical responses. An article presenting a subset of these results has been published in the proceedings of the 44$^{\text{th}}$ IEEE International Conference on Acoustics, Speech and Signal Processing [26].

4. The effect of steady, speech-shaped ambient noise on the design specification of speech privacy control systems has been investigated. This was accomplished by simulating a noisy environment, superimposing the zonal sound fields produced by a range of simulated speech privacy control systems, then predicting the intelligibility in each zone using the SII metric. For each simulated system, the speech and masker levels were adjusted using the method described in Contribution 1 to optimally achieve a given level of speech intelligibility in each zone. The results demonstrate that if the effects of artificial masking and ambient noise are combined, this reduces the level of acoustic contrast that is required to achieve a certain degree of privacy performance, compared to the case where privacy is obtained by relying on the ambient noise alone. For the tested array configurations, the reduction in required contrast was between 6 and 8 dB, and the more onerous the speech intelligibility constraints, the greater the benefit of combining artificial masking and ambient noise. Results from a simplified simulation which assumes frequency independent acoustic contrast and steady noise conditions were presented at the 10$^{\text{th}}$ IEEE Sensor Array and Multichannel Signal Processing Workshop [27].

5. Recorded, real-world ambient noise samples have also been considered as maskers in order to generalise the results discussed in Contribution 4, and similar trends are observed, except that the masking effect of these signals can be more inconsistent and challenging to predict instrumentally, limiting their effectiveness. However, the background noise in a space is more temporally and spatially uniform, meaning that this component of the ambient noise can be relied upon in predictions of masking within private listening zones. System designers can either opt for a static masker based on the maximum and minimum expected background noise levels, or use an adaptive algorithm based on live background noise measurements. The former approach is simpler, but requires greater reliance on

the artificial masker, whereas the more complex adaptive system can reduce the artificial masker when it is not needed, yielding a more acceptable acoustic environment in the dark zone.

Chapter 2

Personal Audio

The work presented in this thesis centres around the considerations that must be made to develop a practical audio system for the reproduction of private speech. This objective is a specific case among the many and varied applications of personal audio, a selection of which are shown in Figure 2.1. This chapter will provide a review of the techniques, applications and challenges of personal audio, specifically focussing on the areas most strongly connected to the task of private speech reproduction. As illustrated in Figure 2.1, personal audio is itself a subset of the broader discipline of Sound Field Control, which encapsulates a great breadth of applications and techniques involving the generation and manipulation of acoustic fields [28]. These include active noise control [29], the absorption or generation of acoustic reflections [30, 31], and techniques used for reproducing immersive, spatial sound fields [32, 33]. Each of these areas has its own extensive body of literature, and unique challenges, so these will not be covered in this review. However, many of the considerations that must be taken into account when designing a personal, zonal audio system are common to the general problem of sound field control.

A cursory analysis of the speech privacy control problem described in Chapter 1 suggests that successful systems must employ a method for controlling the spread of sound between regions where different people are located. Accordingly, the present chapter begins with a review of methods that can be used to generate individual listening zones. In recent years, some of these approaches have been refined by incorporating aspects of human sound perception into their design. This focal shift has yielded a transition away from the production of technical demonstrations of what is possible to achieve, towards practical embodiments of the technology and evaluation using subjective testing. After reviewing a selection of these projects, the chapter concludes by discussing the specific task of providing private listening zones and detailing the results of a small number of recent studies in which this problem has been approached.

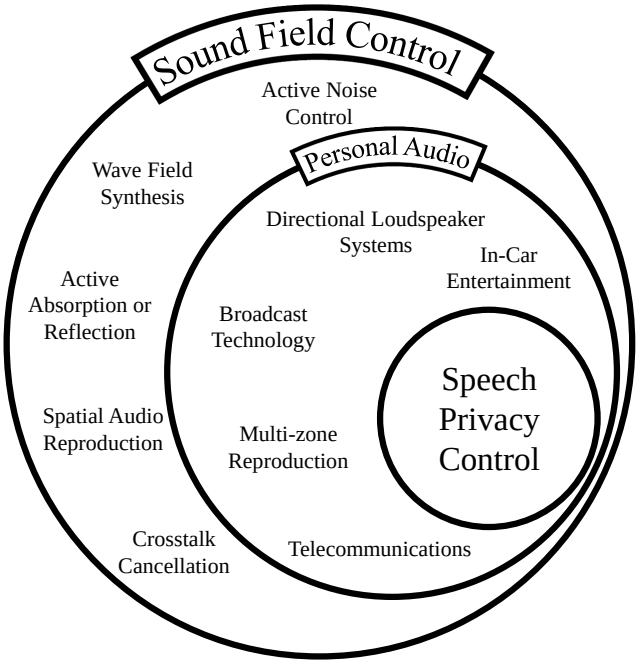


FIGURE 2.1: Examples of sound field control and personal audio applications and techniques.

2.1 Controlling Zonal Sound Fields

Personal audio depends on the ability to generate spatially distinct regions with different sound levels, and this can be achieved using several different techniques. Druyvesteyn and Garas [4] proposed that the problem of personal audio reproduction could be separated into three frequency regions, each with a preferred strategy for providing the greatest level of isolation between spaces: active sound control for low frequencies, loudspeaker array processing for mid-frequencies, and individual loudspeaker directivity and reverberation control at high frequencies. Whichever combination of methods is used, individual sound zones can be layered using linear superposition to generate a multi-zone sound field, in which the occupants of each zone ideally hear their own desired programme material with minimal interference from the programme reproduced in the other zones. This multi-zone, multi-listener scenario has been investigated using large arrays of loudspeakers distributed around the boundary of the reproduction environment, or surrounding the listeners [34], and advantageously, this arrangement also allows the perceived directionality of the reproduced sound field to be controlled [35]. However, the case considered in this thesis is geometrically simpler, with a single target listener and a nearby region where privacy must be prioritised, and correspondingly, this invites the use of more compact, directional sound reproduction systems. Additionally, as the problem of privacy control is restricted to the speech frequency range, the scope of this thesis will likewise focus on solutions that solely involve loudspeaker array processing. To verify this choice of approach, a range of alternative methods and techniques for controlling zonal sound fields are reviewed in the following subsections.

2.1.1 Passive Acoustic Treatment

A simple method for maintaining acoustical separation between listening zones in a shared space is to provide passive sound absorption throughout. Speech presented to a target listener within a particular listening zone is more likely to be audible or intelligible elsewhere if nearby surfaces are acoustically reflective, compared to the case where passive acoustic treatment absorbs the incident sound. Reduction of the reverberation time is not necessarily the sole aim of passive acoustic treatment of this type, rather, passive absorption contributes to a reduction in the spatial decay of sound pressure level, which in turn can improve the privacy between neighbouring spaces [36].

This technique is distinct from the others described in this section, as it does not relate to control of the source directivity. Rather, passive treatment may be installed in an attempt to reduce the negative consequences of sources that have an unavoidably wide directivity, including human talkers. Consequently, passive acoustic treatment can be used in combination with other techniques in order to maximise their performance. In Chapter 8, the effect of varying the absorption coefficient of the boundaries of a closed room is discussed in terms of the acoustic contrast and speech intelligibility contrast between listening zones.

2.1.2 Directional Loudspeaker Systems

The directivity of a single, conventional loudspeaker is chiefly governed by its size compared to the acoustic wavelength [37]. A typical sealed box loudspeaker will radiate approximately omni-directionally at low frequencies, and increasingly directionally at high frequencies. Specifically, at low frequencies, the moving components responsible for the fluctuating acoustic pressure are significantly smaller than the acoustic wavelength being produced, thereby appearing as a single point source of pressure. In practice, however, the finite-sized driver generates a pressure distribution over its surface, and when the acoustic wavelength is smaller than the size of the driver, constructive and destructive interference from different areas of the driver causes the far-field directivity of the source to narrow towards the axis [38]. This high frequency directivity can form a useful component in the production of zonal sound fields, but individual loudspeakers would need to be complemented with other, lower frequency methods, as suggested by Druyvesteyn and Garas [4], to privately reproduce speech. Individual sealed loudspeakers may be used over a wider bandwidth for speech privacy control if they are placed very close to the ears, but this approach relies on their proximity to the user, rather than their directivity [39].

Greater control of low frequency directivity can be achieved through the use of phase-shift or gradient loudspeakers [40]. Rather than relying on phase cancellation due to the radiation from a single loudspeaker diaphragm, gradient loudspeakers can be formed using pairs of drivers with a fixed delay relationship [40]. Alternatively, both sides of a loudspeaker diaphragm can be left open to the air, such that the interaction between the front and rear radiation forms a compound source with dipole directivity. In a generalisation of this approach, the rear radiation from the driver can be modified by selecting the acoustical resistance of a rear port and the compliance of the surrounding cabinet [41]. This controls the phase relationship between the front and rear radiation, which leads to control over the loudspeaker directivity at wavelengths significantly longer than the dimensions of the device.

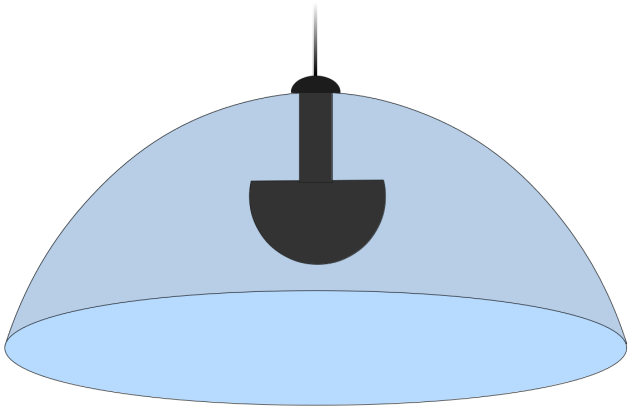


FIGURE 2.2: Diagram of a parabolic reflector loudspeaker. The source radiates upwards into the dome, which reflects approximate plane waves to listeners below.

A less sophisticated solution to confining sound to a narrow region of space is provided by dome-reflector loudspeakers. Usually suspended above a listening zone, a single loudspeaker is situated close to the focal point of a curved reflector. This increases the effective aperture of the loudspeaker, increasing directivity at lower frequencies. The concept has been used extensively in museums to provide accompanying speech to exhibits [42], though a unit designed for use in Public Address systems is manufactured by *Meyer Sound* that is claimed to be capable of delivering 110 dB SPL at 100 metres, with a nominally flat frequency response from 500 Hz to 15 kHz [43, 44].

Whilst all of the alternative approaches mentioned above can be used to generate localised listening zones, they are also comparatively limited, compared to the flexibility offered by loudspeaker arrays. The directional responses of the sources described in this section are fixed to their physical axes, which restricts the range of positions where such systems can be installed in order to insonify a particular area. Loudspeaker arrays need not be configured to radiate on-axis, allowing for greater flexibility in placement and the ability to account for changes in the target listener position in real-time [45]. Additionally, due to the linear acoustic paradigm, signals from a single loudspeaker array may be superimposed to generate two or more distinct listening zones [46]. This flexibility allows the system to focus speech towards one listener, and a masking signal towards another, as described in Chapter 1, thereby increasing the potential for privacy compared to the case with a single sound zoning process.

2.1.3 Parametric Acoustic Arrays

Despite the potential to use the front and rear radiation from standard electrodynamic drivers to improve their directivity, there remains a strong dependence on the physical size of a source compared to the acoustic wavelength. However, parametric acoustic arrays [47] use the non-linear interaction of two ultrasonic waves to generate an audible signal at the difference frequency between the waves, thereby significantly increasing the directivity. Development of this technology for airborne use was almost abandoned in the 1980s due to high levels of harmonic distortion, but improvements in signal processing [48] and ultrasound transducer design [49] resulted in

the commercial development of parametric loudspeakers with similar distortion levels to conventional loudspeakers, albeit at fairly low sound pressure levels. Analysis of the commercially produced range of *Audio Spotlight* loudspeakers shows that the frequency response decays at 40 dB/decade (12 dB/octave) below 1 kHz [50], corroborating the claim from Directional Audio, a company that offers the Audio Spotlight for hire, that the units are "best used when ambient [sound] levels are low and the audio content is mainly speech" [51]. This restriction significantly reduces the applicability of parametric arrays for speech privacy control in noisy public spaces. Furthermore, the high levels of ultrasonic signals required to generate audible sound have frequently been questioned on health and safety grounds [52], suggesting that this technology has much to prove before being widely implemented. Although arrays of conventional loudspeaker drivers cannot match the directivity performance of parametric arrays, they are less restricted in terms of their output level, frequency response and distortion characteristics.

2.1.4 Loudspeaker Array Processing

The concept of generating sound zones using arrays of loudspeakers is a natural extension of active noise control, whereby a localised region of low acoustic potential energy is obtained by manipulating the output of one or more loudspeakers. In active noise control, the secondary source(s) tasked with reducing the noise from a primary disturbance may be driven using either a known reference signal correlated with the primary disturbance, or from an error sensor located within the region of interest. In both cases, control can only be achieved when these signals are appropriately filtered, based on the characteristics of the acoustical environment and the positions of the sources and sensors [29]. The same is true for the loudspeaker array processing required to generate personal sound zones; the problem of reproducing an acoustic signal for an individual in a shared space can be reduced to the problem of designing a series of filters, the complex coefficients of which are described as loudspeaker weights. Several processes for generating sound zoning filters have been developed, but these can be broadly categorised into two types: processes that define the listening zones in terms of their energy, such as Acoustic Contrast Control (ACC) [5, 14] and Acoustic Energy Difference Maximisation [53], and methods in which the sound fields in each zone are explicitly defined, such as the Pressure Matching method (PM) [54] and Wave Field Synthesis [32]. A high-level description of these techniques will be provided in the remainder of this section.

Following the observation that active noise control techniques can create a localised quiet region by minimising the acoustic potential energy, the *acoustical bright zone* was defined by Choi and Kim [14] as an area of high acoustic potential energy, compared to the average energy in the wider region of interest. This description of a difference in energy between regions led to the definition of the acoustic contrast, both as a measurement parameter to judge the effectiveness of sound zone creation, and as an explicit optimisation goal in the process that would become known as Acoustic Contrast Control (ACC) [5]. ACC is generalised from multichannel active noise control, and has the objective of maximising the acoustic potential energy in a bright zone whilst placing an upper bound on the energy in a designated dark zone. The dark zone may itself be a constrained region within a larger space, or be defined as all points not contained within the bright zone. In the latter case, the ACC problem can be reduced to that of array

directivity control [2, 55], a subject with extensive literature coverage regarding audio and radio-frequency applications [56]. Loudspeaker arrays are fundamentally limited in their directivity as they must use a discrete number of finite-size transducers to represent a theoretical, continuous distribution of acoustical sources [57]. Further to this, the electrical power available to the array can limit the achievable acoustic contrast [14], and this power limitation is particularly significant in the case of portable devices where over-driving the loudspeakers may cause damage or failure [2]. The derivation of ACC presented in Section 5.2.1 shows that by manually taking the precaution of limiting the array output power via regularisation, systems also become more robust to environmental changes and errors in loudspeaker positioning and sensitivity [58]. In the case of speech privacy control, there is a documented correspondence between the level of acoustic contrast and the degree of speech intelligibility contrast [9] achieved by a system, suggesting that regularised ACC is an appropriate technique for use in this application.

The PM method differs from ACC, as rather than maximising the acoustic potential energy within the bright zone, a target sound field is specified and an optimisation process adjusts the loudspeaker weights to minimise the difference between the target and reproduced fields in terms of magnitude and phase [54, 59]. A mathematical derivation for the filters produced using the PM method is provided in Appendix A. The additional phase constraint in the PM method can result in reduced levels of acoustic contrast, compared to when acoustic contrast is directly optimised for in ACC [60]. However, using PM to explicitly define the target sound field within the bright zone is perceptually beneficial in system geometries where the loudspeaker array surrounds the listeners, as the sound in the bright zone can be made to appear from a steady direction [60, 61]. When this same circular geometry is used with ACC, it has been reported that the uncontrolled phase can lead to the undesirable impression of the sound source swirling around the listener [62], but this effect is greatly reduced when a linear loudspeaker array is used [60]. Line arrays are inherently limited in the range of possible plane wave directions that can be synthesised, so as a consequence, the perceived direction of arrival from a linear loudspeaker array is invariably co-located with the physical loudspeakers[a]. Comparisons of sound field planarity between linear and circular arrays showed that above 580 Hz, planarity scores in excess of 90% were achieved by linear arrays using ACC [16]. Furthermore, the 580 Hz limit is not necessarily indicative of non-planar sound field reproduction below this frequency, and is rather related to limitations in the resolution of the planarity control measure [16, 35].

These experimental comparisons into different sound zoning methods [16, 60, 61] provide an example of how both practical and psychoacoustical insight into a design problem can be used to determine the most appropriate approach for a given scenario. In the case of conventional personal audio systems for use as entertainment devices, the trade-off between acoustic contrast and audio quality is an important consideration. Hybrid sound zoning methods that draw on elements from both ACC and PM have enabled fine-grained control between these two objectives. One example is the Weighted Pressure Matching method; by adding weightings to the control points that demarcate the sound zones, different levels of priority can be assigned to the target sound field, in terms of reproduction accuracy [63, 64]. Conventional bright and dark zones would be represented with weightings of 1 and 0 respectively, with lower-priority *grey zones* taking values in between. The key advantage of this formulation is that the trade-off between

[a] This is true for zonal audio systems where the zones are at least as large as a listener's head. Inter-aural crosstalk cancellation systems can create virtual images away from the loudspeakers but this topic is outside the scope of the present thesis.

directivity and reproduction accuracy in a system with limited input energy can be smoothly varied, perhaps by the user of a system. A similar approach that exploits the mathematical similarities between the ACC and PM methods was described by Chang and Jacobsen [65]. In their formulation, a single parameter could be smoothly varied to adjust the priority given to energy minimisation in the dark zone and minimising the mean-square reproduction error in the bright zone. In an independent, but similar parametrisation of the sound field control problem, described by Lee et al. [66], two parameters could be chosen to control the trade-off between acoustic contrast and reproduction accuracy. The method, known as Variable Span Trade-off filters (VAST), has both ACC and PM as special cases.

The primary objective of the speech privacy control systems described in this thesis is to produce listening zones with a high degree of acoustic separation, thereby improving privacy for the target listener - other considerations such as improving target quality and reducing the potential for noise annoyance are of secondary importance. By virtue of the optimisation cost function used, ACC should produce a higher level of contrast than PM, but in practice, the achievable levels of acoustic contrast are strongly dependent on the level of regularisation used in each method. From a practical standpoint, the most pertinent difference between the two methods from the perspective of the personal audio system designer is the need to supply a target sound field across the zone of interest in the PM method. ACC does not require this additional information as it uses the transfer responses alone. Given that there is no significant perceptual advantage in terms of spatial stability when linear arrays are used [60], the requirement in PM for a target sound field merely adds complexity, by adding another variable that must be considered and justified in the optimisation of the system. This consideration also affects hybrid methods such as VAST [66], the Weighted Pressure Matching method [64] and Planarity Control [16], all of which require the specification of target sound fields and / or additional parameters. Therefore, to minimise the number of parameters that must be optimised, whilst maintaining high levels of acoustic contrast, regularised ACC is selected as the sound zoning technique that will be used throughout this thesis.

2.2 Including Perception

Given that the concept of producing personal audio zones has been proven, recent progress has shifted the focus towards practical implementation [5, 7, 67–70], with many authors quantifying the performance of personal audio systems using metrics based on the human perception of sound, and validating designs using formal listening tests. Based on this focal shift, it is expected that future personal audio systems will integrate an understanding of psychoacoustics into the design of the system, rather than simply assessing their performance using subjective studies retrospectively. Recent work from the Perceptually Optimised Sound Zones (POSZ) project [71] has made steps in this direction, with the aim of producing practical audio systems that provide distinct sound zones, whose performance is evaluated and enhanced based on models of perceived sound quality.

The POSZ project set out to investigate the performance of loudspeaker array-based personal audio systems in realistic rooms, evaluating their performance using perceptually relevant metrics, and ecologically valid stimuli, i.e. with listening tests using programme material representative

of audio heard at home. Publications from POSZ spanned both engineering and psychoacoustical research [71]. Key engineering contributions include methods for selecting an optimal set of loudspeaker positions in challenging acoustical conditions [12, 72] and studies on the effect of regularisation on different zoning methods [16, 73]. These comparisons led to a hybrid of PM and ACC being developed, formalised as the control of the sound field's planarity [35, 74]. Rather than explicitly specifying a planar sound field to reproduce in a zone, as in PM, Planarity Control describes a window of acceptable angles for plane waves. The mathematical formulation becomes identical to ACC if a 360° angular window is specified, and similar to PM if only one angle is selected.

Psychoacoustical investigations by the group have centred around matching results from listening tests to perceptual models, with the goal of ultimately being able to automate design decisions such as loudspeaker placement [18] and the choice of sound zoning method [60]. Distraction was identified as the most relevant attribute to listeners subject to audio-on-audio interference which sound zoning methods seek to alleviate [3]. This led to the design of a distraction model consisting of a range of perceptual attributes [75]. While this represented a significant improvement on previous models [76], which performed poorly against validation datasets, the authors indicated that further improvement may be afforded by extracting higher level contextual features from audio and speech and reducing the computational complexity of the distraction model, so that it could be applied in real-time [77]. The quality of speech material was evaluated in terms of the acceptability of an interfering programme, a quantity which was well predicted by the SNR [13, 78], though significant differences in results were found with different types of interferer programme material, a conclusion which was common to many of the studies conducted by the group.

The VAST framework [66] described above has also been extended to include a perceptual weighting function and the facility to adapt the control strategy based on the signals input into each zone [17]. Heuristically speaking, these modifications allow the system to exploit the masking effect of the target signals by permitting greater leakage from other zones in the parts of the spectrum that are masked. Analysis of the reproduced signals in anechoic and weakly reverberant environments showed that these amendments to the VAST method improved perceptual metrics and maintained objective performance indicators, such as acoustic contrast. The group have only carried out informal listening studies to date, so no comment can be made regarding the magnitude of the perceived improvement, compared to the additional task of selecting the requisite parameters.

The speech privacy control systems described in this thesis may be similarly optimised based on the properties of the speech sound fields which they must reproduce. The typical frequency range and dynamic range of human speech and music [79] are represented as regions in Figure 2.3. The restricted domain of speech compared to music demonstrates an opportunity that may be harnessed when designing private personal audio systems, namely, that system performance can be concentrated into regions occupied by speech. As well as having a characteristic frequency and intensity distribution, speech is also distinguished from other types of acoustical signals in terms of its modulation spectrum; octave-band specific modulations between 0.5 and 16 Hz have been shown to be important for carrying speech information [80]. The depth of these modulations may reduced by the addition of noise or through temporal smearing of the signal, either due to room reverberation or poorly designed sound zoning filters [81]. In the speech privacy control

problem, these effects must be analysed across both listening zones, as a beneficial increase to the intelligibility in the bright zone is detrimental to the objectives in the dark zone, and vice-versa.

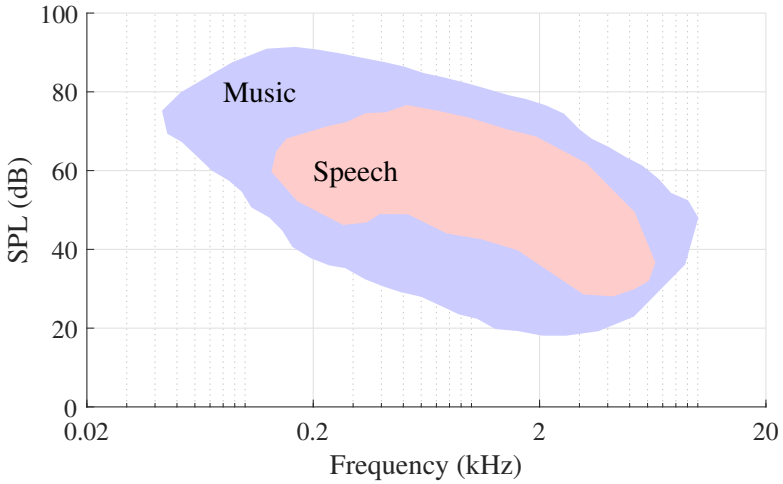


FIGURE 2.3: Representation of the frequency range and dynamic range of speech. Data from Ref. [79]

Donley and Ritz [82] have approached the problem of reproducing multi-zone speech sound fields by considering the perceptual relevance of speech material leaking from one zone into another. They argue that the perceived performance of a personal audio system can be adversely affected if the conventional constraint of setting the dark zone pressure to zero is applied [82]. Alternatively, they suggest allowing a certain level of leakage from the bright zone into the dark zone, and using the resulting improvement in system efficiency to improve the quality of reproduction. Using this approach, a similar trade-off between reproduction quality and inter-zone separation to that found by Olivieri et al. [64] was found. By adjusting this balance to minimise reproduction error, with the leakage matching a standard human hearing threshold curve, spatial inaccuracy and mean-square error were significantly reduced compared to previous methods, while using less electrical power.

Each of these contributions indicate that it is possible and indeed beneficial to incorporate an understanding of masking and the functions of the human auditory system into a personal audio system design. While the integration of human sound perception is merely beneficial in the design of conventional personal audio systems, speech privacy control systems *must* integrate an understanding of speech perception into their design. Otherwise, the claim of privacy cannot reliably be made. The following section describes recent research into this specific application.

2.3 Speech Privacy Control

The scope of this thesis is restricted to investigating the physical and perceptual aspects of the speech privacy control problem, i.e. the private delivery of a spoken message to a target listener in a shared space. Much of the literature regarding the control of speech privacy relates to understanding and improving the experience of workers in open plan office environments (e.g. [83–85]), but there are many parallels between this situation and the problem of providing private sound zones. Open-plan offices can be made more appropriate for private work by

arranging the location of workstations based on instrumental predictions of speech intelligibility [86, 87], and through the introduction of artificial masking noise [88–90]. Likewise, in a speech privacy control system, the performance is dependent on the desired locations of the bright and dark zones with respect to the loudspeaker array, and any additional masking sound must be designed in such a way that it delivers the required level of intelligibility reduction, without itself becoming distracting or annoying. This objective is shared in open-plan office sound masking design [84, 90, 91].

As the field of personal audio has developed, terminology has gradually evolved for the description of the sound zones produced by a personal audio system. In a number of publications the word "private" has been used, e.g. "private sound" [15, 64], "private listening zone" [92], "private audio system" [93] and "private listening space" [94]. However, none of these referenced examples include either subjective or metric-based verification for the claim of privacy. Instead, they implicitly assume a link between inter-zone separation and privacy.

The work of Donley and Ritz into perceptually weighted multi-zone control [82] was later developed through the use of objective speech intelligibility metrics to directly evaluate inter-zone privacy [8]. A filtered random noise masking signal was focussed into the dark zone in order to improve speech intelligibility contrast without significant degradation of the bright zone signal, measured using PESQ [95]. The simulated system used in Ref. [8] to demonstrate this consisted of a circular array of 295 loudspeakers. Further work by the same authors [9] has shown how performance varies with more practically sized linear arrays, including experimental validation of their simulations in an anechoic chamber, which show good agreement. Low-pass filtering of the masking signal was proposed as a means to improve performance by reducing the level of high frequency grating lobes which may impinge on dark zones. The paper does not present results of listening tests to support the use of sound quality metrics or specify in detail the choice of intelligibility level, determined using the Short-Time Objective Intelligibility metric [96], that corresponds to confidential speech privacy levels. Furthermore, the perceptual implications of radiating a masking signal into the reproduction environment are not investigated. Nevertheless, the line of enquiry taken by Donley et al. is the closest in scope to that presented in this thesis found to date, and provides a valuable reference approach for comparison with the methods developed in this thesis.

In addition to the work on privacy control using conventional loudspeaker arrays, two alternative methods for private speech transmission have recently been proposed [97]. Rather than being degraded by room reverberation, these methods are wholly dependent on the presence of multiple echo-paths between a small set of distributed loudspeakers and the positions of target listeners in a room. In the first method, impulse response measurements are used to create appropriate filters, which are applied to Gaussian noise bursts or short sections of target speech [98]. After echoing along the multiple paths in the room, these signals recombine at a target point to yield the desired speech signal, and are uncorrelated elsewhere, resulting in poor intelligibility. The second method is inspired by jamming eavesdroppers in wireless communication systems; specially crafted noise signals are radiated in addition to target speech such that the noise signals destructively interfere at target points whilst successfully masking messages elsewhere. Measurements showed that speech intelligibility at a measurement point 10 cm away from the target location was significantly impeded, on the one hand showing excellent privacy provision,

yet also indicating that the robustness of the system to environmental changes or even the movement of people or furniture within the room could be extremely limited in practice.

In this thesis, conditions relating to privacy will be defined using the results from objective speech intelligibility testing and correlation with speech intelligibility metrics. This follows the approach taken in research into open-plan office privacy. A key finding from this research is that seemingly high levels of intelligibility, measured as the percentage of correctly identified words in a listening test, can still correspond with subjectively acceptable levels of privacy. For example, Bradley and Gover [99] compared subjective ratings of privacy against the results of sentence intelligibility tests and instrumental predictions using the SII metric. They found that in conditions characterised by an SII value of 0.15, privacy was rated as "Acceptable" on average, but 50% of respondents still achieved greater than 40% words correct in a test composed of low-predictability Harvard sentences [100].

It is important to note that an "Acceptable" level of privacy in an open-plan office is conceptually different to that expected between closed rooms [101], and likewise, the wide range of potential applications for speech privacy control systems also results in different levels of privacy expectation. Nevertheless, personal audio technology offers a method to directly influence the sound fields in each listening zone, and can therefore be reasonably expected to provide a greater degree of privacy than the simpler, global approaches that are conventionally used in the sound design of open-plan offices, such as electronic sound masking and passive acoustic treatment. One way to impose this stricter expectation of privacy is to set a significantly lower threshold for the percentage of correctly identified words, for example, 5%. However, carrying out speech tests to accurately assess this intelligibility level has a number of difficulties, such as the effects of subject fatigue after listening to many poorly intelligible sentences. The proposed solution to this problem is to conduct an intelligibility test with a restricted set of possible words, such as a matrix sentence test [102]. The additional information given in a matrix test results in higher intelligibility scores for the same test conditions; a 50% correct score in a matrix test corresponds to a score of 5% correct in a test of connected sentence comprehension [103]. This correspondence between different speech tests is illustrated in Figure 3.7, and the process of converting between various descriptors of intelligibility and privacy are described in detail in Sections 3.2 and 3.3. In Chapter 7, this definition of privacy, i.e. 50% words correct in a matrix test, is converted to an equivalent SII value, which is used in later chapters to constrain the maximum intelligibility allowed within the dark zone.

2.4 Summary

Personal audio systems use loudspeaker array processing to generate spatially separated listening zones in shared spaces. The applications of personal audio technology include shared office spaces and museum exhibits [4], targeted advertising [104], television systems [5, 6], headrest-mounted loudspeaker systems [67] in-car entertainment [7] and mobile devices [2]. Recently, the use of loudspeaker array-based systems for private speech reproduction has also been investigated [8, 9], demonstrating that by introducing an additional masking signal in a multi-zone reproduction setup, it is possible to produce listening zones with distinct levels of speech intelligibility.

There are two key practical barriers to the implementation of systems designed with this specific purpose in mind that are not necessarily evident from the general study of personal audio system design. Firstly, in conventional personal audio systems, the degree of separation between listening zones can be measured objectively using the acoustic contrast metric [14], whereas in speech privacy control, the objective is instead a contrast in the speech intelligibility between the bright and dark zones. Conducting intelligibility measurements using human subjects is time-consuming and expensive, and thus is impractical to include in a system design process. However, metrics can be used to convert between objectively measurable or predictable properties of the zonal sound fields and the intelligibility of speech. This approach is discussed in detail in Chapter 3.

The second barrier to successful implementation of a practical private audio system concerns how the masking signal is perceived. This additional masking is necessary to achieve sufficient levels of speech privacy with practically-sized loudspeaker arrays, but could prove annoying or distracting to nearby listeners if their viewpoint is not considered. Similar to the intelligibility measurement discussed in the previous paragraph, this type of perceptual evaluation is impractical to carry out directly using human trials for all system implementations and designs. Nevertheless, psychoacoustic metrics can be used to obtain an estimate of the perceptual effect of making certain design changes. Several of these metrics are presented, and a scheme for combining them is described in Chapter 4.

Chapter 3

Objective Metrics for the Assessment of Speech Intelligibility and Privacy

In the early years of the 20$^{\text{th}}$ century, the rapid development and distribution of telephone systems demanded sudden progress in the science of speech perception. Manual procedures for testing the communication quality and accuracy of prototype telephone systems were suggested as early as 1910 [105] and were standardised by Bell Telephone Laboratories in 1929 by Fletcher and Steinberg [106], two authors who would go on to make many significant contributions, not only within the field of speech perception, but the understanding of the human hearing system in general. In 1947, French and Steinberg published an article in the Journal of the Acoustical Society of America entitled "Factors Governing the Intelligibility of Speech Sounds" [107]. The article summarised several years of research into the basic relationships between the physical properties of speech signals and their intelligibility, and was the first publication to formalise the concept of an "Articulation Index" (AI); a means to predict the intelligibility of speech sounds from objective measurements alone. The multitude of national and international standards that still use the AI, e.g. [108–111], almost completely unmodified from its original 1947 formulation, are testament to the significance of Fletcher, French and Steinberg's experimental results.

Accurate instrumental measures of intelligibility and privacy are essential to the development of private personal audio systems, as these metrics can be used to set target intelligibility levels within the bright and dark listening zones, which in turn can be used to decide the level and spectral shape of the masking signal produced by the system. The alternative approach to acquiring this information involves carrying out speech intelligibility tests using human subjects. This costly and time-consuming process can yield accurate information regarding the performance of a single system, but extensive testing is required to generalise these results to new designs. Instrumental intelligibility metrics, on the other hand, can be rapidly calculated based on simulations of prototype systems. This provides near-instant feedback on how changes to the design of a system, the choice of masking signal, or the reproduction environment affect the intelligibility of speech, and hence the privacy, in each zone.

The present chapter begins by describing the key features of the AI metric and subsequent refinements that have been designed for specific applications. A selection of more recent alternatives to AI that rely on more complex computational models and signal processing techniques are also discussed. Two methods of converting between objective SNRs and intelligibility scores are described - a direct method based on listening tests, and the preferred indirect method which uses objective intelligibility metrics as an intermediate step. The chapter concludes with an overview of metrics and standards that have been used to quantify how intelligibility is related to the concept of speech privacy. These connections form an essential component of the masking signal design process described later in Chapter 7.

3.1 Objective Intelligibility Prediction

The intelligibility of speech under certain acoustical conditions can be defined and measured in many different ways. The basic premise is to carry out a blind test in which certain items of speech are presented to a listener, who must then identify what has been presented; the percentage of correct responses in the test is defined as the intelligibility rating, or score for that particular listener and test condition. Speech intelligibility tests can assess the intelligibility of individual phonemes [105, 112], single words [113, 114], words in the context of a sentence [115, 116], or even complex ideas across a short passage of speech [106, 117], depending on the quantity of interest. A more thorough review of speech intelligibility testing is provided in Chapter 6, as the present section focuses on instrumental speech intelligibility metrics.

An important distinction must be made at the outset between the output of speech intelligibility metrics and the results of speech intelligibility tests. As described in the previous paragraph, intelligibility has numerous definitions and, due to redundancies in spoken language, it is easier to understand whole words within the context of a grammatically correct sentence, than it is to understand single, nonsense syllables in the same acoustical conditions. Therefore, the intelligibility score in a given listening test is a function of the testing format, in addition to the particular degradations to the speech signal, such as noise, reverberation and distortion. These types of degradations can all be quantified by objective intelligibility prediction metrics, but their outputs are only intended to be monotonically related to the listening test result. Most intelligibility metrics actually assess the *audibility* of speech cues [118], and a mapping function is required to convert the output of a metric into results that are comparable with intelligibility scores found using a particular listening test format. This mapping is known as the Audibility-Intelligibility Transfer Function (AITF) [119].

A related but distinct transfer function commonly encountered in the study of speech perception in noise is the *psychometric function* - a mapping between the SNR and the intelligibility score in a listening test. Like the AITF, the psychometric function is dependent on the speech material and the listening test format [120]. To illustrate the differences between the two functions, and to show how they may be obtained in practice, Figure 3.1 shows a diagram of an example test where the listener correctly identifies three out of five keywords from a sentence presented in noise. The same evaluation is also performed by a speech intelligibility metric. In this example, the metric uses a reference copy of the speech signal and the degraded signal to produce its result, though other modes of operation are available and these are discussed below. Repeated

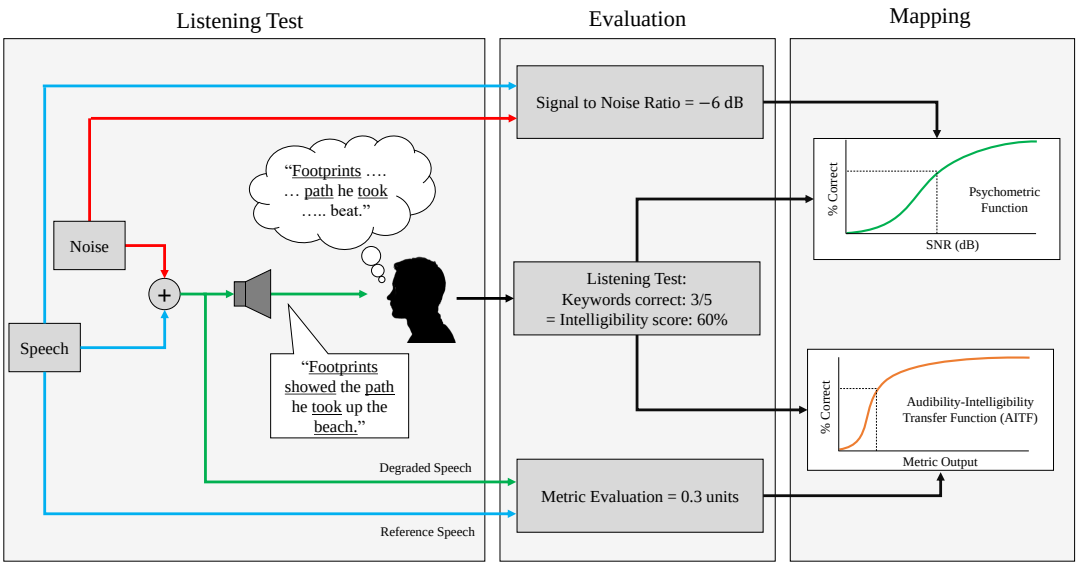


FIGURE 3.1: Example test methodology for the production of a psychometric function and an AITF.

application of this process at a range of SNRs will yield a psychometric function and an AITF by combining the listening test results with the input SNRs and the metric results respectively.

The example metric shown in Figure 3.1 uses a copy of the reference signal and a measurement of the degraded signal to evaluate the intelligibility. This approach is taken by a number of intelligibility metrics, such as the (Extended) Short-Time Objective Intelligibility metric, (E)STOI, [96, 121]. Other metrics, known as "single-ended" measurements, use only the degraded signal, and assumed properties such as the expected spectrum of speech. The Speech Transmission Index (STI) metric [122, 123] uses this technique. The AI [108], and its successor, the Speech Intelligibility Index (SII) [124] operate using the spectra of speech and noise in isolation, but information is provided on the expected spectrum of speech for normal, raised, loud and shouted speech, which can be substituted for a speech spectrum measurement should one not be available.

The approaches referenced above each require different levels of information about the transmission channel under test. Single-ended measurements require the least information, and are thus suited to implementation in the widest range of applications, at the expense of accuracy, due to the assumptions made about the spectrum of speech. Metrics that require isolated samples of speech and noise are more suited to research and audiological applications where this can be provided. Although the applications of personal audio technology referenced in this work may include real-time sound transmission equipment, for the purposes of research, or configuring a system in advance of real-time use, it is assumed that a clean reference copy of the audio to be transmitted is available in advance, and that similarly, isolated samples of the noise received in each zone can also be measured or calculated. Correspondingly, all categories of metric are potentially applicable to the present work. The following subsections describe the operation of several speech intelligibility metrics, to determine the metrics that are most fit for purpose in the design of private personal audio systems.

3.1.1 Articulation Index and Speech Intelligibility Index

AI [108] and its successor SII [124] are designed to produce intelligibility ratings based on three frequency dependent factors: the spectral level of the speech, the spectral level of any noise or equivalent degradations due to distortion and reverberation, and the hearing threshold of listeners. Figure 3.2 shows a simplified block diagram of the SII calculation process for conditions where reverberation is negligible. These conditions are similar to the range of applicability of the older AI standard, and except for some adjustments to the model coefficients, the metrics are essentially equivalent when reverberation can be neglected [125]. The SII, pictured at the right of the block diagram is composed of a sum of Band Audibility Functions, A_i, weighted by the relative importance, I_i, of each frequency band to speech intelligibility. The standard allows for various divisions of the frequency spectrum into bands, for ease of computation or finer-grained detail. The audibility of speech cues in a frequency band is dependent on the distortion of the speech from a standard speech spectrum, L_i, and on a generalised SNR, K_i. The "noise" portion of K_i is formed by comparing the masking effect of the additional noise and the speech itself, Z_i, with the average hearing threshold of the population, X'_i, to form the Equivalent Speech Disturbance Level, D_i.

The noise levels input to the model, N'_i, must be measured and corrected for listener position and whether a monaural or binaural intelligibility rating is required. In the absence of measured speech levels, the standard provides estimated spectral levels for four vocal effort levels. A similar correction procedure gives the equivalent speech spectrum level, E'_i. The metric also offers flexibility on the definition of the equivalent hearing threshold level, T'_i, either from a series of measured hearing thresholds, or based on the reference hearing threshold level, defined as 0 dB across frequency.

The SII differs from the original AI by improving the modelling of the upward spread of masking from low frequencies to high frequencies, and recognising that different patterns of speech, such as full sentences, phonetically balanced words and nonsense syllables, have different band importance functions [125]. A major difference comes in the handling of reverberant conditions, which were not considered in the original AI due to its origins in the evaluation of telephone systems. In conditions where the reverberation is significant, the values of the model inputs E'_i, N'_i and T'_i are calculated using a different method based on the preservation of intensity modulations by the transmission channel. This model is inspired by the Speech Transmission Index standard [122, 123], which is described in detail in the following subsection.

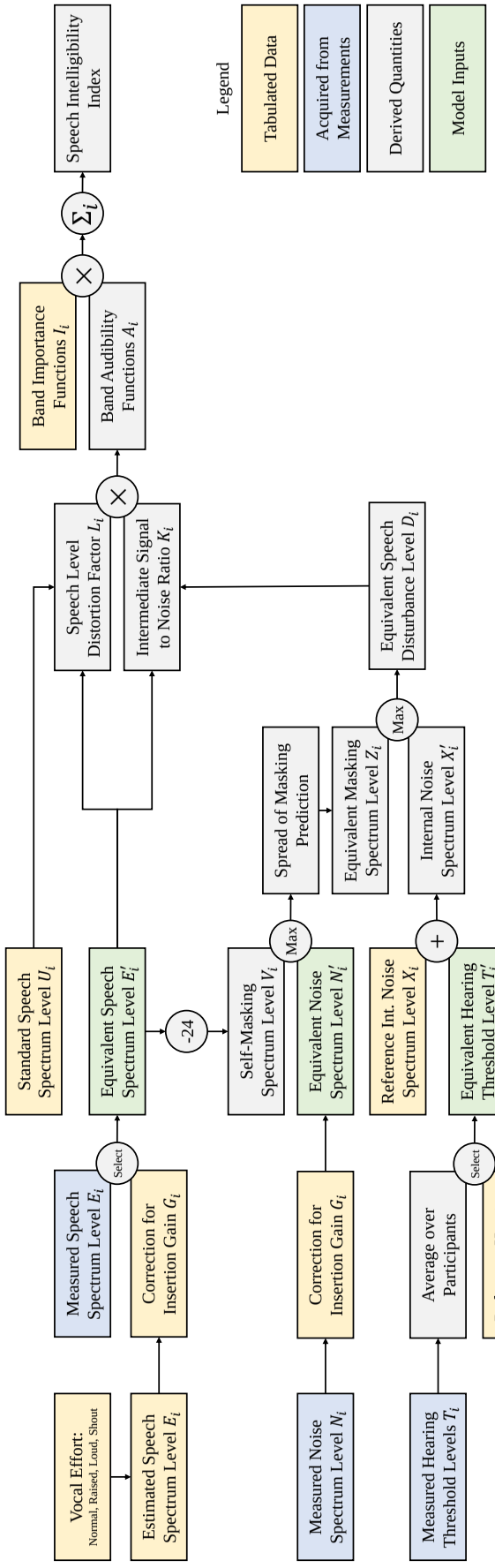


FIGURE 3.2: Block diagram of SII calculation process, in the case of negligible reverberation, as described in Section 5.1 of Ref. [124]. Variables subscripted with an i are calculated for each frequency band.

3.1.2 Speech Transmission Index (STI)

The Speech Transmission Index (STI) is a single-number rating that quantifies the effect an acoustical or electrical signal path will have on the intelligibility of speech signals. It is codified in IEC 60268-16:2011 [122], and many example applications are provided, such as the measurement of public address systems, communication devices, and direct (i.e. unassisted) speech communication in reverberant spaces. Three methods are presented in the standard, but the standard states that,

> "None of the methods are suitable for the measurement and assessment of speech privacy and speech masking systems, as STI has not been validated for conditions that represent speech privacy applications". [122].

This suggests that the STI is unsuitable for use as an intelligibility metric for the present work, as this measure of intelligibility is intended to be used as a proxy for privacy. However, the STI is directly used in the 2012 revision of ISO 3382-3 [36] to assess the distance at which speech privacy can be claimed in open plan offices; in this standard a value of STI < 0.2 is deemed to correspond to private communication conditions. ISO 3382 supports this with references to two publications, from 2005 [126] and 2007 [127]. These references predate the current revision of the STI standard, IEC 60268-16:2011. Two international standards, therefore, appear to disagree over the validity of the STI for privacy measurements. Nevertheless, the techniques and background materials used in the STI calculation provide important information about the effects of various types of audible degradation on human speech perception, so a brief review of the STI method is provided below.

STI is founded upon the concept that speech intelligibility is maximised when the intensity envelope of speech is preserved by a transmission channel. This time domain approach contrasts with the operating mode of the AI, which is based on the SNR in each frequency band. Rather than comparing a clean input speech signal with the output from the transmission channel, the full version of STI [122] uses a set of amplitude modulated noise signals to form a modulation transfer function matrix, which is then processed to yield the STI. Figure 3.3 shows a block diagram of the method.

The laborious frequency-by-frequency analysis required by the original implementation of STI was later replaced by an indirect method, which uses the impulse response of the transmission channel to populate the modulation transfer function matrix, hence characterising signal degradation with a single measurement. This indirect measurement method is often used in building acoustics to assess the need for, or effectiveness of, public address and voice alarm systems [122, 128], but it relies on the assumption that the transmission channel is linear and time-invariant. This fast, indirect method would be inappropriate for use in many speech privacy control settings, as it is possible for the masking signal added to the target speech to vary with time. In this case, the degradation to the speech cannot be represented as an impulse response.

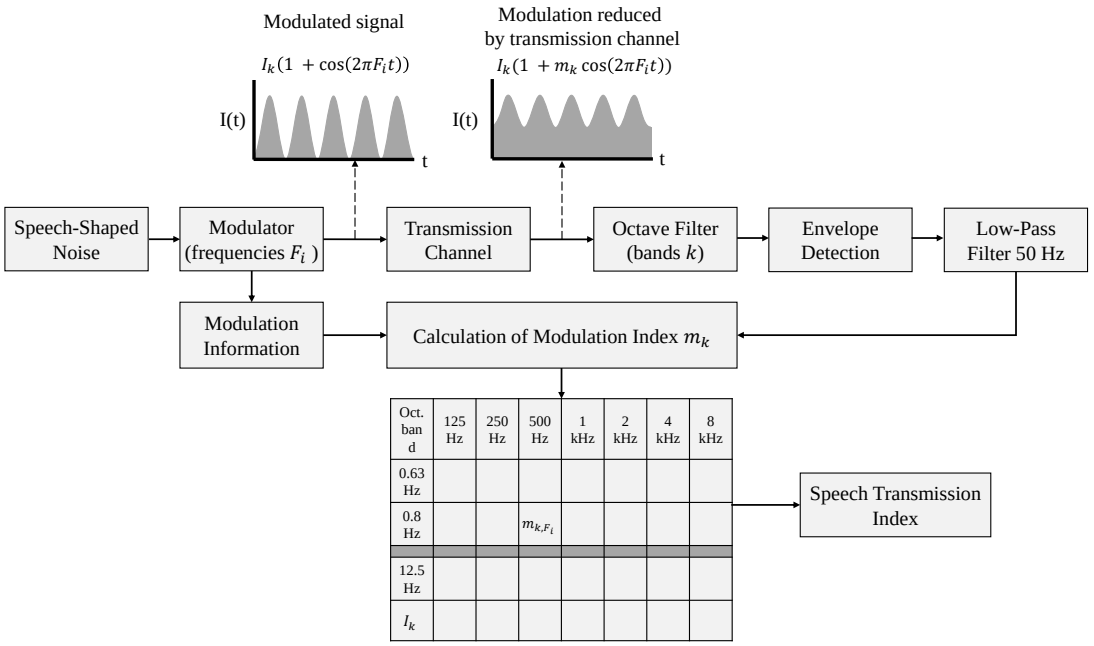


FIGURE 3.3: Block diagram for the calculation of the STI [122] associated with a transmission channel.

3.1.3 Metrics for Fluctuating Degradations

Metrics such as AI and SII, which use the global statistics of speech and noise to form their result, tend to underestimate the intelligibility in cases where the masking noise fluctuates with time, such as when the masking consists of one competing talker. Listeners with normal hearing are able to take advantage of these spectrotemporal "dips" in the masker, where the SNR in a narrow frequency band is, for a short duration, much higher than the average across the signal [129]. This phenomenon was formulated into the "Glimpse Proportion" metric by Cooke [130]. The metric compares modified spectrograms of speech and noise signals, and characterises "glimpses" as regions of sufficient area on these spectrograms where the target speech dominates the noise. The conversion function between the proportion of valid glimpses and the resulting intelligibility score was obtained using objective speech intelligibility tests. The glimpse proportion metric was later extended to include a frequency dependent distortion weighting [131], recognising that as well as restricting the audibility of a target speech signal, additive noise also reduces intelligibility by producing a distorting effect [132]. The result was an improvement to the performance of the metric under a wider range of steady and fluctuating masking conditions. A further extension enabled the prediction of binaural intelligibility by considering the distortion weighted glimpses in each ear, thereby accounting for the effects of better ear listening and binaural unmasking [132]. When evaluated in anechoic conditions with speech-shaped noise and competing speech maskers, the metric showed better predictive power to four other binaural speech intelligibility metrics. In reverberant conditions, the binaural glimpse proportion metric showed greater consistency between different types of masker when compared against alternative metrics [133].

Another refinement to the glimpse proportion metric is provided by the Short-Time Objective Intelligibility (STOI) algorithm, which was developed in 2011 [96] in response to a need for an improved objective measure of intelligibility in the case of fluctuating degradations, including non-linear processing, which the original glimpse proportion metric was unable to resolve [121].

In contrast to SII, a measurement of the additional degradation is not required in isolation by STOI. Rather, a clean reference copy of the speech, and the speech signal corrupted by noise are input to the algorithm. STOI divides these signals into short time segments of 384 ms to capture the effect of short-duration or time-varying degradations. This value is chosen to reflect the temporal integration time of the human auditory system. Frequency resolution is provided by further subdividing these time-windows into ⅓-octave frequency bands, and the metric forms its intelligibility estimate by calculating the correlation coefficient between the reference and degraded signal envelopes in each sub-band, then averaging the result across frequency. This final step implies independent contributions to intelligibility from each band, an assumption which is in opposition to other established metrics, such as SII, which recognise differences in the importance of each frequency band, and the upward spread of masking.

To rectify this assumption, an extension to the STOI algorithm was published in 2016. This Extended STOI (ESTOI) algorithm [121] trades off a slight increase in computational complexity for a further reduction in the assumptions about the form of any signal degradation. In ESTOI, rather than calculating temporal correlation coefficients, then averaging across frequency, the operations are transposed - spectral correlation coefficients are calculated for each sub-band envelope, before being time-averaged across each frame. Figure 3.4 shows a block diagram of the ESTOI algorithm, and the reference implementation of the ESTOI algorithm produced by Jensen and Taal is available at Ref. [134].

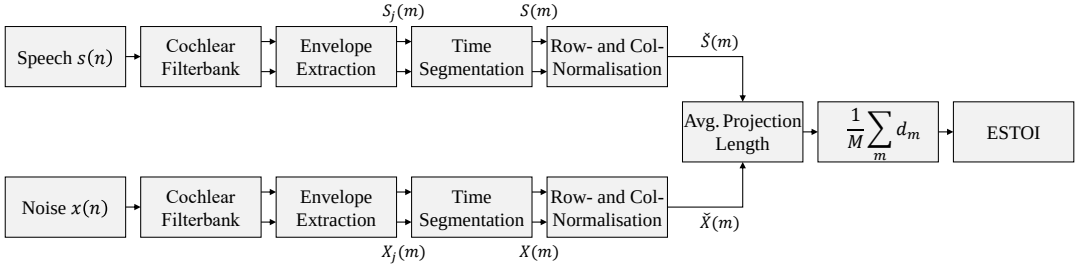


FIGURE 3.4: Block diagram of the ESTOI algorithm [121].

First, the clean and processed digital signals, $s(n)$ and $x(n)$ are passed through a ⅓-octave filterbank, then the temporal envelopes are extracted. The resulting spectrograms $S_j(m)$ and $X_j(m)$ are divided again into shorter time segments before being normalised in time and frequency to form the matrices $\check{S}_m$ and $\check{X}_m$ for the clean and processed signals respectively. The intermediate intelligibility indices for each time-segment, d_m, are formed by calculating the vector projection of the columns of $\check{S}_m$ onto the columns of $\check{X}_m$. If the degradations to the reference speech signal are insignificant, then the $\check{S}_m$ and $\check{X}_m$ matrices will be similar and the row and column normalisation will result in a value of d_m close to 1. This indicates that 100% of speech cues are audible in the degraded or processed signal, and therefore the speech is fully intelligible. In high-noise situations, the degraded and reference signals are essentially uncorrelated, i.e. the vectors formed by the columns of $\check{S}_m$ and $\check{X}_m$ are orthogonal, resulting in a value of $d_m \approx 0$. Negative values of d_m are mathematically possible if the projections of $\check{S}$ and $\check{X}$ are in opposite vector directions, but this is extremely uncommon in practice and the input signals $s(n)$ and $x(n)$ would have to be specifically constructed to have these properties.

Internally representing signals as spectrograms in the STOI and ESTOI algorithms preserves both time and frequency information. Consequently, these methods, and in particular, ESTOI, are more capable of predicting the intelligibility of speech degraded by strongly time-varying signals than other methods [121]. By its authors own admission, ESTOI does not attempt to model the physics of the human auditory system, instead making algorithmic choices based on the loss of speech information that a particular degradation may cause [135]. This enables the algorithm not only to capture the effect of the energetic masking that degrades a signal, but also the loss of intelligibility due to informational masking [136, 137]. The perceptual design requirements in certain speech privacy control scenarios may necessitate the use of time-varying masking signals, for example to account for changes in the background noise level, or to match the characteristics of a certain type of ambient background noise that is already present in a space. In these situations, an intelligibility metric that is consistent under many different forms of additive degradation is required; ESTOI fulfils this demand.

3.2 Conversions Between Metric Outputs and Intelligibility

The speech privacy control systems described in this thesis use masking signals to provide privacy for their target listeners. A key consideration in the design of such systems is the relationship between the level of a particular masking signal, which controls the SNR, and the intelligibility reduction that is achieved in the dark zone. This relationship, between SNR and intelligibility, is characterised by the psychometric function, as described above in Figure 3.1. On the surface, this observation suggests that intelligibility metrics are unnecessary, as the transfer function between SNR and intelligibility could be looked up in the literature, solving the problem in a single step. However, while there is a wealth of published psychometric functions available, these are all specific to particular listening test formats, sectors of the population, speech corpora and types of noise [120]. Furthermore, as will be made clear in Chapter 5, the use of a loudspeaker array and the associated signal processing necessarily implies bandwidth limitations and other artefacts to both the speech and noise signals, which in turn will affect the conversion between SNR and intelligibility. Finally, even if an appropriate psychometric function could be found, this cannot provide information on the perceived privacy, only the intelligibility.

Figure 3.5 graphically displays the links between the SNR, the output of metrics, intelligibility and privacy. Each link is described by a transfer function or qualitative relation, and each is dependent on certain variables being fixed. For example, to fully specify the aforementioned psychometric function, which can be used to convert SNRs into intelligibility scores, the speech corpus, listening test format and the type of masking must all be known. In other words, a psychometric function is specific to a given combination of these variables.

Speech intelligibility metrics are useful as they provide a secondary path to convert between the level of degradations to a speech signal, i.e. the SNR, and the resulting intelligibility. To provide this relationship, two conversion steps are required, as shown by the red arrows in Figure 3.5. The first step describes how changes in the SNR affect the output of the metric, and for most metrics this is dependent on the audibility of speech cues within a signal [118]. The second step describes the relationship between the output of the metric and the true intelligibility, in terms

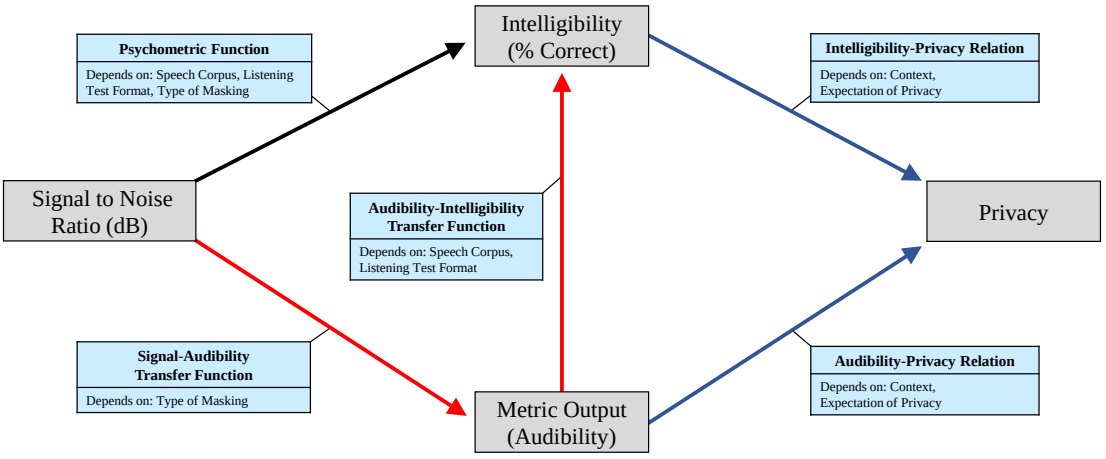


FIGURE 3.5: Links between variables in the Speech Privacy Control problem. The dependencies of each link, i.e. the conditions that must be specified to fully define the function or relation, are provided below the name of each link.

of correctly understood syllables, words or sentences. In the literature, these two relationships are rarely addressed together within the same publication, and there is therefore no standard terminology to disambiguate between them. Commonly, both conversions are simply referred to as the "transfer function", either between a given SNR and the output of a metric, e.g. [138], or between the output of a metric and the percent-correct score, e.g. [103, 124, 139].

To avoid confusion with the electroacoustical transfer function, which describes the transmission path between the input to a loudspeaker and the output from a microphone, the links between the variables discussed in this section will be described using distinct terms - these are given in the blue boxes in Figure 3.5. The transfer function between the output of a metric and the percent-correct score has been described by Scollie as the Audibility-Intelligibility Transfer Function (AITF) [119], as most intelligibility metrics quantify the proportion of audible speech cues in a degraded speech signal [118]. This relationship is in principle independent of the type of signal degradation, as this effect should be absorbed into the metric. The AITF is therefore only specific to the type of intelligibility which is required, e.g. syllables, words or sentences. Following the same naming convention, the transfer function between the SNR and the output of a metric will be termed the Signal-Audibility Transfer Function (SATF). The SATF is specific to the type of degradation, i.e. the spectral and temporal composition of the masker, but is independent of the speech content. By combining the two functions, a psychometric function can be generated.

The two-stage approach to construct the psychometric function from the SATF and AITF is extremely useful in the design of personal audio systems where a particular level of speech intelligibility is required in each zone. This can be visualised in Figure 3.5 by following the red arrows from the SNR to the intelligibility score, via the metric output. The task of a personal audio system designer is to specify the SNRs in the bright and dark zone at which appropriate levels of intelligibility are achieved for each listener. Without the use of speech intelligibility metrics, separate psychometric functions would need to be constructed from listening test data, for each type of masking noise that could potentially be employed by the system. Additionally, any future modifications to the system that could affect its bandwidth, or other features associated with the transmission quality, would invalidate all previous results as these factors would

have been incorporated into the psychometric function. The solution is provided by codifying the requirements of the system in terms of speech intelligibility metric values in each zone, rather than quoting SNRs directly. With this approach, standard AITFs from the literature can be used to convert between metric values and intelligibility scores, and SATFs for new masking signal designs or system configurations can be rapidly generated using simulations. This is the basis of the masking signal design process described later in Chapter 7.

Despite this major advantage, carrying out the conversion between SNR and intelligibility using two transfer functions introduces uncertainty. While most intelligibility metrics are deterministic, the output of the SATF may vary with the fine structure of input signals, such as with different sentences or random masking noise samples, even if the SNR is identical. This random variability is also seen in the algorithms for evaluating perceptual quantities such as roughness and fluctuation strength, which are discussed in Sections 4.3 and 4.4 respectively. The AITF also possesses inherent uncertainty, as it is defined based on the statistical results of listening tests; some sentences are easier for humans to understand in continuous noise than others, for example due to the familiarity of words.

To illustrate the variability of the SATF, Figure 3.6 shows the outputs of the STOI, ESTOI and SII algorithms applied to speech degraded by speech-shaped noise with a 50 dB variation in the SNR. Each data point represents the intelligibility of a different recorded sentence from the Harvard Sentence Corpus [21]. The data is fitted to a logistic function of the form

$$y = Q/(1 + e^{-\alpha(x-x_i)}) + u. \tag{3.1}$$

where x is the SNR and α and x_i are parameters that control the rate of decay and the position of the symmetric inflection point of the logistic respectively. Q is a scaling factor and u shifts the logistic, for cases where the standard range of the logistic, between 0 and 1, does not match the data to be fitted, as can be seen in the rightmost panel of Figure 3.6, where the STOI metric outputs a value around 0.4 for -30 dB SNR. This value represents the lowest value of the STOI metric for additive noise, but lower values of the STOI metric can be obtained when the speech signal is processed non-linearly. For each SATF shown in Figure 3.6, 95% observation bounds are given for the fitted logistic curves - this indicates the likely range of metric values associated with each SNR, when steady, speech-shaped noise is used to mask speech. In terms of metric values, the SII has the smallest confidence intervals, as the output of the metric is only dependent on the overall spectra of the input signals, and is thus less affected by changes in the fine structure of the random noise or the selected sentences. The central plot in Figure 3.6 can provide an approximate indication of how the SNR is related to the SII in each zone of a speech privacy control system. In order to achieve SII = 0.05 in the dark zone and SII = 0.75 in the bright zone, a difference in SNR of 25 dB between the zones is required. Using compact loudspeaker arrays, this level of acoustic contrast is difficult to achieve with a single sound zoning process. However, with two independent sound zoning processes for the speech and masking signals, as shown in the block diagram in Figure 1.2, the required SNR difference can be shared between the two processes. In a system with symmetrical bright and dark zones, this halves the acoustic contrast level (in dB) required by each process. This analysis is only approximate as it does not include the effects of the sound zoning process on the frequency response and temporal structure of each

signal, but it does verify the significant advantage of the proposed method over conventional beamforming techniques.

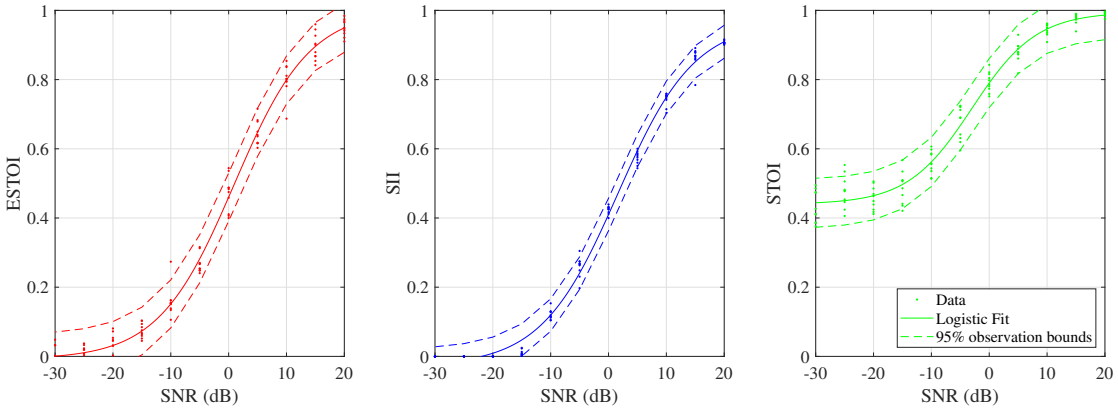


FIGURE 3.6: Signal-Audibility Transfer Functions between SNR and three objective intelligibility measures. Target material is spoken Harvard sentences from the Hurricane natural speech corpus [21], and interferer is speech-shaped noise.

It is important to note that the indices presented in Figure 3.6 are not numerically comparable with one another; for example, an ESTOI value of 0.3 does not necessarily represent the same level of intelligibility as STOI = 0.3 or SII = 0.3. Instead, each metric must be mapped to the true intelligibility score using independent AITFs, as described in Figure 3.1. A selection of AITFs for the SII metric are presented in Figure 3.7. The wide variation in AITFs for different speech materials displayed in Figure 3.7 shows how redundancies in spoken language can assist with the understanding of speech in challenging acoustical conditions. Achieving a score of 50% correct in familiar sentences, or restricted sets of words such as the Hagerman matrix test [102] can be achieved at an SII from 0.15-0.20, whereas to achieve the same score when tested against nonsense monosyllables requires significantly more favourable conditions, characterised by SII values between 0.40 and 0.45. The former can therefore be interpreted as being easier to understand in a given condition than the latter. Additionally, the AITFs for speech materials that are easier to understand also have steeper slopes, i.e. a small change in SII results in a large change in the intelligibility. This reinforces the conclusion that these speech materials possess greater redundancy, such that listening either results in understanding everything, or nothing, over a relatively small range of SII [99]. As referenced in Section 2.3, this invites the use of matrix tests as methods for determining private listening conditions, as a 50% correct score in this type of test corresponds to significantly lower levels of intelligibility in real-world passages of speech, such as those included in the Connected Speech Test [103].

By considering both the SATF and AITF for sentences, it can be seen that there is a rapid change in predicted intelligibility with small changes in the masking signal level. This relation is also observed in published psychometric functions between SNR and sentence intelligibility. This has significant implications for the design of speech privacy control systems, as it implies that the masking signal level must be specified precisely in order to ensure that the desired level of (un)intelligibility is achieved in the dark zone of a designed system. Figure 3.5 shows that one conceivable way to establish this link would be to form a mapping between the quantitative intelligibility and the qualitative privacy delivered by a system, but this is likely to be highly dependent on the context and expectations of listeners, and the details of the particular listening test that yielded the results. Instead, the approach taken by standards committees

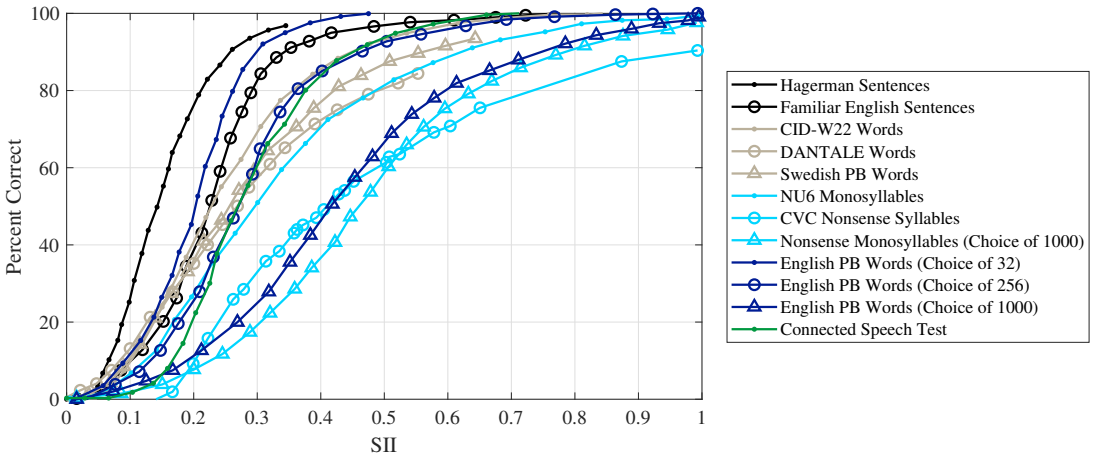


FIGURE 3.7: Audibility-Intelligibility Transfer Functions for the SII metric and a range of speech intelligibility tests, grouped by the type of speech material. Black lines indicate sentence-length speech tokens [102, 140]; grey lines are used for tests of single word intelligibility [118, 138]; light blue lines indicate monosyllable tests [118, 139, 140] ; dark blue lines indicate tests that use a restricted set of phonetically balanced (PB) words [140], and the green line shows the Connected Speech Test [103]

and researchers into open plan office privacy has been to form direct links between the output of intelligibility metrics and the results from subjective experiments into privacy, thus avoiding the additional complexity of including speech intelligibility test results. The following section describes the processes and standards that have been developed to convert between intelligibility metrics and levels of privacy.

3.3 Conversions Between Intelligibility Metrics and Privacy

There is a clear correspondence between intelligibility and privacy in sound zones; that is, as the proportion of words that are understandable by an eavesdropper increases, privacy decreases for the target listener. However, both are dependent upon the syntax and vocabulary in the spoken message and the amount of listening effort the eavesdropper uses. The link between intelligibility and privacy was observed by Fletcher and Steinberg in 1929 - in the following quotation from "Articulation Testing Methods" [106], the "discrete sentence intelligibility" refers to the ability of listeners to comprehend the substance of sentences, or respond intelligently to spoken questions.

> "It will be seen that for changes in distortion, the changes in the discrete sentence intelligibility will be small for systems having syllable articulations greater than 30 per cent, but very large for systems having syllable articulations below 20 per cent. It is for systems in this latter class that these test sentences are useful. A case in point is the measurement of the degree of secrecy obtained in sound proofing telephone booths, or in dealing with cross-talk." [106]

Much of the research into understanding the links between intelligibility and privacy has been considered in the context of office design. In an early review of speech privacy in offices, and

in order to provide practical guidance to architects and acoustic consultants, Cavanaugh et al. proposed a single-number rating scheme for privacy based on readily available architectural data and simple acoustical measurements [86]. The method showed a strong correspondence between objective speech intelligibility and the impression of speech privacy. This was evidenced by the favourable comparison between the single-number ratings and observed reactions of occupants in 37 office environments, as well as through independent validation [141].

A further insight from the studies of Cavanaugh et al. [86] is that different situations demand different levels of privacy, and it is important to confirm the expectations of listeners in studies when asking them about the privacy of different acoustic conditions. In their experiments that led to the generation of the rating system, listeners could adjust the level of interfering speech from another room until, firstly, it would interfere with day-to-day work, and secondly, the privacy of company classified material would be compromised. The notion of "everyday/acceptable" and "confidential" privacy, and the corresponding values of speech intelligibility metrics have been carried forward into more recent research into speech privacy [142, 143] and standards concerning the assessment of speech privacy in different spaces [101, 109, 144].

Two of these standards, ASTM E2638 [144] and ASTM E1130 [109] refer to closed and open plan offices respectively, and the different methods used in each standard reflects the usual requirements of each type of space. In open plan offices, methods aimed at reducing speech intelligibility are usually implemented to minimise distraction between neighbouring workspaces, whereas in closed offices, confidentiality may reasonably be expected between adjacent rooms. Accordingly, the recommendations in the corresponding standards use two different indices to describe privacy. In open plan offices, the Privacy Index (PI) is used - this is defined directly from the AI as

$$\mathrm{PI} = 1 - \mathrm{AI}. \tag{3.2}$$

In closed offices, privacy levels are described using the Speech Privacy Class (SPC), a measure that is based on the transmission loss through the boundaries of the space under test, $T(f)$, and the spectral level of the background noise in the receiving environment, $L_N(f)$. The frequency index f refers to sixteen ⅓-octave bands, such that the SPC can be calculated as

$$\mathrm{SPC} = \frac{1}{16} \sum_{f=160\,\mathrm{Hz}}^{5\,\mathrm{kHz}} [T(f) + L_N(f)]. \tag{3.3}$$

Despite differences in the frequency weightings between PI and SPC, Müller-Trapet and Gover present an approximate relationship between the two indices, which has been validated across the expected operating range of both metrics using 1800 datasets collected in North American offices [145]. The relationships are calculated as [145]

$$\mathrm{PI(SPC)} \approx 50 + 50\ \mathrm{erf}\left(\frac{6.65}{100}\ \mathrm{SPC} - 2.8\right), \tag{3.4}$$

and,

$$\mathrm{SPC(PI)} \approx \frac{100}{6.65}\ \mathrm{erf}^{-1}\left(\frac{\mathrm{PI}}{50} - 1\right) + 42.2, \tag{3.5}$$

where erf and erf^{-1} refer to the standard error function and its inverse.

A similar exercise to provide a relationship between the AI and SII was carried out by Bradley [142], in order that the privacy recommendations that relied on the old AI standard could be updated for use with the newer SII metric, provided that the acoustical conditions under test were appropriate for both standards. The empirical relationship was found to be well-represented by a fourth-order polynomial for AI values less than 0.5, which is given as [142]

$$\mathrm{SII} = 0.019 + 1.94\mathrm{AI} - 5.26\mathrm{AI}^2 + 11.73\mathrm{AI}^3 - 9.25\mathrm{AI}^4. \tag{3.6}$$

Equations 3.2 - 3.6 allow the recommendations made in each of the aforementioned standards and articles to be gathered and numerically compared. These are collated in Table 3.1.

AI	SII	SPC	Privacy / Intelligibility Description	Context	Reference
$\leq \mathbf{0.05}$	≤ 0.10	≥ 60	Confidential privacy	Open Plan	[109]
$\mathbf{0.05 - 0.20}$	$0.10 - 0.28$	$51 - 60$	Normal privacy	Open Plan	[109]
$> \mathbf{0.20}$	> 0.28	$48 - 51$	Speech becomes more readily understood	Open Plan	[109]
$> \mathbf{0.30}$	> 0.37	< 48	Unacceptable privacy	Open Plan	[109]
$> \mathbf{0.40}$	> 0.46	< 45	Essentially no privacy	Open Plan	[109]
$\mathbf{0.10 - 0.20}$	$0.17 - 0.28$	$51 - 56$	Speech is noticeable but not understandable	Open Plan	[146]
$< \mathbf{0.15}$	< 0.23	> 53	Acceptable privacy; freedom from intrusion	Open Plan	[88]
$> \mathbf{0.05}$	> 0.10	< 60	Most critical 10% of subjects feel a lack of privacy	Closed Room	[86]
$\approx \mathbf{0.10}$	≈ 0.17	≈ 51	Everyday privacy requirements satisfied	Closed Room	[86]
< 0.01	0.03	**70**	Minimal Speech Privacy - One or two words will be intelligible at most once each 3 minutes, and speech sounds will frequently be audible (at most once each 0.6 minutes)	Closed Room	[144]
< 0.01	0.02	**75**	Standard Speech Privacy - One or two words will be occasionally intelligible (at most once each 18 minutes) and frequently audible (at most once each 2 minutes)	Closed Room	[144]
< 0.01	0.02	**80**	Standard Speech Security - One or two words will very rarely be intelligible (at most once each 2.3 hours) and occasionally audible (at most once each 12.5 minutes)	Closed Room	[144]
< 0.01	0.02	**85**	High Speech Security - Speech essentially unintelligible (at most once each 16 hours) and very rarely audible (at most once each 1.5 hours)	Closed Room	[144]
< 0.01	0.02	**90**	Very High Speech Security - Speech not intelligible and very rarely audible (at most once each 11 hours)	Closed Room	[144]

TABLE 3.1: Values in literature pertaining to the instrumental appraisal of privacy. Original values from the referenced standards or publications are printed in **bold**, others have been calculated using Equations 3.2 - 3.6.

The entries from ASTM E2638 [144] in Table 3.1 give quantitative, as well as qualitative descriptions of the privacy obtained in each of the described scenarios, e.g. "Standard Speech Security - One or two words will very rarely be intelligible (at most once each 2.3 hours)". These descriptions lock the subjective appraisal of speech privacy to measurable quantities. Limited information is available with regard to numerical intelligibility scores at different levels of privacy in open environments, but it is expected that privacy descriptors in these spaces relate to higher levels of intelligibility, due to the lower expectations of privacy in these open spaces, compared to between closed rooms. Bradley and Gover [99] report that at AI values of 0.15 the median speech intelligibility, using low-predictability Harvard sentences [100], was 88%, but this score was expected to decrease with less-well enunciated, everyday speech. At the same AI and SII values, speech privacy was subjectively rated by listeners as just below "acceptable", on an ordinal scale of speech privacy where 1 = "None", 2 = "A little", 3 = "Acceptable", 4 = "Moderately Good" and 5 = " Confidential". The study's authors describe SII = 0.20 as the "corner" value on the curve between intelligibility (or privacy) and SII; for lower SII values, the intelligibility scores decrease rapidly and thus correspond to significant improvements to privacy. On the other hand, solutions that provide SII values greater than 0.2 essentially offer no additional privacy compared to doing nothing [99].

3.4 Summary

The objective of the privacy control considered in this thesis is to ensure that speech intended for the target listener in the bright zone is not overheard by any listeners in the dark zone. This is achieved through a combination of focussing speech towards the target listener, and radiating a masking signal into the dark zone. These signals can be input into intelligibility metrics to inform how changes to the design of the loudspeaker array and masking signal affect how speech is perceived in each zone. The present chapter has discussed the construction, advantages and disadvantages of a number of such metrics, but the SII is the most appropriate for the purposes of practical private audio system design.

The SII is based on extensive research into the factors affecting speech intelligibility, spanning over seventy years, and has been widely adopted in both research and clinical settings. This has led to a wide range of AITFs being made available for many different intelligibility tests, as shown in Figure 3.7, so the conversion from SII to percent-correct scores can be carried out with confidence and specificity to the type of intelligibility required. Furthermore, as the SII is directly based on the frequency-weighted SNR, the conversion between broadband SNR and SII described by the SATF has a narrower confidence interval than competing metrics with more complex processing algorithms. This therefore reduces the uncertainty of intelligibility predictions. The most significant benefit of SII, however, is that links have been established between SII and both qualitative and quantitative descriptions of privacy, as discussed in Section 3.3.

The use of SII to standardise privacy ratings for both open and closed offices gives confidence that the metric can also be used to evaluate the privacy in sound zones. It is clear from the comparisons between the indices in Table 3.1 that there is a significant perceptual difference between the expected levels of privacy in open plan and closed spaces. "Acceptable" or "normal" levels of privacy are associated with significantly lower SII values in closed rooms compared to

open plan areas, and this suggests that it is potentially unnecessary to expect a speech privacy control system to provide similar levels of intelligibility reduction to that provided by a closed room. The implications of this will be discussed in Chapter 7, where an SII limit is specified for the dark zone.

A potential limitation of using the SII metric is that the algorithm is not designed to operate on speech that is masked by strongly fluctuating noise. However, for the purpose of initial investigations into the factors affecting speech privacy in sound zones, it is beneficial for the initial experiments to be restricted to simple, stationary random noise maskers, as the broader effects of masker level and frequency spectrum can then be explored in isolation. The additional complexities of evaluating the effectiveness of fluctuating masking signals are discussed in Chapter 9, both in terms of their ability to mask speech, and also the perceptual effects of introducing this type of masking noise into an environment.

A key engineering trade-off that arises from the selection of a dark zone SII target is that, for a personal audio system design with a given level of acoustic contrast performance, demanding a higher level of privacy will also require an increase in the level of the masker. The perceptual effects of additional masking can be quantified, for both stationary and non-stationary maskers, using subjective metrics, and this additional consideration is essential for the production of practical sound zoning systems. In the present chapter, intelligibility metrics have been used to consider the effects of additional masking on the privacy of the target listener, without regard for the experience of other listeners. This single-minded approach risks developing systems that are unacceptable for use in public spaces. Therefore, it is necessary to also include perceptual evaluation in the design of speech privacy control systems, and this will thus be discussed in the following chapter.

Chapter 4

Subjective Metrics for Perceptual Evaluation

The discussion presented in Section 2.2 shows that it is desirable to integrate an understanding of the human perception of sound into the process of designing a personal audio system, alongside other engineering objectives. In order to justify the installation of a speech privacy control system in a space, it must not only execute its primary function of controlling the intelligibility of speech in each zone, but do so without adverse side-effects. Due to limitations in the directivity of practical loudspeaker arrays, speech intended for the target listener can remain intelligible to other nearby listeners, necessitating the introduction of masking noise into the dark zone, as depicted in the block diagram in Figure 1.2. If this additional noise is too loud or unpleasant, privacy and practicality become irreconcilable. In Chapter 7, a process is described for the design of an optimised masking signal, which minimises the likelihood of adverse listener reactions, whilst maintaining privacy between the listening zones. A necessary component of this process is a means to instrumentally evaluate how a particular masking signal may be perceived by listeners.

In the previous chapter, metrics were used to predict the intelligibility of speech using the physical properties of the signals received in each zone. The metrics described in this chapter operate on similar principles, in this case with the aim of providing a mapping between measurable, objective parameters of a signal, the *stimulus*, and the expected subjective response, the *sensation*, from a population [79]. Subjective metrics can be loosely ranked or categorised based on the complexity of the relation between stimulus and sensation. For example, the subjective experience of loudness is primarily related to intensity, with further spectral and temporal effects [147]. Complex sensations, such as the annoyance experienced due to noise, can be modelled by combining simpler effects, using the results from subjective experimentation to determine the relative influence of each factor.

Following previous work by Widmann, [148], in this chapter the complex psychoacoustic sensation of noise annoyance is broken down into four elementary attributes; loudness, sharpness, roughness and fluctuation strength. The following four sections provide the mathematical definitions for metrics that predict these perceptual attributes, based on measurable properties

of input signals. This will provide information on how a masking signal could be adjusted to improve its acceptability in a personal audio context. The implications of designing a masking signal with a low level of each attribute will be discussed, in order to highlight the trade-offs between the various requirements of private personal audio systems.

4.1 Loudness

The sensation of loudness is correlated with intensity, but varies with the spectral content and temporal profile of a signal. Much has been written about the correlation between intensity and loudness, indeed some of the earliest psychoacoustic experiments of the 19^{th} century sought to quantify the various physical factors that contribute to the experience of loudness [147]. Formalised experiments led to the now-standard equal-loudness contours for continuous tones, as shown in Figure 4.1 [149, 150]. A range of loudness models have been developed, each with subtly different rationale and intended application. Some models seek to emulate loudness by accurately reproducing the physical processes in the auditory system [151]. This allows effects such as hearing loss to be accounted for - an important step required in the fitting and calibration of hearing aids. Some of these models have been adapted to capture the effects of time-varying loudness [152], or re-formulated to focus on computational efficiency [153]. Often, models are left with free parameters that can be adjusted to fit statistical results from listening tests [154].

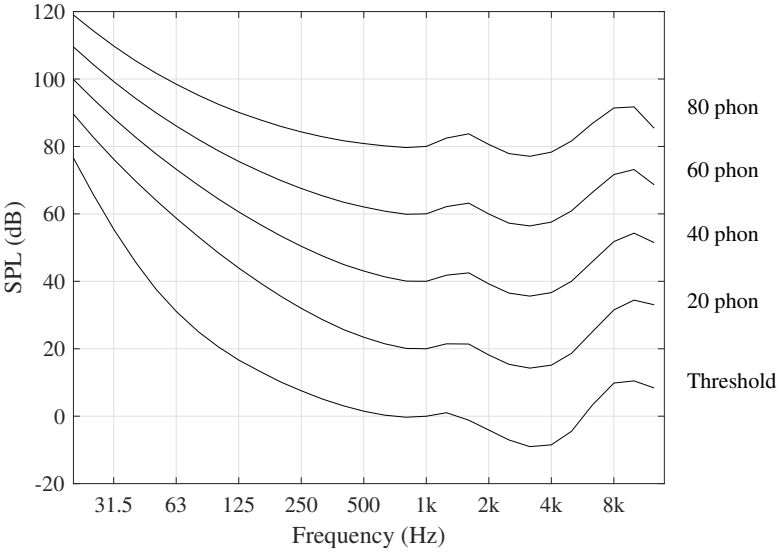


FIGURE 4.1: Equal-loudness contours according to ISO 226.

The equal-loudness contours shown in Figure 4.1 can be applied to the long-term spectrum of a stationary noise to determine a measure of the loudness. Two interchangeable units are commonly used for loudness measurement. Figure 4.1 describes *loudness levels*, L_N, in phons; a 1 kHz tone at the mean threshold of hearing is defined as 0 phons, and the equal-loudness contours are defined at 20 phon intervals, corresponding to 20 dB increases in the Sound Pressure Level (SPL) at 1 kHz. The definition of the equal-loudness contours makes the loudness level in phons a convenient measure to describe sounds that are equally as loud as each other. To compare the difference in the subjective loudness of sounds, the phon scale can be converted to a second unit, the sone, N. One sone is defined as 40 phons at any frequency, and a perceived

doubling of the loudness is represented by a doubling in the number of sones - for pure tones, this approximately corresponds to an increase of 10 phons [79]. For narrowband signals, the relationships between sones and phons are defined in ISO 532-1:2017 [155] as

$$L_N = 40 + 33.22 \log_{10}(N), \text{and} \tag{4.1}$$

$$N = 2^{0.1(L_N - 40)}, \tag{4.2}$$

when the signal is greater than 1 sone in loudness, or has a loudness level greater than 40 phons. To account for the change in human perception of quiet sounds, below a loudness of 1 sone, or a loudness level of 40 phons, the relationships are

$$L_N = 40(N + 0.0005)^{0.35}, \text{and} \tag{4.3}$$

$$N = (L_N/40)^{2.86} - 0.005. \tag{4.4}$$

These relationships can be used to determine the loudness of stationary broadband sounds, but in order to capture the perceived loudness of time-varying sounds, an interim measure known as the instantaneous loudness must be calculated [156]. Sampled at 500 Hz intervals, the instantaneous loudness produces a loudness envelope signal that can be indexed to determine the percentile loudness, N_x, which is the instantaneous loudness exceeded for x percent of the signal duration. For non-stationary sound, the fifth percentile loudness, N_5, is reported to correlate more strongly with the perceived loudness of the source than the time-average of the instantaneous loudness [79]. This measure has been internationally standardised [155], though other methods for aggregating short term loudness data into a single-number descriptor are also mentioned in the literature [157–159].

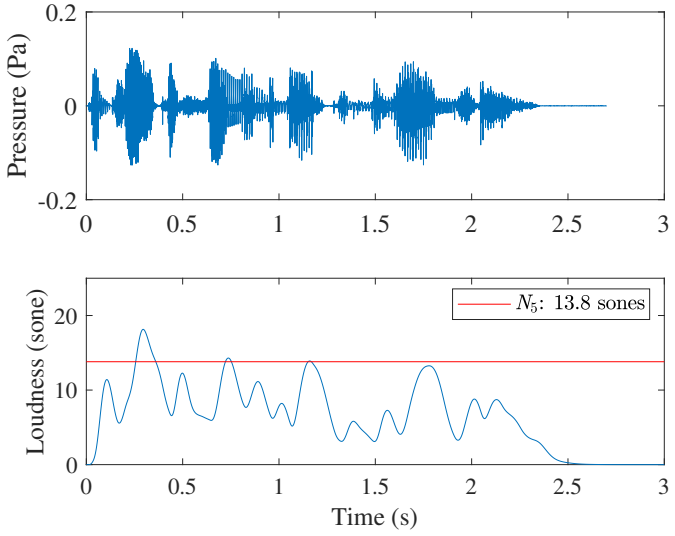


FIGURE 4.2: Male speech signal, (upper panel) and the associated instantaneous loudness in sones (lower panel). The fifth percentile loudness N_5 is indicative of the overall perceived loudness of the signal.

To illustrate the calculation of the N_5 index, Figure 4.2 shows the instantaneous loudness of a spoken sentence, normalised to have an average SPL of 60 dB, calculated using a MATLAB implementation of Zwicker's instantaneous loudness model [160]. For a total of 5% of the signal duration, the loudness exceeds a value of 13.8 sones, and this contribution comes almost completely from a single spoken word. Similar to how variations in prosody can have significant implications for the overall intelligibility of a sentence, the perceived loudness of a temporally varying signal can be dominated by infrequent peaks in the loudness. As can be seen from the decay of loudness with time at the end of the speech segment, the sensation of loudness continues beyond the physical stimulus. This effect is demonstrated more clearly in Figure 4.3, which shows the instantaneous loudness of a 50 ms burst of white noise. When considering the perceptual evaluation of masking signals, the ability to model this non-simultaneous masking effect is important as a masker can remain effective outside of its actual temporal envelope.

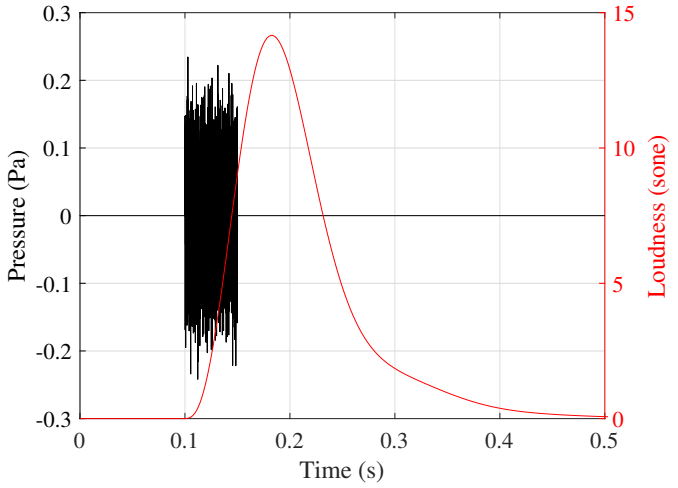


FIGURE 4.3: Instantaneous loudness (red, sone) of a 50 ms white noise burst (black, Pascal).

A measure of the overall loudness of a signal can also be aggregated from the loudness measured in separate frequency bands. Perceptual studies often divide the audible frequency range into critical bands, as opposed to octave- or ⅓-octave bands, as this better reflects the frequency selectivity of the ear [161, 162]. The critical bandwidth is defined as the range of frequencies present within a broadband noise that contribute to the masking of a tone. Critical bands are approximately 100 Hz wide below 500 Hz and increase in width above 500 Hz, such that the bandwidth is approximately 20% of the centre frequency. This continuous definition can be more conveniently represented using a standard set of critical bands, known as the Bark scale [163]. The equal perceptual contribution to loudness from each critical band allows the specific loudness from each band, N' in sones/Bark, to be integrated into a single loudness rating in sones. Two plots showing the contributions of specific loudness to overall loudness, reproduced using data from Zwicker and Fastl [79] are shown in Figure 4.4. Despite specific loudness readings in the lower panel exceeding those in the upper panel, the wider bandwidth of the broadband noise represented in the upper panel gives this signal a greater perceived loudness than the narrowband sound. The implication of this for the design of masking signals is that all frequency regions contribute to the impression of loudness. This highlights the importance of matching the spectra of the masking noise to that of the speech, as any additional energy outside of the frequency range

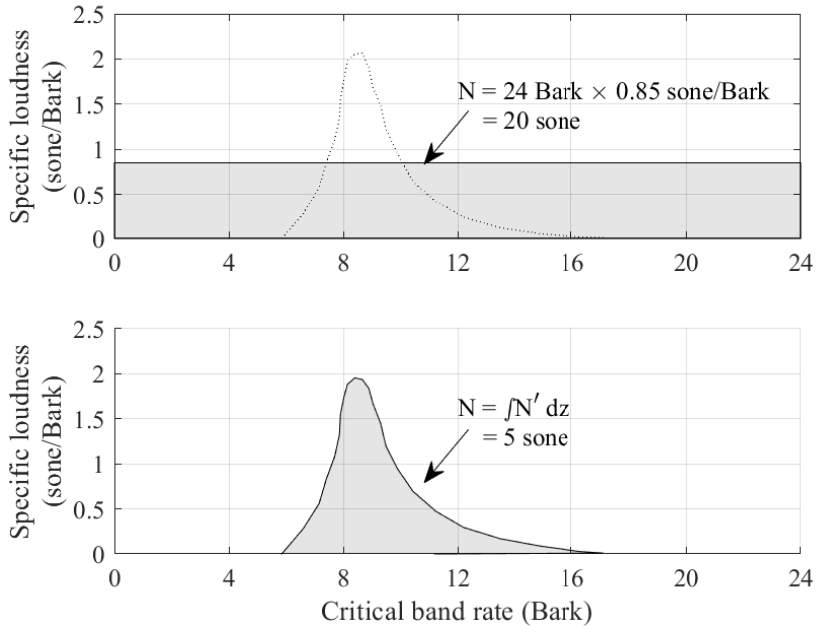


FIGURE 4.4: Specific loudness, N', as a function of critical band rate for uniform exciting noise (upper panel, 20 sone) and 1 kHz critical band wide noise (lower panel, 5 sone). The shaded areas indicate the areas contributing to total loudness in each panel, and the dotted curve in the upper panel matches the area of total loudness of the narrowband noise in the lower panel for comparison.

of speech will increase the perceived loudness of the masker without necessarily contributing to a reduction in speech intelligibility.

Instantaneous loudness can be combined with specific loudness to give an indication of how a signal varies in terms of both time and frequency, known as the specific instantaneous loudness. This provides similar information to a spectrogram, except the frequency divisions and time response are more closely aligned with how sounds are processed by the human auditory system. Figure 4.5 shows the specific instantaneous loudness of a single spoken word. The expected distribution of high and low frequency excitation with, respectively, the consonant and vowel sounds in the word can be seen. This visualisation may be used to demonstrate that to successfully reduce the intelligibility of a spoken word, the entire excitation pattern produced by the word must be masked. As discussed in Section 3.1.3, if a fluctuating masking signal is used, it is possible for "glimpses" of the underlying word to be audible when the brief excitation caused by a particular speech sound aligns with a spectrotemporal "dip" in the masker. The main strength of the specific instantaneous loudness, however, is that it provides spectral and temporal information that can be processed to calculate metrics associated with other perceptual sensations, such as fluctuation strength and sharpness.

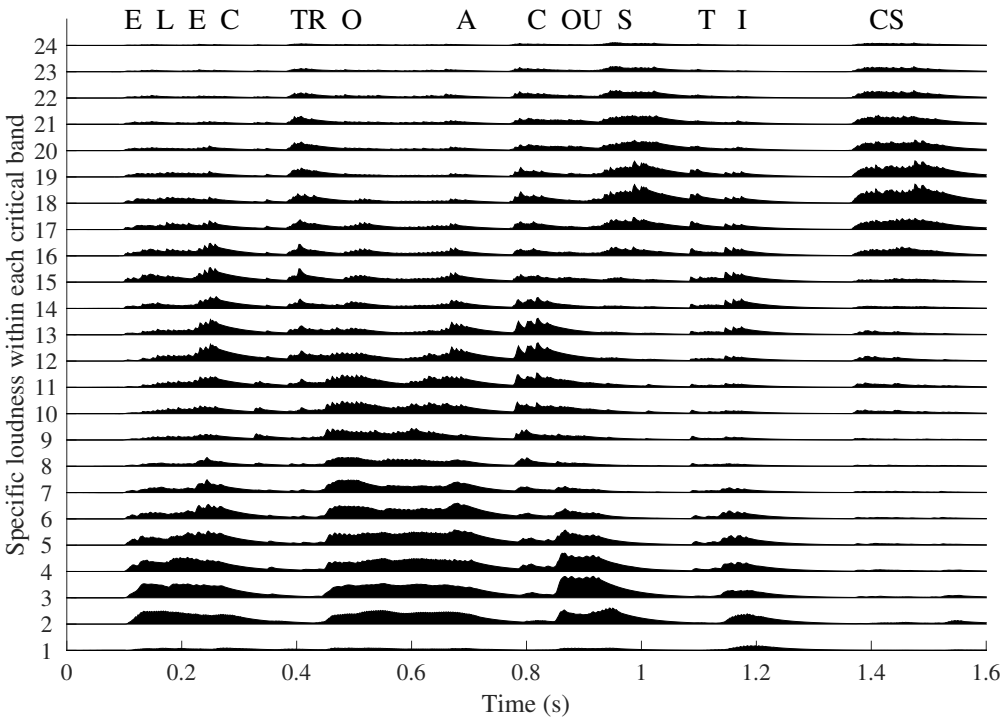


FIGURE 4.5: Specific instantaneous loudness of the spoken word "Electroacoustics".

4.2 Sharpness

Sharpness is an unpleasant sensation that is dependent on both the intensity of a signal and its frequency spectrum; signals with a higher proportion of energy above approximately 3 kHz exhibit a greater degree of sharpness. In order to account for the non-flat frequency response of the ear, the proportion of energy in each frequency band is found by processing the specific loudness, N' of the signal. From this, the corresponding sharpness, S can be calculated as [79]

$$S = 0.11 \frac{\int_0^{24\text{Bark}} N' g(z) z \, \mathrm{d}z}{\int_0^{24\text{Bark}} N' \, \mathrm{d}z} \tag{4.5}$$

where the Bark-band weighting $g(z)$, displayed graphically in Figure 4.6, is given by

$$g(z) = \begin{cases} 1, & z < 14 \\ 0.00012z^4 - 0.0056z^3 + 0.1z^2 - 0.81z + 3.51, & z > 14 \end{cases}. \tag{4.6}$$

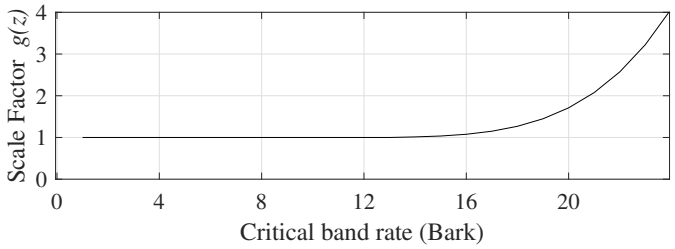


FIGURE 4.6: Sharpness weighting function $g(z)$ (Equation 4.6) as a function of critical band rate.

The factor 0.11 in Equation 4.5 calibrates the sharpness metric to a fixed point, 1 *acum*, which is the sharpness of a critical band limited noise centred at 1 kHz, and reproduced at 60 dB SPL. To demonstrate the types of sounds that elicit a high level of perceived sharpness, the instantaneous sharpness of a spoken sentence has been calculated using a MATLAB implementation of the sharpness metric [164], and is presented in Figure 4.7. The instantaneous sharpness is formed by inputting the specific instantaneous loudness into Equation 4.5. During fricative sounds such as "sh","s" and "f", critical bands above $z = 14$ are excited, as demonstrated for the "s" phoneme in Figure 4.5. These bands are emphasised by the sharpness weighting function $g(z)$, so the predicted instantaneous sharpness of these sounds is increased. Whilst informative, the instantaneous sharpness value is unwieldy for large scale perceptual evaluations, and a single-number rating representative of the overall sharpness is required, similarly to how N_5 represents the overall loudness from the instantaneous loudness. A comparison of three methods for aggregating the instantaneous sharpness into a single-number rating [165] found that the arithmetic mean of instantaneous sharpness values was better correlated with perceptual ratings than the geometric mean or the maximum sharpness.

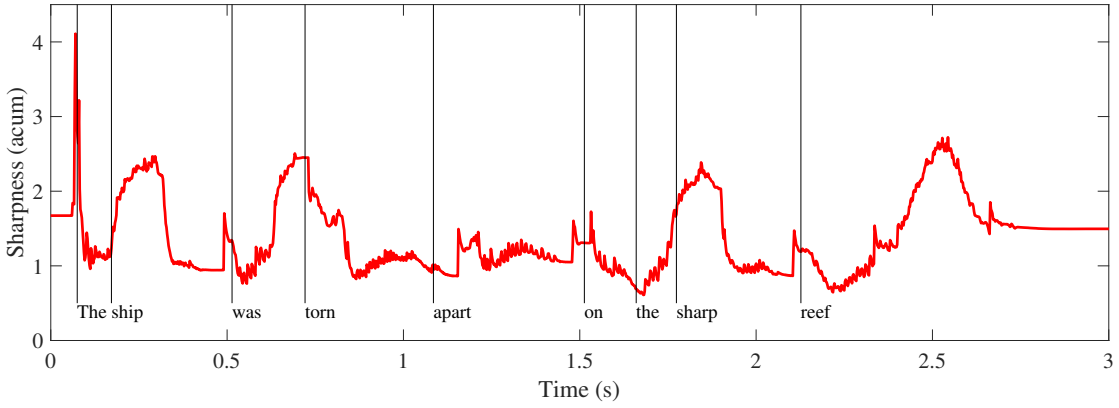


FIGURE 4.7: Instantaneous sharpness of a sentence from the Harvard Sentence Corpus [100, 166]. Vertical bars indicate the beginning of each word.

4.3 Roughness

Roughness is usually regarded as an unpleasant sensation, and thus is included in the formulation of the psychoacoustic annoyance metric [79]. A number of methods for the prediction of psychoacoustical roughness have been proposed [79, 167–169], with various degrees of algorithmic complexity, ranges of application and published validation with subjective testing. The

various models agree that the sensation of roughness is evoked by sounds that modulate, in amplitude or frequency, at a rate of between a few tens of Hertz and the low hundreds of Hertz [79]. The model by Aures [167], with optimisations by Daniel and Weber [168] is selected for use in this thesis due to its flexibility at characterising the roughness of both broadband and tonal signals. The model assesses the modulation depth and frequency of critical-band filtered channels to determine the specific roughness, similar to the specific loudness described in Section 4.1. Additionally, the model calculates interaction effects between adjacent frequency channels, to account for the finding that the overall roughness sensation can be reduced if the excitation in one frequency region masks another [167]. An implementation of this model has been made available as part of the PsySound3 project [170].

The perception of roughness in modulated signals exists between two other sensations, in terms of the frequency of the modulation. Low frequency modulation, below around 15 Hz, is discernible as a smoothly varying change in either amplitude or frequency. At higher modulation frequencies, above around 100 Hz, the modulation begins to manifest itself as additional tones, which can be perceived separately to the carrier signal. These two frequency components, f_{mod} and f_c, can be seen in the waveform of a sinusoidally amplitude modulated tone,

$$p(t) = p_0[1 + m\cos(2\pi f_{mod}t)\cos(2\pi f_c t)], \tag{4.7}$$

where m is the modulation index.

Subjective experiments find that with $m = 1$, the maximum roughness level of a tone with $f_c < 1$ kHz is experienced at a modulation rate of 70 Hz [79]. Lower frequency tones exhibit maximum roughness at lower modulation frequencies, as shown in Figure 4.8, which shows the roughness of several 100% amplitude modulated tones. In order to provide a fixed point on the scale of roughness, the perceived roughness of a 1 kHz critical band-limited noise, at 60 dB SPL, fully amplitude modulated at 70 Hz is defined as 1 asper. This value is approximately maximal for amplitude modulated narrowband noises or tones, though values as large as 5-6 asper are produced by amplitude modulated broadband tones or rapid, wide frequency modulation [79].

As described in Section 3.4, the masking signals considered in the majority of this thesis will consist of stationary random noise, and thus do not include explicit amplitude modulation at the frequency ranges that are associated with the perception of roughness. Nevertheless, it is important to consider the roughness of these signals as there is no requirement for the modulation of a signal to be periodic for roughness to be perceived [79]. The random envelope modulation of stationary random noise inherently produces a small sensation of roughness, and this is most significantly affected by the bandwidth of the noise [168]. In a given random signal, the number of envelope maxima per second, n, is proportional to the bandwidth, B; on average, $n = 0.64B$ [167]. By interpreting n as a modulation frequency, and assuming peak roughness coincides with a modulation frequency around 70 Hz, as shown in Figure 4.8, unmodulated noise with a bandwidth of 70 Hz/0.64 $\approx$ 100 Hz will provide the greatest sensation of roughness.

Speech-shaped noise has a significantly wider bandwidth than this limit, and when reproduced at 70 dB, has an estimated roughness of 0.06 asper. This is close to the threshold of roughness perception for amplitude modulated tones [168]. The random nature of the signal means that

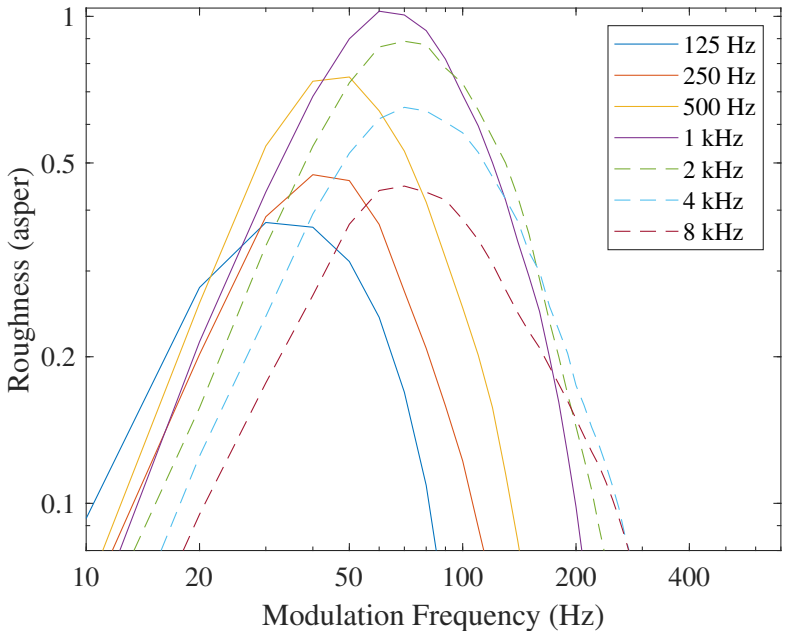


FIGURE 4.8: Predicted roughness of 100% sinusoidally amplitude modulated pure tones at centre frequencies from 125 Hz to 8 kHz.

the roughness value is also subject to variation. To illustrate this, Figure 4.9 shows a histogram of roughness evaluations estimated from 523 short segments of speech-shaped noise.

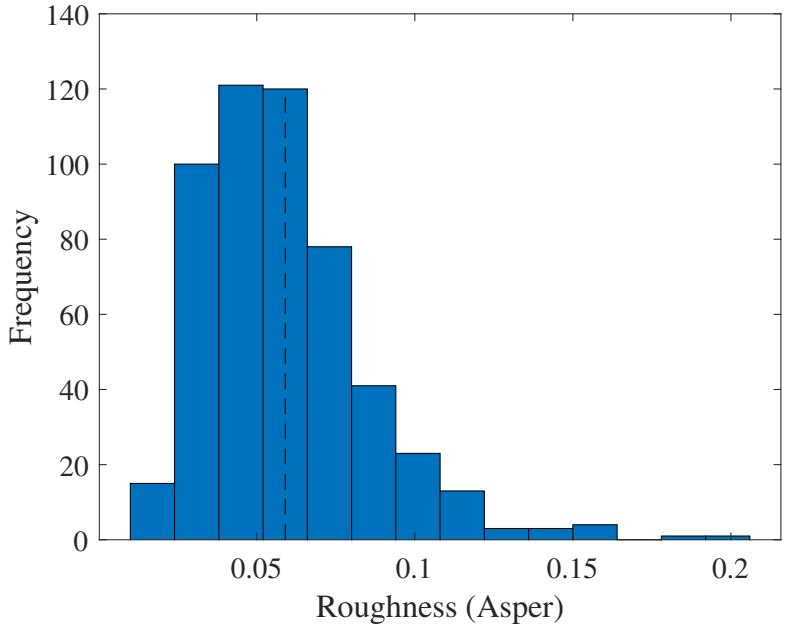


FIGURE 4.9: Histogram of roughness scores evaluated for 523 short segments of noise matching the spectrum of the VCTK speech corpus [22]. The dashed line indicates the mean of the distribution.

As roughness is considered to be an undesirable feature of environmental noise, is it likely to be advantageous to reduce the roughness of the masking signals output by the personal audio system. However, unlike with sharpness or loudness, where simple linear filtering of the signal can have a profound effect on the perception of these sensations, the roughness of a signal is affected by both the spectral and temporal properties of a signal. This makes it difficult to reduce

the roughness of a signal without also affecting other perceptually important parameters. Although stationary, random noise maskers have an inherently low level of roughness, as evidenced in Figure 4.9, there are further perceptual considerations to be made regarding the perceived appropriateness or naturalness of a masking signal, with reference to the ambient background noise in the reproduction environment. As this contextual effect cannot be calculated based on the properties of the signal alone, it cannot be included in a purely instrumental analysis of psychoacoustic annoyance.

The variations in amplitude associated with multi-talker babble noise or other speech-like masking signals such as the ICRA series of modulated speech-shaped noises [171] occur over much longer timescales, and relate to the perception of level fluctuation, rather than roughness. The mechanics and modelling of this sensation are described in the following section.

4.4 Fluctuation Strength

Another acoustical stimulus that can result in an unpleasant sensation is related to the strength of the amplitude fluctuations in the signal [79]. The fluctuation strength of a signal is distinguished from its roughness by the frequency of the modulations. While a peak in the roughness response is reported at a modulation frequency around 70 Hz, maximum fluctuation strength is experienced with an amplitude modulation of 4 Hz. Short-term memory effects at these timescales mean that the actual modulation depth of a signal and the perceived depth of the modulation, ΔL are not identical [79]. This is illustrated qualitatively in Figure 4.10. The difference between the true and perceived modulation depth increases as the modulation frequency increases, to a point at around 20 Hz where the modulation begins to be perceived as roughness. A 1 kHz critical band limited noise at 60 dB, fully amplitude modulated at 4 Hz, is defined as having a fluctuation strength of 1 vacil. The noticeability of sirens and other warning signals can be attributed to their high level of fluctuation strength as these signals fluctuate in both amplitude and frequency.

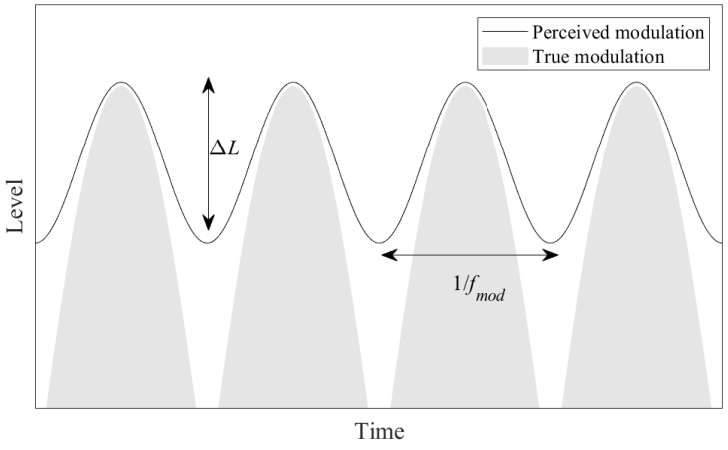


FIGURE 4.10: Qualitative diagram of perceived modulation depth ΔL of a masking signal, sinusoidally amplitude modulated at frequency f_{mod}.

No standard implementation of the fluctuation strength metric was available from other sources, so a new implementation was written [164], based on the algorithm described by Zwicker and Fastl [79]. The fluctuation strength model depends on an assessment of the perceived modulation depth and modulation frequency. This is achieved by first passing the signal through the loudness algorithm [160], to produce a set of instantaneous loudness envelopes for 240 frequency bands, each with a width of one tenth of a critical band. As the loudness algorithm includes non-simultaneous masking effects, these loudness envelopes approximate the black trace in Figure 4.10. These 240 signals are aggregated by summing them into 24 critical bands. The peaks and troughs of each envelope are found, and the level difference ΔL between each peak and the deepest adjacent trough are averaged across signals then summed across Bark bands to produce the fluctuation strength, which is scaled by a constant factor to calibrate the algorithm to an output of 1 vacil with the prescribed input signal.

This implementation of the fluctuation strength algorithm used in this work can be compared against target values from subjective tests results presented by Zwicker and Fastl [79], for fluctuating broadband noise. Figures 4.11 to 4.13 show that the implementation of the fluctuation strength algorithm captures the dependence of the metric on modulation depth, modulation frequency and the overall SPL of the signal, respectively.

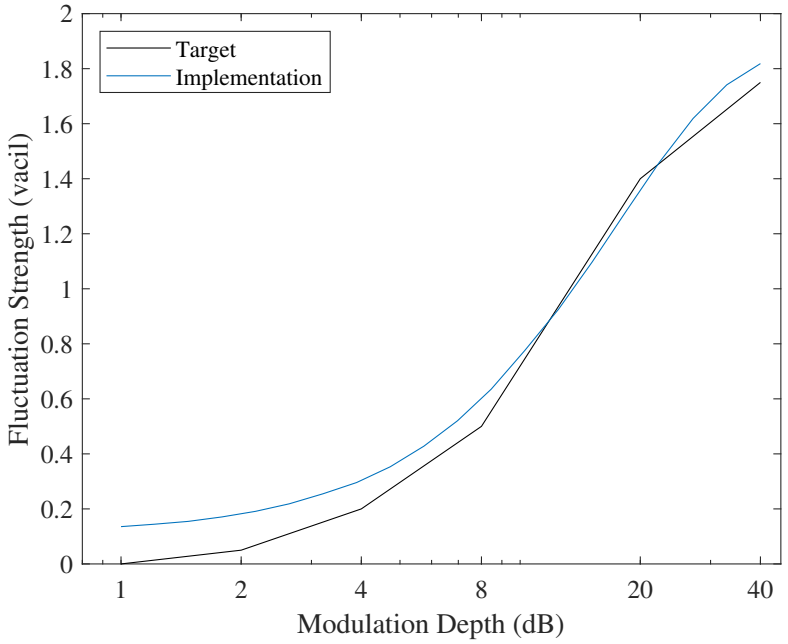


FIGURE 4.11: Variation in the fluctuation strength metric with modulation depth for 60 dB SPL amplitude modulated speech-shaped noise, measured as the ratio in decibels between the maximum and minimum of the signal envelope. Target levels from Ref. [79].

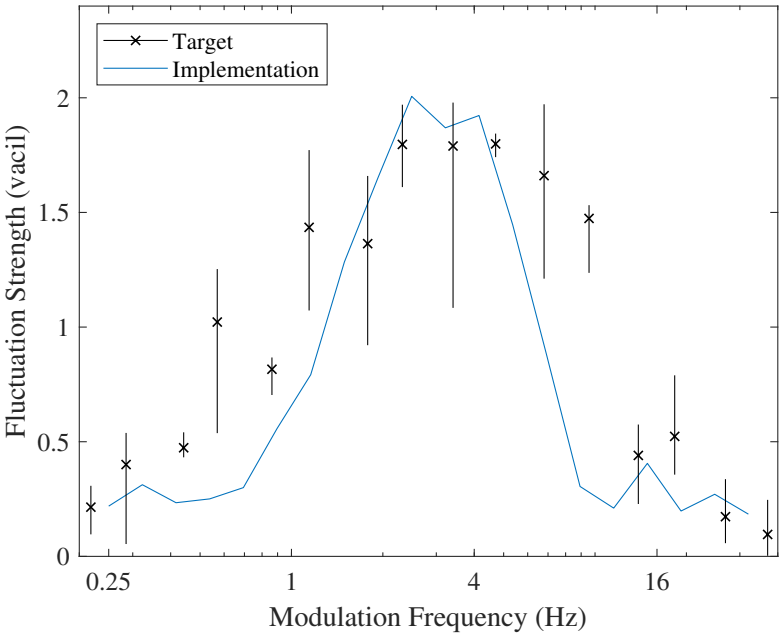


FIGURE 4.12: Variation in the fluctuation strength of 60 dB SPL amplitude modulated speech-shaped noise with modulation frequency, at a modulation depth of 40 dB. Target levels from Ref. [79].

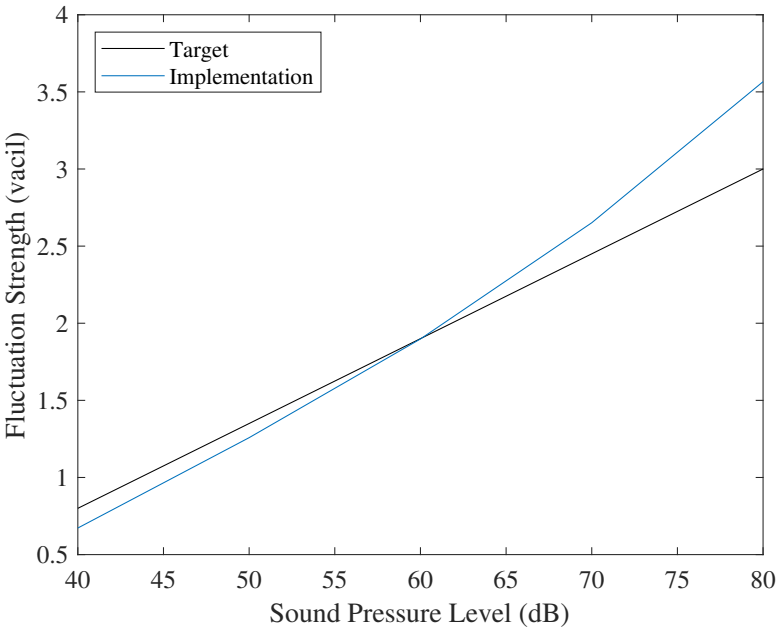


FIGURE 4.13: Variation in the fluctuation strength of 4 Hz sinusoidally amplitude modulated speech-shaped noise with signal level, at a modulation depth of 40 dB. Target levels from Ref. [79].

The perceived fluctuation of stationary broadband noise is negligible, but the effects of fluctuation may still be encountered when the practicalities of personal audio system design are considered. Systems that are not used continuously for the delivery of private speech risk becoming sources of noise pollution if the masking signal is left at a constant level. One approach could be to adjust the level of the masking signal to track changes in the output speech level, ensuring

that when the system is not outputting speech, it does not output noise either. If the time constant of these amplitude modulations is set to react quickly to changes in the speech level, due to natural fluctuations in prosody, or gaps between sentences, this could result in unpleasant fluctuation in the masking noise level. In this case, a metric such as fluctuation strength could be utilised to automatically limit the rate at which a system adapts to these changes. Another example of where the fluctuation strength metric could be used is to instrumentally distinguish the perceptual advantages and disadvantages of using stationary masking signals over those that mimic the temporal structure of speech, such as the ICRA series of speech-shaped noises [171] or that contain distinguishable speech samples, such as multi-talker babble. These types of signals are efficient maskers due to their ability to cause informational, as well as energetic masking, i.e. they can be reproduced at a lower level than steady state maskers for the same reduction in intelligibility [172]. Consequently, the fluctuation strength, along with the loudness, sharpness and roughness metrics described earlier in this chapter, provide simple ways to characterise the wide range of perceptual sensations that could be experienced by listeners situated in the dark zone of a private personal audio system, when different types of masking signal are used. Each of the four sensations described in this chapter may be considered separately, as they are distinguishable by the human auditory system [79], but for further simplicity, these metrics can also be aggregated together into a single metric, the psychoacoustic annoyance. The process for forming this metric is described in the following section.

4.5 Psychoacoustic Annoyance

Throughout this chapter, attention has been paid to increasingly high-level sensations, with progressively more complex relationships to the objective stimuli that cause them. The sensation of annoyance deviates slightly from this sequence, as the experience of annoyance due to noise is controlled by many factors, not all of them acoustical. Reactions such as fear of the (unseen) sound source or the elicitation of unpleasant memories all contribute to the sensation of annoyance, but are not a direct consequence of the physical parameters of the sound [173]. The contribution to the overall sensation of annoyance that is controlled by auditory parameters is termed the psychoacoustic annoyance [79, 148]. Typically, the annoyance of environmental noise is managed by controlling the sound level; European Union guidelines for the control of environmental noise cite the reduction of annoyance caused by noise as one motivation for recommending an upper bound on the time averaged A-weighted SPL during night-time hours [174]. The use of such crude measures to capture annoyance is arguably justified by the additional time, expense and computation required by more complex metrics [173]. However, for the purposes of research into novel speech privacy control systems, the additional detail and insight provided by an analysis of psychoacoustic annoyance is justifiable, particularly when simulations of the signals received in each listening zone can be rapidly generated. The four subjective metrics described in the previous four sections of this chapter can be combined to produce the psychoacoustic annoyance, PA, as

$$PA = N_5 \left(1 + \sqrt{w_S^2 + w_{FR}^2}\right), \tag{4.8}$$

where

$$w_S = (S - 1.75) \times 0.25\log(N_5 + 10) \tag{4.9}$$

for $S > 1.75$ acum, and

$$w_{FR} = 2.18/N_5^{0.4}(0.4F + 0.6R) \tag{4.10}$$

where N_5 is the instantaneous loudness exceeded for 5% of the signal duration, and Sharpness S, Fluctuation Strength F and Roughness R are measured in acum, vacil and asper respectively. For sharpness less than 1.75, the contribution to annoyance from w_S is zero.

The coefficients for the psychoacoustic annoyance metric were determined by Widmann by fitting data to listening test results [148]. In these tests, participants rated the annoyance of broadband and narrowband noise samples with and without amplitude modulation, at different absolute levels. Although synthetic signals were used to calibrate the metric, published results [79] show that when the metric is applied to a range of recordings of natural and technical sounds, a strong correlation is found with the subjectively measured annoyance. Despite giving a quantitative result, the context-dependence of annoyance means that the metric is best used to compare the annoyance of similar sounds, reproduced in a similar context [175], rather than making face-value judgements such as "This masking sound is more annoying than a lawnmower", for example. This limitation of the annoyance metric may prove significant when comparing how different masking signals are perceived in the presence of ambient noise, a situation described later in Chapter 9.

In this thesis, it will frequently be necessary to evaluate the psychoacoustic annoyance of the signals received in the listening zones produced by a personal audio system. These signals will consist of mixtures of speech and noise. Figure 4.14 shows how the various metrics described earlier in this chapter, and the overall SPL, vary when the level of a speech-shaped noise is changed relative to a fixed-level speech signal at 60 dB SPL. Firstly, and unsurprisingly, the upper left panel of Figure 4.14 shows that as noise levels approach, and then increase beyond the level of the speech, the perceived loudness of the combined signal increases dramatically. Compared against the loudness, the perceived sharpness increases more gradually as the noise level increases. This is because the sharpness metric is based on a frequency-weighted time-average over the specific instantaneous loudness, as described in Equation 4.5. At low noise levels, the combined signal is dominated by speech, and the sharpness of speech is concentrated into short time frames, as shown in Figure 4.7. When time-averaged, these contributions to the overall sharpness are lower than the sharpness of speech-shaped noise, which dominates the combined signal once noise levels increase. The sharpness levels for the signals tested here fall below the threshold of 1.75 acum required by Equation 4.8 for inclusion in the psychoacoustic annoyance metric, though bandwidth limitations and irregularities in the frequency response imposed by the sound zoning process may increase the overall sharpness of signals emitted by practical systems.

The middle row of panels in Figure 4.14 show that the roughness and fluctuation strength of the combined signal decrease significantly as the noise levels increase. The temporal modulation present in speech, on timescales corresponding to both the sensation of roughness and fluctuation, is smoothed by the broadband speech-shaped noise as it becomes more dominant in the combined signal. Some fluctuation and roughness remains perceivable in stationary broadband noise, as these sensations are caused by the small variations in level that are inherent in random signals.

As the overall signal level increases, so do the scale of these variations - this explains the slight increase in predicted roughness and fluctuation strength at very high noise levels.

The resulting psychoacoustic annoyance rating, in the lower left panel of Figure 4.14, reaches a minima at an SNR of 0 dB, i.e. when speech and noise are both reproduced at 60 dB. When speech at a constant level dominates the noise, the contribution to annoyance from loudness is constant but the fluctuation strength and roughness of the signal result in higher psychoacoustic annoyance. As the noise level increases, psychoacoustic annoyance values are dominated by the increase in loudness, as evidenced by the inclusion of loudness in every term of Equation 4.8, which describes the composition of the annoyance metric. This initial indication, that reducing the loudness of the masker is the most significant contributor to the the reduction of annoyance, will be investigated more fully in Chapter 6. In the previous chapter, a reduction in the level of the masker was also linked to an undesirable increase in the intelligibility of speech in the dark zone. In other words, satisfaction of both objective intelligibility and subjective acceptability require changes to the masking signal level in opposite directions. Given this constraint on the loudness, it will be necessary to also investigate how to reduce the smaller contributions from the sharpness, roughness and fluctuation strength metrics to the overall annoyance.

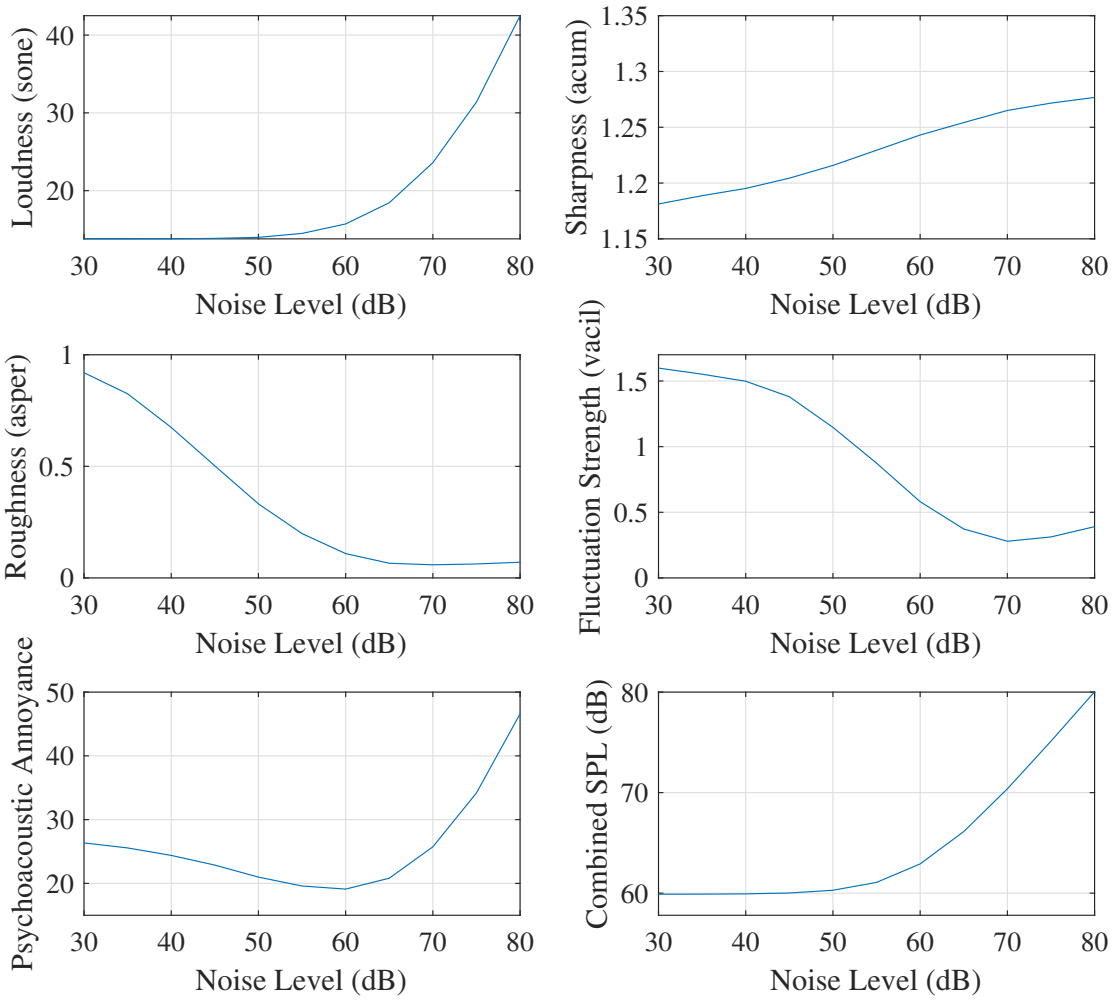


FIGURE 4.14: Loudness, sharpness, roughness, fluctuation strength, psychoacoustic annoyance and SPL, evaluated for a combined speech and noise signal. The speech level is held constant at 60 dB SPL and the noise level is varied from 30 to 80 dB SPL.

4.6 Sound Quality

An alternative criterion which has been frequently posed as a desirable feature of personal audio systems is high audio quality in the bright zone. This approach is not surprising, as the acceptability of novel audio technologies by early adopters depends on both form and function [175]. For a personal audio system, the latter may be expressed in terms of a well-known figure of merit such as total harmonic distortion, or a particular level of acoustic contrast between the bright and dark zones. Quality, on the other hand, is appraised subjectively, so is dependent on the context and the expectations of listeners [175]. Consequently, a universal sound quality appraisal method, which only takes input of audio signals, would be impossible to design and unusable in practice; the scope must be precisely defined.

To a certain extent, designing the system to minimise the psychoacoustic annoyance in the dark zone is a decision that is based on a desire to increase product quality or acceptability [175]. However, psychoacoustic annoyance is specifically limited in scope to give an output which is proportional to features that listeners may find annoying, as opposed to the broader challenge

of responding to any degradation that would reduce overall quality. Nevertheless, metrics which purport to produce an output that correlates with mean opinion scores of basic audio quality do exist, and are used in the telecommunication industry.

The International Telecommunications Union (ITU) has produced a range of recommendations for the evaluation of speech quality in telecommunications systems, which exist in "full reference" form where the algorithm has access to both input and output signals of a telecommunications channel, and "no reference" form, which only use the degraded output of a telecommunication system [176]. Similar categories of objective intelligibility predictors were described in Section 3.1. The most recent "full reference" method, ITU-T P.863: Perceptual objective listening quality assessment (POLQA) [177] replaces older speech quality assessment methods P.862: Perceptual Evaluation of Speech Quality (PESQ) [95], and P.861: Perceptual Speech Quality Measure [178], with increased applicability to super-wideband speech (50 - 14000 Hz) and sensitivity to the type of signal degradations present in contemporary and near-future digital telecommunications systems, such as time warping, packet loss and the use of audio compression codecs. The POLQA algorithm takes input of a clean reference signal and a degraded signal, which are then segmented into time frames. The frames of the degraded signal are time-aligned to the reference, then both signals are converted to an internal representation, which is analogous to the form audio signals are understood to take within the human auditory system, including details of subjective loudness in sones and perceptual frequency on the Bark scale. Various signal processing operations remove perceptually irrelevant details from the degraded signal, before six quality indicators are computed. These indicators measure differences between the reference and degraded internal representations of the signal in terms of frequency response, noise level, reverberation and three other measures of difference in the time-pitch-loudness (perceptual) domain.

Each ITU-T P86x recommendation includes a reference software implementation. This user-friendliness has led to widespread use of the methods, potentially without due concern for their range of applicability or the types of signals that may be input into the software for which reliable or truly meaningful operation is expected. This includes examples of PESQ being used as a sound quality evaluation method for personal audio systems [9, 19, 94]. This edition of the recommendation [95] does not specifically limit use of the method to telecommunications equipment, though the descriptions of validation experiments exclusively consider the use of telephone handsets (whether physical or simulated) as input and output devices for signals processed by the PESQ algorithm. The more recent recommendation P.863 (POLQA) [177] is likewise designed for telecommunications network and equipment testing, and specifically states that it is not intended to be used to assess the effect of acoustic noise in the receiving environment, a case which is only implied in PESQ. In both recommendations, the only consideration of background noise is made for noise in the sending environment which is transmitted to the receiver by the channel under test. This suggests that this type of quality evaluation is likely to be unsuitable for the use-case intended in this work, as the personal audio system generates masking noise in the environment where the listener is situated, which in many anticipated use-cases will itself be noisy.

Nevertheless, in ancillary investigations where the effects of ambient noise are not significant, such as assessing the extent of any audible degradation of the speech signal by the zoning methods themselves, PESQ and POLQA remain highly attractive metrics. In these cases, including

information about speech quality, as well as speech intelligibility, is a potentially valuable addition. This is particularly important when the logistic relationship between intelligibility metrics and percent-correct scores saturate at high SNRs, as shown for the SII in Figure 3.7, as a quality metric could be used to distinguish between conditions with perfect intelligibility.

4.7 Summary

For a personal audio system to be implemented successfully, the experience of all nearby listeners must be taken into account. This means that it is necessary to evaluate the psychoacoustical impact of the system, as well as its technical and privacy-related performance. In Chapter 3 it was shown that speech intelligibility can be estimated based on properties of the signals received in the bright and dark zones, using objective intelligibility metrics. This approach reduces the dependence on lengthy and costly listening tests when provisioning a new system. The present chapter has described a similar approach that can be used to reduce or eliminate the features of the masking signal that may be perceived as annoying, thus raising the perceptual acceptability of a given system.

A number of established methods for extracting perceptually relevant information from signals have been described in this chapter. These metrics output single-number ratings that correspond to the sensations of loudness, sharpness, roughness and fluctuation strength, which in turn can be combined to form a value that corresponds with the psychoacoustic annoyance [79]. The implementations of these metrics have been benchmarked against published results from subjective tests and have been applied to speech and noise signals representative of those emitted by personal audio systems to find the signal conditions where each of the metrics dominantly contributes to the psychoacoustic annoyance.

Consideration of the results from this chapter and Chapter 3 indicates that loudness is a significant contributor to both the psychoacoustic annoyance of a signal, and its efficacy as a masker. This trade-off between objectives marks the necessity of designing the masking signal based on both intelligibility and annoyance predictions simultaneously, using simulated recordings of the signals in each listening zone. However, in order to acquire these signals, an understanding of the physical performance limitations of the employed loudspeaker array must also be known. The following chapter discusses the physical performance evaluation of a linear loudspeaker array in a room, including an assessment of how this is affected by changes in the geometry of the array and the zones.

Chapter 5

Loudspeaker Array Performance Evaluation

Understanding the physical performance limitations of loudspeaker array-based sound zoning systems is a vital step towards the provision of speech privacy control. In Chapter 3, a link was demonstrated between the SNR difference between listening zones and the speech intelligibility contrast. This difference in inter-zone SNR can be achieved by focussing speech and masking noise into separate listening zones, and the degree of separation can be characterised by the frequency dependent acoustic contrast. This chapter contains a description of a prototype loudspeaker array, and provides details on the levels of acoustic contrast that can be achieved with different zonal configurations.

The process of calculating electroacoustical transfer responses from this array to a range of positions in an acoustically treated room is presented, using both omnidirectional pressure microphones for sound zone evaluation, and dummy head recordings for use in listening tests. Further to this discussion, the mathematical derivation of the Acoustic Contrast Control method is presented, and the performance of the loudspeaker array is described in terms of the acoustic contrast for several zone positions within the room. Two smaller arrays, created using subsets of the full array, are used to investigate the effects of inter-element spacing and array aperture on the frequency dependent acoustic contrast, and these array designs are used as a platform for the perceptual studies presented Chapter 7.

5.1 Experimental Setup

Early investigations into the development of personal audio systems used simulations with various degrees of fidelity and complexity. For example, loudspeakers have been modelled as point monopoles [2, 14] and higher order directional sources [70], and important reflections in listening environments have been simulated using image source modelling [60] and the boundary element method [179]. These simulations are an important step towards identifying and understanding the fundamental limitations of a given system.

When a system is realised, its practical limitations become evident, and the achievable performance in realistic reproduction environments becomes clear. The net effect of the true loudspeaker array directivity, the matching between loudspeaker drivers, any colouration by the signal processing hardware, and the room reverberation can be characterised by the set of electroacoustical transfer responses between the loudspeaker array elements and microphones placed in the room. This section documents the process of measuring these electroacoustical transfer responses, starting with a description of the loudspeaker array prototype that is used in the remainder of this thesis.

5.1.1 27-Channel Loudspeaker Array

Circular and spherical loudspeaker arrays that encompass the listening space have been used in many investigations relating to multi-zone sound field reproduction [8, 16, 34, 53]. A circular geometry is mathematically convenient for certain zonal control formulations such as orthogonal basis expansion methods or where sound fields are represented by their spherical harmonic decomposition, however, this mathematical convenience does not translate into practical functionality. Line arrays, and to a lesser extent compact cylindrical arrays [64], can often offer comparable performance in a form-factor that is much easier to integrate into a given reproduction space.

Consequently, a line array is selected as the platform for the investigations presented in the remainder of this thesis. The employed 27-channel array was originally developed for in-car use and has been previously described by House et al. [180]. Engineering drawings of the array geometry with important dimensions highlighted are provided for reference in Figure 5.1.

A fundamental limitation of all uniform line arrays is the upper frequency limit at which grating lobes in the directivity of the array begin to develop. For horizontal beamforming, this frequency limit is reached when the horizontal spacing between array elements δ is equal to half a wavelength. Equivalently, a target spatial aliasing frequency limit f_{max} can be achieved by setting δ such that $\delta = 0.5c/f_{max}$, where c is the speed of sound. In practice, δ is limited by the physical size of the loudspeaker drivers. In the presented array, alternate drivers are vertically displaced from each other to allow the horizontal spacing of the drivers to be decreased, compared with mounting them side-by-side, which in turn raises the aliasing frequency limit of the line array.

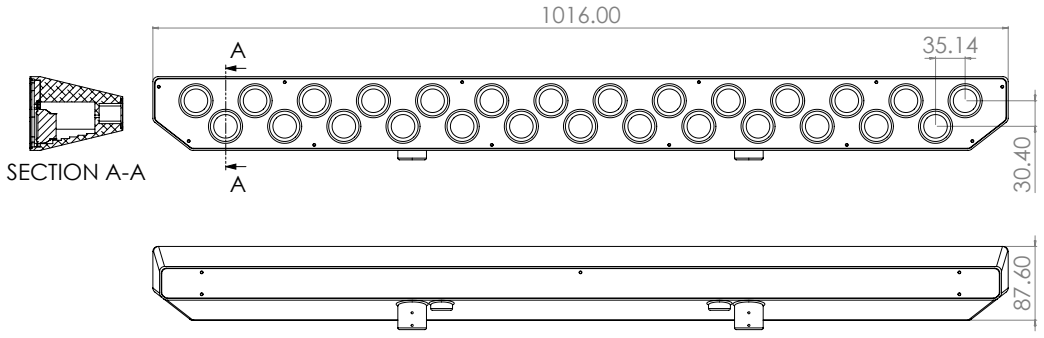


FIGURE 5.1: Engineering drawing of the 27-channel loudspeaker array, produced from original CAD files supplied by Charlie House. Dimensions in mm.

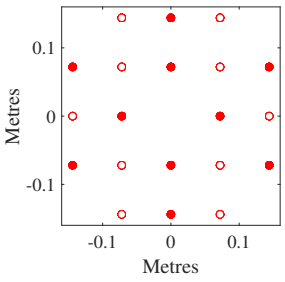
The process of generating sound zones requires knowledge of the electroacoustical transfer responses between each of the individual loudspeaker array elements and microphones located within each desired sound zone. The process of acquiring these measurements can be significantly accelerated by using an array of microphones to simultaneously capture the transfer responses from each speaker in turn.

5.1.2 Microphone Arrays

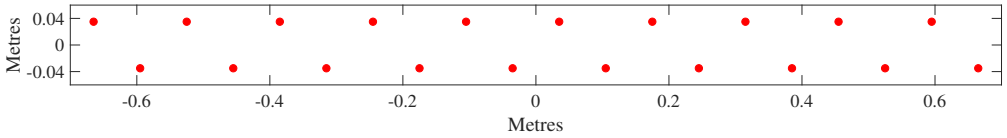
To operate a sound zoning system, it is necessary to output signals from multiple loudspeakers simultaneously, i.e. to use a loudspeaker array, as the physical interactions between these acoustical signals causes the desired bright and dark zones. Unlike active noise control systems, which use microphones as error measurement devices during operation, the proposed system does not require inputs from microphones once a system has been set up, as it assumes that the acoustical conditions in the reproduction space can be measured once and used throughout a devices operational life. Transfer responses to multiple points in the room are required to produce spatially separated sound zones, and in principle these can be captured using a single microphone which is moved after each measurement. If an array of microphones are used, the transfer responses to multiple locations can be captured simultaneously, reducing the acquisition time. Note that no microphone array processing, e.g. beamforming, is used - the response from each microphone in the array is considered independently.

Two arrays of PCB Piezotronics Model 130F20 ¼ inch pre-polarised electret microphones are used as sensors in the transfer response measurements. The first array, a square grid of 20 microphones, spans an area comparable to the size of a human head, and is thus intended to simultaneously capture the transfer responses required for the definition of individual listening zones. The second array is formed of two lines of microphones, with the length similar to the width of the source loudspeaker array. The geometry of the dual line array is designed such that a regular grid of microphone positions can be built up by moving the whole array. This process enables transfer responses from a large area in the room to be captured, in turn facilitating the production of contour plots that show how various perceptual and physical quantities vary over space. Diagrams of the two microphone arrays are displayed in Figure 5.2.

Further to the responses measured with omnidirectional microphones, transfer responses were measured using a Knowles Electronics Mannequin for Acoustics Research (KEMAR) head and torso simulator [181]. These measurements enable simulated playback of the array over headphones, which is used in the listening tests described in Chapter 6.



(A) Square grid microphone array



(B) Dual line microphone array

FIGURE 5.2: Positions of microphones within the two measurement microphone arrays used for transfer response measurements. In the upper panel, the filled and empty markers indicate microphones used for the optimisation of zonal filters, and the evaluation of system performance respectively.

5.1.3 The ISVR Audio Laboratory

Transfer response measurements were carried out in a purpose-built audio laboratory. The room, pictured in Figure 5.3, measures 3.7 × 4.4 × 2.3 metres, and all walls are treated with fabric panels containing mineral wool insulation. The floor is carpeted and the ceiling is finished with suspended ceiling tiles and open-cell foam insulation. This provides a mid-frequency reverberation time $T_{60,mf}$ of 0.11 seconds, given by the arithmetic average of 500 Hz, 1 kHz and 2 kHz octave band reverberation times [182]. The reverberant properties of the audio laboratory can also be described by considering the direct and reverberant sound fields that are produced when a source operates within the room. The intensity of the direct component decays geometrically with distance, whereas the diffuse reverberant component is, to a first approximation, uniformly distributed throughout the room. The distance from the source at which the intensity of these two components is equal is described as the critical distance, d_c and is given by [183]

$$d_c = 0.1\sqrt{\frac{GV}{\pi T_{60}}}, \tag{5.1}$$

where V is the room volume in m^3 and G is the directivity of the source. G is defined in this case as the ratio of the maximum intensity in a given direction, usually on-axis with the source, to the average intensity over a sphere surrounding the source. For a monopole source, i.e. $G = 1$, the critical distance for the listening room is 1.04 metres.

An extruded aluminium structure spanning the walls and ceiling of the laboratory supports a 39-channel array of KEF HTS3001 loudspeakers, which are individually addressable through a MADI interface. These loudspeakers were not used in the experiments presented in this thesis.

FIGURE 5.3: The ISVR audio laboratory.

5.1.4 Transfer Response Measurements

Sections 5.1.1 to 5.1.3 have described the loudspeaker array, microphones and reproduction environment which facilitate and characterise the transfer response measurements. Assuming linearity and time-invariance, the electroacoustical transfer responses between the loudspeaker and microphone arrays in the room capture all the information required by the sound zoning algorithms to generate sound zoning filters. This includes, but is not limited to, the acoustic propagation delay from each source to each sensor, the frequency response and sensitivity variation between each driver, and the effects of reflections from room boundaries. The propagation delay, and for simple cases the directivity of individual array elements, may be estimated from the geometry of the loudspeaker array and zones under test. However, the other features of the impulse response are more challenging to approximate. The effects of using measured or modelled transfer responses on acoustic contrast and speech privacy are discussed further in Chapter 8.

The purpose of the response measurements was to provide data that could later be used by the sound zoning algorithms to optimise loudspeaker filters, and also to simulate playback from loudspeaker arrays in listening tests, without requiring the subjects to sit in front of the loudspeaker array operating in real-time. The 27-channel loudspeaker array described in Section 5.1.1 was used as the primary source in the measurements. Table 5.1 provides a summary of the measurements taken.

Microphone Array	**Microphone Channels**	**Positions**	**Total**
Square Grid	20	19	10260
Dual Line Array	20	4	2160
KEMAR Mannequin	2	59	3186
			15606

TABLE 5.1: Transfer response measurement details. The totals in the rightmost column count the number of individual impulse responses recorded for each loudspeaker-microphone pairing.

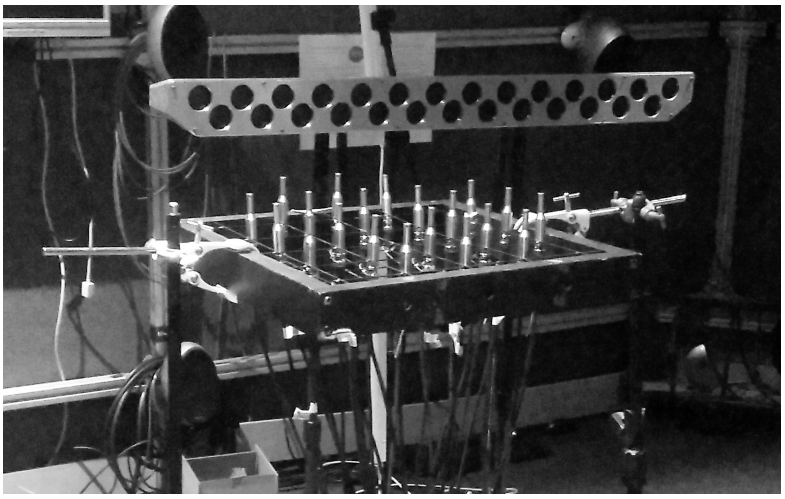


FIGURE 5.4: Foreground: Measurement microphone grid, Background: 27-channel loudspeaker array.

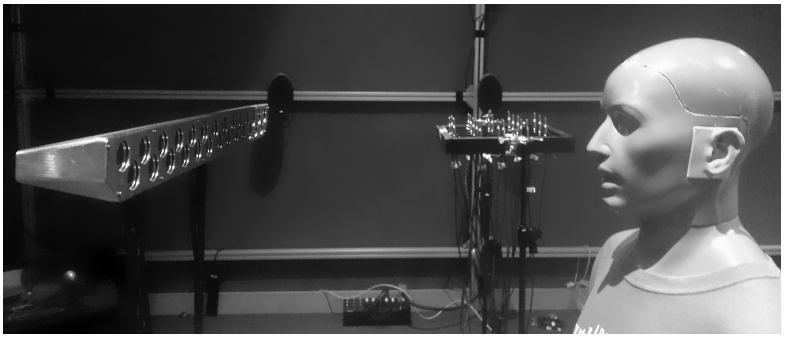


FIGURE 5.5: Left: 27-channel loudspeaker array, Centre: Measurement microphone grid, Right: KEMAR mannequin.

Figures 5.4 to 5.7 show images of the loudspeaker and microphone arrays during the measurements. Binaural room impulse responses captured using the KEMAR mannequin include two different orientations; the mannequin was either positioned facing forwards or facing the centre of the array. These different orientations provide different impressions of the location of the array in auralisations, allowing the flexibility to present speech content from the front, as is standard in spatial listening tests [184], or to provide spatial cues to the listener as to whether they are situated in the left or right listening zone. The channel numbering of the microphone grid depicted in Figure 5.6 ensures that channels 1-10 and 11-20 span the entire grid, so that contiguous subsets of ten channels could be selected to respectively optimise the sound zoning filters and evaluate their performance.

Measurements were automated by outputting a 10-second logarithmic sine-sweep through each channel of the source array in turn, whilst simultaneously recording the input from the connected microphone array. Figure 5.8 shows the various equipment connections necessary for taking measurements from the 27-channel array. After capturing recordings from all connected microphones, impulse responses were recovered by convolution with a pre-calculated inverse filter [185]. The associated frequency domain transfer responses were then calculated using the Discrete Fourier Transform.

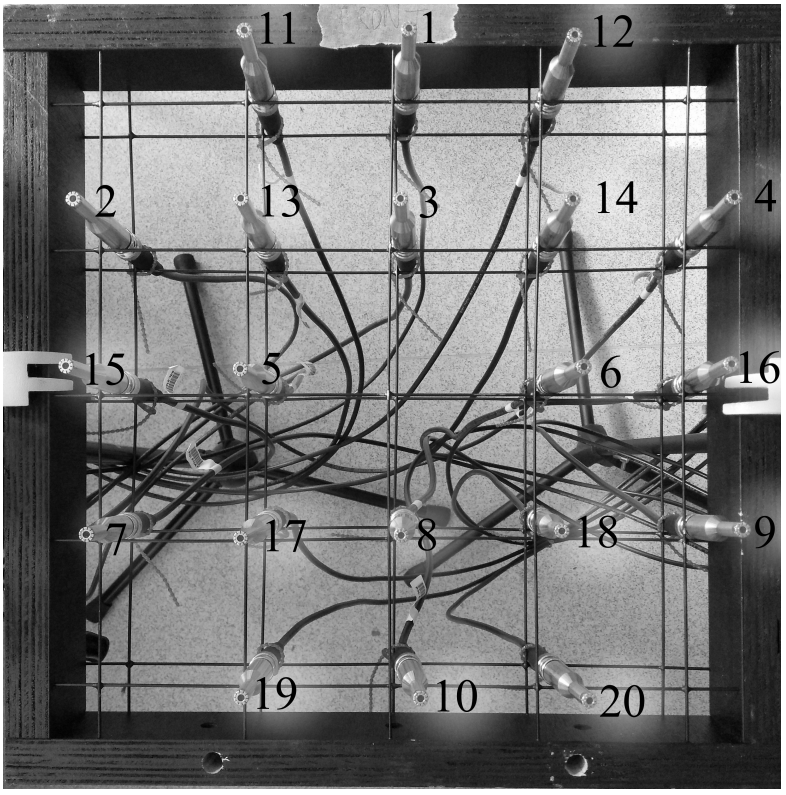


FIGURE 5.6: Square, 72 mm pitch measurement microphone grid, with microphone channel numbers indicated.

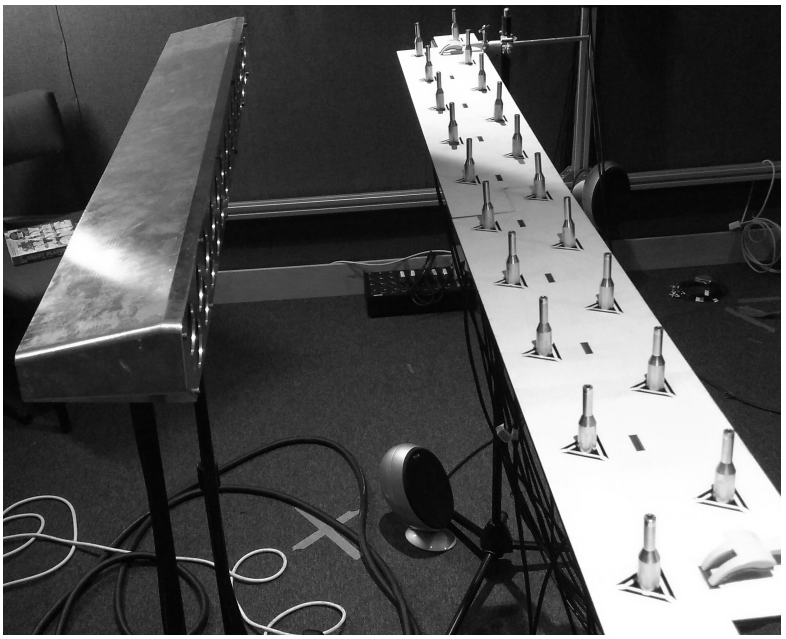


FIGURE 5.7: Left: 27-channel loudspeaker array, Right: Measurement microphone line array.

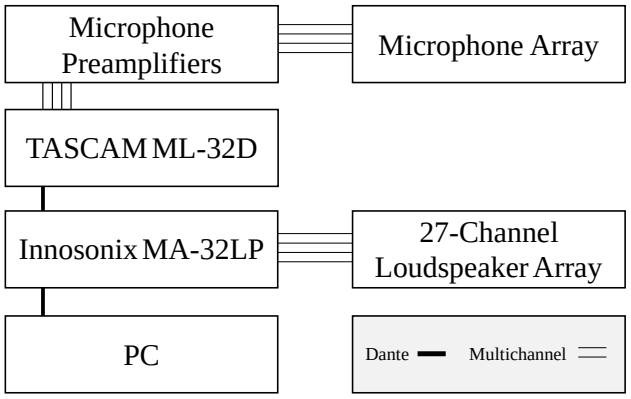


FIGURE 5.8: Equipment connections for transfer response measurements from the 27-channel loudspeaker array.

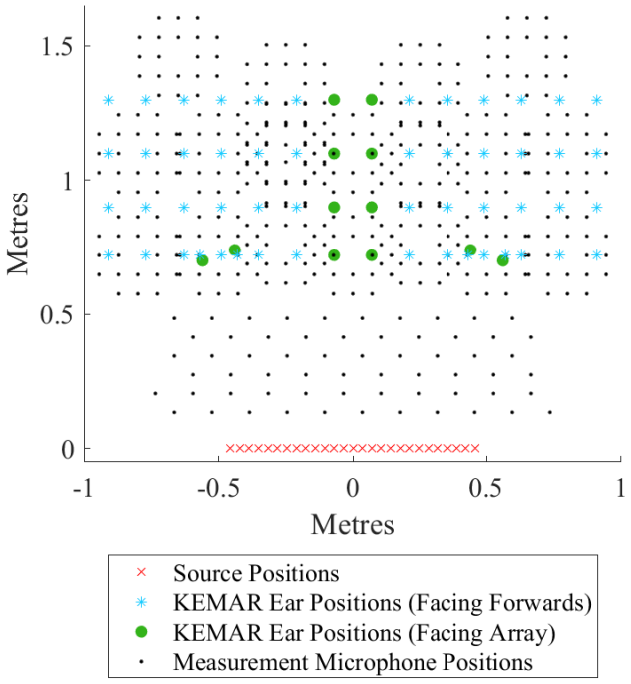


FIGURE 5.9: Map of source and microphone positions for the transfer response measurements. All points are 1.22 metres above floor level.

The full range of microphone positions for which transfer response measurements were carried out are indicated in Figure 5.9. The coordinate system origin is coincident with the centre of the loudspeaker array and all microphones were positioned 1.22 metres above floor level.

5.2 Sound Zoning Filter Design

Once transfer response measurements have been acquired, they can then be used in a sound zoning algorithm. As mentioned in Section 2.1.4, the array weights necessary for loudspeaker array-based sound field reproduction can be calculated using a variety of methods. Frequency domain methods can be categorised into those that optimise based on the energy within each sound zone, disregarding phase, and those that seek to minimise the error in the reproduction of a specified sound field. The prototypical examples within these two categories are Acoustic Contrast Control (ACC) and the Pressure Matching (PM) method. ACC is selected as the most appropriate method for the specific goal of speech privacy control, as it directly maximises the acoustic contrast between listening zones, a quantity which is well correlated with the level of speech intelligibility contrast [9]. Furthermore, using ACC simplifies the analyses presented in Chapters 7, 8 and 9 as there is no requirement to specify a target sound field, as is the case with PM, or optimise the tuning parameters required by hybrid sound zoning methods, as previously discussed in Section 2.1.4.

In the case of the proposed speech privacy control system, two sets of loudspeaker weights must be calculated. The first follows conventional personal audio nomenclature by maximising the level of a speech signal in the bright zone and minimising the level in the dark zone. The second process has the opposite goal; the level of a masking signal must be maximised in the dark zone whilst minimising leakage into the bright zone, where the target listener is situated. The principle of linear superposition allows the signals from both sound zoning processes to be combined. The following derivation of the ACC filters is presented for the first process, which focuses the speech signal into the bright zone. The derivation for the secondary process is identical, except for the labelling of the zones. For comparison with the ACC approach, a derivation of the PM method is provided in Appendix A.

5.2.1 Acoustic Contrast Control

The process of ACC can be performed in the frequency domain, but the eventual goal is a set of time domain FIR filters that drive each loudspeaker in the source array. Accordingly, the process detailed below is carried out at equally spaced frequencies, up to the Nyquist frequency, then this information is aggregated into a FIR filter via the frequency sampling method.

At each frequency, the system must determine the amplitude and phase that each loudspeaker must be driven at in order to maximise the acoustic contrast. This information is encoded in the complex elements of the source strength vector $\mathbf{q}$, which can be expressed as

$$\mathbf{q} = j\omega\rho_0\mathbf{Q}_m, \tag{5.2}$$

where ω is the angular frequency, ρ_0 is the density of air in kg/m^3 and $\mathbf{Q}_m$ is a vector of source volume velocities in m^3/s.

For the N_b microphones in the bright zone, the pressures $\mathbf{p}_b$ in Pascals are given by the product of the complex transfer response matrix $\mathbf{Z}_b$ and the source strengths $\mathbf{q}$, that is,

$$\underset{(N_b\times 1)}{\mathbf{p}_b} = \underset{(N_b\times M)(M\times 1)}{\mathbf{Z}_b \quad \mathbf{q}} . \tag{5.3}$$

Likewise, for the dark zone,

$$\underset{(N_d\times 1)}{\mathbf{p}_d} = \underset{(N_d\times M)(M\times 1)}{\mathbf{Z}_d \quad \mathbf{q}} . \tag{5.4}$$

The transfer response matrices $\mathbf{Z}_b$ and $\mathbf{Z}_d$ may be specified from impulse response measurements, as described in Section 5.1.4 or estimated based on the geometry of the loudspeaker array and zones.

The objective of ACC is to maximise the ratio of the mean squared pressure in the bright zone to that in the dark zone. This ratio, the Acoustic Contrast C, can be expressed using the transfer responses and source strengths as follows:

$$C = \frac{N_d \mathbf{q}^H \mathbf{Z}_b^H \mathbf{Z}_b \mathbf{q}}{N_b \mathbf{q}^H \mathbf{Z}_d^H \mathbf{Z}_d \mathbf{q}}, \tag{5.5}$$

where the superscript $\{\}^H$ indicates the complex conjugate (Hermitian) transpose. The acoustic contrast is usually quoted in decibels, by taking ten times the base 10 logarithm of the value of C calculated in Equation 5.5,.

Maximising acoustic contrast alone can lead to solutions for the source strength vector $\mathbf{q}$ with very high signal strengths. In a realised system, this can result in high pressure levels away from bright and dark zone control points, heavy cancellation between adjacent array elements, and ultimately reduced available dynamic range due to amplifier and loudspeaker power limitations [58]. It is therefore prudent to place a constraint on the array effort E, defined as the sum of the modulus squared signals driving the array, normalised by the input signal required to drive a single element at the centre of the array so that the mean square pressure in the bright zone is the same as that when the array is being driven by $\mathbf{q}$, that is

$$E = \frac{\mathbf{q}^H \mathbf{q}}{|q_0|^2}, \tag{5.6}$$

where q_0 is the source strength of a single monopole that produces the same mean square pressure in the bright zone as the array.

Optimal values of $\mathbf{q}$ can be calculated by writing a constrained optimisation problem, which can be solved using the method of Lagrange multipliers. For the purpose of finding the weights that minimise the pressure in the dark zone, one constraint is that the bright zone pressure $\mathbf{p}_b^H \mathbf{p}_b$ is equal to some freely selectable constant B. The array effort constraint $\mathbf{q}^H \mathbf{q} = E$, is maintained by varying the value of β_2. However, the formulation of the Lagrange multipliers in the following derivation sees β_2 also serving as a regularisation parameter to improve the numerical conditioning of the solution. Therefore, the value of β_2 can be chosen to be any value greater than or equal to that which provides the maximum permitted effort E. The Lagrangian for the bright zone optimisation problem is

$$L = \mathbf{p}_d^H \mathbf{p}_d + \beta_1(\mathbf{p}_b^H \mathbf{p}_b - B) + \beta_2(\mathbf{q}^H \mathbf{q} - E). \tag{5.7}$$

The vector of complex differentials of L are given by

$$\frac{\partial L}{\partial \mathbf{q}} = 2(\mathbf{Z}_d^H \mathbf{Z}_d \mathbf{q} + \beta_1 \mathbf{Z}_b^H \mathbf{Z}_b \mathbf{q} + \beta_2 \mathbf{q}), \tag{5.8}$$

which is equal to zero if

$$\beta_1 \mathbf{q} = -[\mathbf{Z}_b^H \mathbf{Z}_b]^{-1}[\mathbf{Z}_d^H \mathbf{Z}_d + \beta_2 \mathbf{I}]\mathbf{q}. \tag{5.9}$$

This is an eigenvalue problem of the form $\mathbf{A}\mathbf{v} = \beta\mathbf{v}$. As the requirement is to minimise the Lagrangian, the eigenvector associated with the *smallest* eigenvalue of

$$[\mathbf{Z}_b^H \mathbf{Z}_b]^{-1}[\mathbf{Z}_d^H \mathbf{Z}_d + \beta_2 \mathbf{I}] \tag{5.10}$$

is the required solution for $\mathbf{q}$. However this formulation may lead to numerical instability due to ill-conditioning of $[\mathbf{Z}_b^H \mathbf{Z}_b]$. A more numerically stable formulation is found by inverting Equation 5.10, which yields

$$[\mathbf{Z}_d^H \mathbf{Z}_d + \beta_2 \mathbf{I}]^{-1}[\mathbf{Z}_b^H \mathbf{Z}_b]. \tag{5.11}$$

The solution $\mathbf{q}$ is then the eigenvector that corresponds to the *largest* eigenvalue of Equation 5.11. Here, β_2 carries a dual purpose, limiting array effort and regularising the matrix $\mathbf{Z}_d^H \mathbf{Z}_d$. These purposes may be separated by partitioning β_2 into $\beta_{\text{min}} + \beta_{\text{eff}}$ (as in Ref. [74]) , where β_{min} is first set to provide sufficient numerical stability, then β_{eff} is subsequently adjusted, if necessary, to limit the array effort so that it does not exceed the pre-specified limit E. β can be freely selected at each frequency. Therefore, regularisation can be specifically emphasised at frequencies where $\mathbf{Z}_d^H \mathbf{Z}_d$ is poorly conditioned. This is achieved by setting $\beta = \beta_0 \kappa$. where κ is the condition number of $\mathbf{Z}_d^H \mathbf{Z}_d$.

5.2.2 Construction of filters from optimal responses

The process described above can be followed to generate a single-sided filter response in the frequency domain. This can be converted to a finite impulse response (FIR) filter for each loudspeaker by concatenating the single-sided response, $\mathbf{q}$, with its index-reversed complex conjugate, to give the two-sided conjugate symmetric frequency response. When the inverse Fourier transform is applied to this two-sided response, the corresponding real impulse response is produced. To ensure that this impulse response is causal, it is necessary to include a modelling delay [186].

When the ACC method is used, there is no constraint placed on the phase of the reproduced signal within the bright zone. When left unconstrained, the resulting time domain filters exhibit a high level of noise, and the impulse, which should occur near the centre of the filter, is smeared.

This time-smearing degrades the intelligibility of speech signals processed using these filters, as the onset of transient sounds such as consonants are smoothed. To illustrate this smearing, Figure 5.10 shows the 27 filter impulse responses that were optimised to focus programme material into the bright zone of the array, using ACC, without any phase constraints being placed on the reproduced sound field.

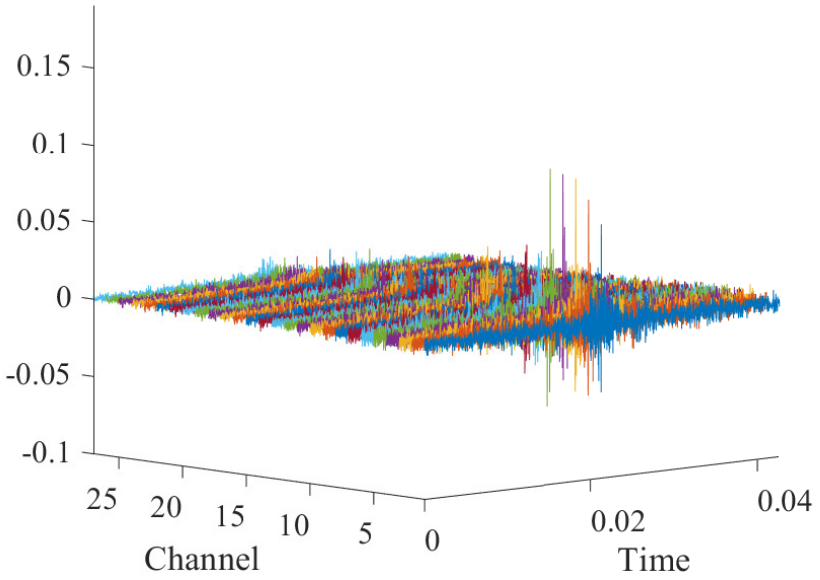


FIGURE 5.10: Waterfall plot of filter impulse responses for focussing speech programme material into the bright zone of the 27 channel array, without applying phase correction.

Using the ACC method provides a degree of flexibility in the construction of the time-domain FIR filters. As the method only constrains the magnitude of the response at each bright zone microphone, it is possible to adjust the phase response at a single microphone within the bright zone. This is achieved by adding a series of steps after the production of the initial frequency domain filters, $\mathbf{q}$. At each frequency, the pressure at the selected microphone p_m is evaluated, by taking the corresponding row of the transfer response matrix, $\mathbf{Z}_{bm}$, and multiplying it by the uncorrected filter, $\mathbf{q}$:

$$\mathbf{p}_m = \mathbf{Z}_{bm} q \tag{5.12}$$

The phase correction ϕ is calculated by calculating the phase angle of p_m and rotating the phase of each element of $\mathbf{q}$ by a corresponding amount:

$$\phi = \angle p_m \tag{5.13}$$

$$\mathbf{q}_{corrected} = \mathbf{q} e^{j\phi} \tag{5.14}$$

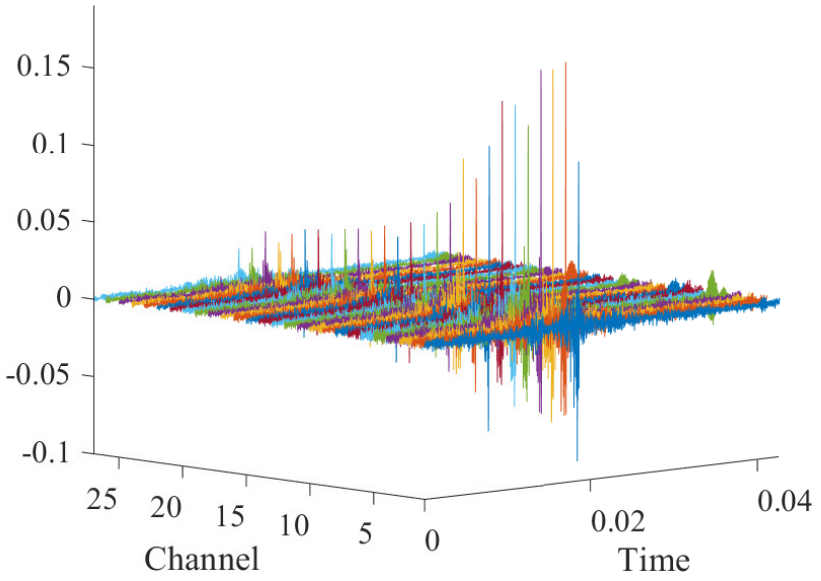


FIGURE 5.11: Waterfall plot of filter impulse responses for focussing speech programme material into the bright zone of the 27 channel array, after applying phase correction.

This has the effect of setting the phase equal to zero at one microphone, and is equivalent to applying an all-pass filter to the time domain filters [63]. Consequently, this has no effect on the acoustic contrast, as the magnitude of the response is unaffected, but cleaner impulse responses are produced when the frequency domain responses are converted to the time domain. This is demonstrated in Figure 5.11, which shows the 27 filter impulse responses optimised to focus programme material into the bright zone of the array, using ACC with phase correction applied. The impulses at the centre of each corrected filter, displayed in Figure 5.11 are significantly more prominent than those generated without the phase correction, shown in Figure 5.10. Consequently, impulsive sounds processed by the array are better preserved at sensor locations when phase-corrected filters are used, compared to the uncorrected case. For this example, the bright zone targeted by the ACC process is situated to the right of the array. Accordingly, the impulse responses for low channel numbers, closest to the bright zone, are much larger in amplitude compared to the opposite side of the array.

5.3 Effects of Zonal Geometry

The previous section has described how sound zoning filters can be optimised to provide the maximum level of acoustic contrast between a pair of listening zones, using information from electroacoustical transfer responses. The level of performance that can be achieved by this optimisation process depends on several factors, including the power deliverable by the array [2], how the system is regularised [58], and the position of the zones with respect to the loudspeaker array [187]. The latter is investigated in this section for the 27-channel loudspeaker array described in Section 5.1.1, in order to understand the physical limitations of this particular device. The wide range of microphone locations recorded in Figure 5.9 allows the effects of zone position

on the acoustic contrast to be evaluated. Firstly, the effect of adjusting the angular span of the zones is explored. The angular span is defined as the angle between the lines that connect the centre of each zone to the centre of the array. The microphones used to optimise the ACC filters are selected randomly from within a circular region with a radius of 0.2 metres at a fixed distance from the centre of the array, as pictured in the left panel of Figure 5.12. A separate set of microphones are chosen from the same regions to evaluate the acoustic contrast, so that no transfer responses are re-used. Disjoint sets of evaluation and optimisation microphones are used to reduce bias in the acoustic contrast estimates [60, 188]. The acoustic contrast results shown in the right panel of Figure 5.12 have been formed by averaging the acoustic contrast achieved with 100 different combinations of 10 optimisation and 10 evaluation microphones within each zone.

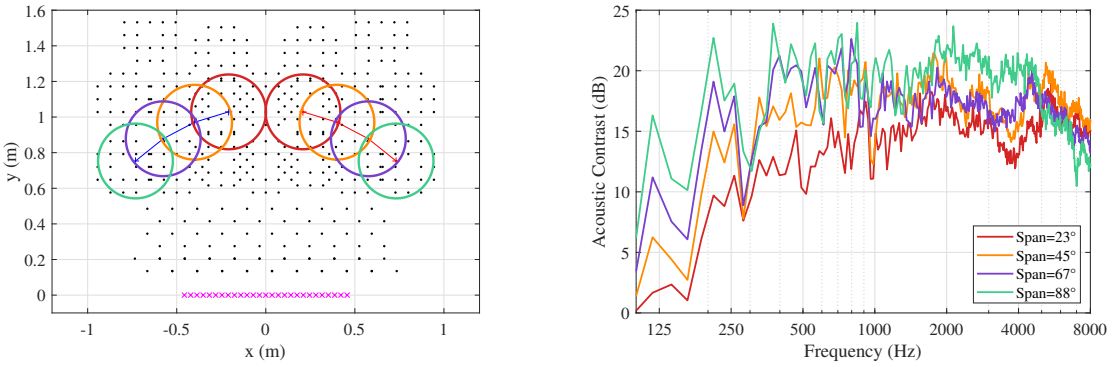
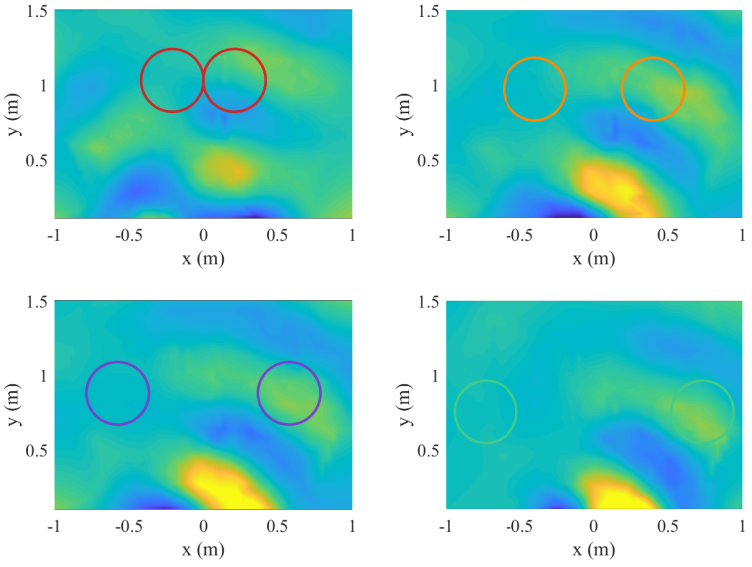


FIGURE 5.12: Left: Bright and dark zone locations with zone centres all situated at 1.05 metres from array centre. Right: Measured acoustic contrast using corresponding zone locations.

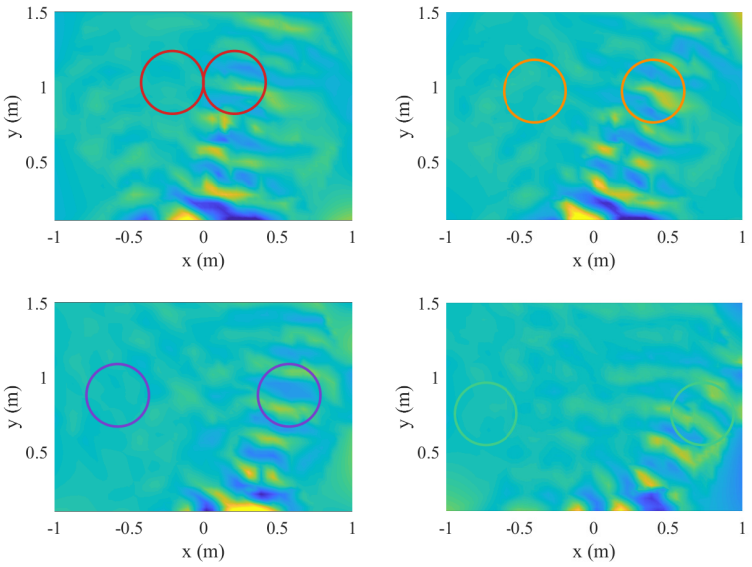
The results in Figure 5.12 show a general trend that when the bright and dark zones are located further apart, i.e. at a wider angular span, the acoustic contrast that is achievable increases. For all considered zonal geometries, acoustic contrast is reduced at low frequencies as the regularised ACC process limits the energy which would be required by the array to drive the pressure in the dark zone to zero [189]. This performance limitation means that the benefit to using wider angular spans, in terms of acoustic contrast, is more significant at low frequencies as the associated longer wavelength prevents the creation of small, localised regions of high or low SPLs. A further low frequency limitation is caused by the restricted aperture of the loudspeaker array. At frequencies above 1 kHz, the difference between the acoustic contrast performance with each angular span is less significant as the size of the zones becomes large compared to the wavelength.

These conclusions can be verified by calculating and visualising the full radiated sound field from the array at particular frequencies. Tonal playback from the array can be simulated by multiplying a new transfer response matrix $\mathbf{Z}_a$, populated with all the microphone measurements, with the filter vector $\mathbf{q}_b$ at each frequency. The resulting pressure fields, at 500 Hz and 1500 Hz are displayed for the four pairs of zones in Figure 5.13; in each contour plot, the bright zone is on the right and the dark zone is on the left. At 500 Hz, when the zones are situated close together, a second beam to the left of the dark zone is produced. This is a consequence of requiring the ACC algorithm to produce zones of high and low SPL that are very close together. Due to the specification of the zones in this example, there is no penalty applied to radiation from the loudspeaker array in other directions, as this does not result in an increase to the acoustic potential energy in the dark zone. However, in reverberant environments or those with strongly

reflective surfaces, this type of spurious radiation can be detrimental to the acoustic contrast. This highlights the importance of including such reflections when capturing or modelling the electroacoustical transfer responses that are used in the sound zoning process [12, 72]. When zones are spaced further apart, the appearance of these side-lobes is reduced, as the ACC process can efficiently produce a single, wide beam that covers the bright zone. At 1500 Hz, shown in Figure 5.13b, the reproduced beams are significantly narrower, enabling a higher degree of control even when the bright and dark zones are close together. The contour plots in Figure 5.13 also confirm that the reproduced sound fields in the bright zone are approximately planar and emerge from the centre of the loudspeaker array, as described in Section 2.1.4.



(A) 500 Hz



(B) 1500 Hz

FIGURE 5.13: Pressure fields produced by the 27-channel loudspeaker array at A) 500 Hz and B) 1500 Hz for the four pairs of zones described in Figure 5.12. In each subplot, the zone locations are indicated by circles, with the bright zone to the right of the array, and the dark zone to the left.

The results shown in Figure 5.13 used four pairs of zones that were set at a constant distance of 1.05 metres from the array, representing a reasonable listening distance that is comparable with the size of the array. This distance also allowed for a wide arc of angular measurement positions to be captured, whilst ensuring all the microphone positions were at least 0.5 metres from the room boundaries, to reduce the effect of reflections. To quantify the effect of source-to-zone distance, another set of four zone locations were selected, in this case spanning a constant 45 degree angle, at a range of distances from the centre of the array. The left panel of Figure 5.14 shows the

locations of these bright and dark zone pairs. With these zonal configurations, the results in the right panel of Figure 5.14 show that acoustic contrast levels are very similar between distances across the frequency range. This supports the explanation given in the previous paragraph that the array is focussing beams of sound towards the centre of each zone, as opposed to exclusively maximising the sound pressure within the zones. The sound field maps in Figure 5.15 confirm this, and show that across the range of array-zone distances considered here, the beamforming pattern produced by the array is very similar. Equivalent acoustic contrast levels are produced as the decay of SPL with distance is constant for both the beam directed towards the bright zone and the null steered towards the dark zone. The closest pair of zones, at a distance of 0.6 metres from the array, has the lowest acoustic contrast among the four alternatives presented in Figure 5.14, again due to the close proximity of the zones to one another.

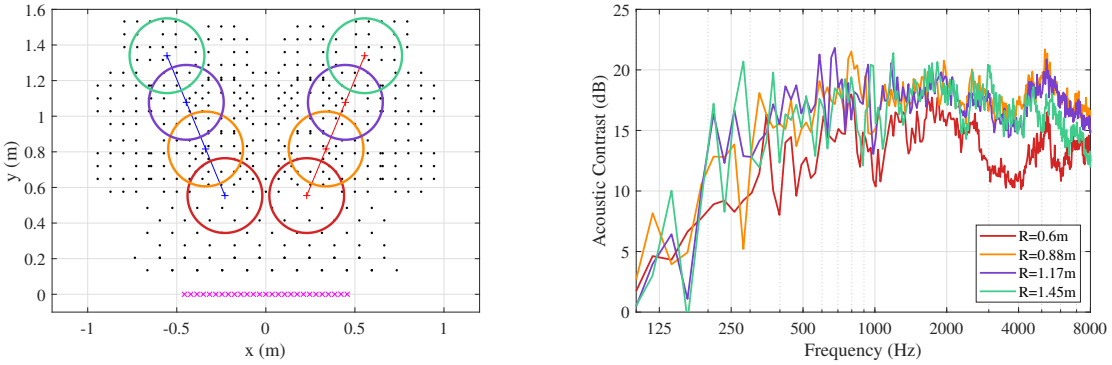
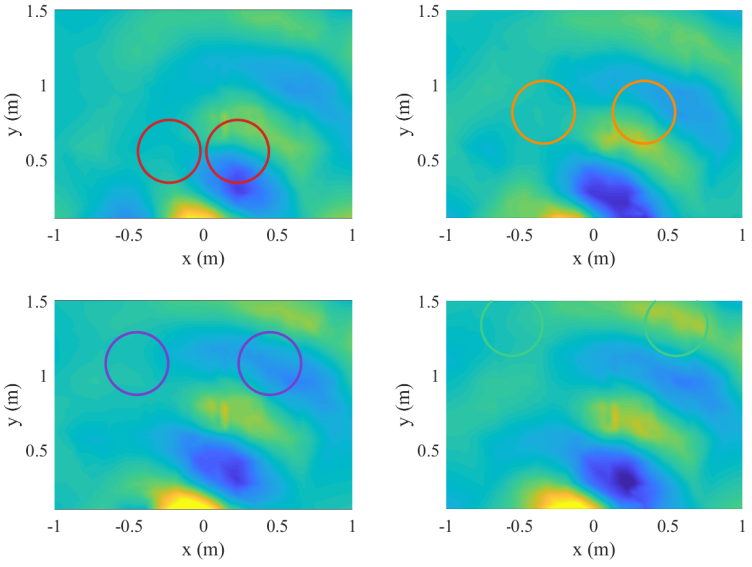
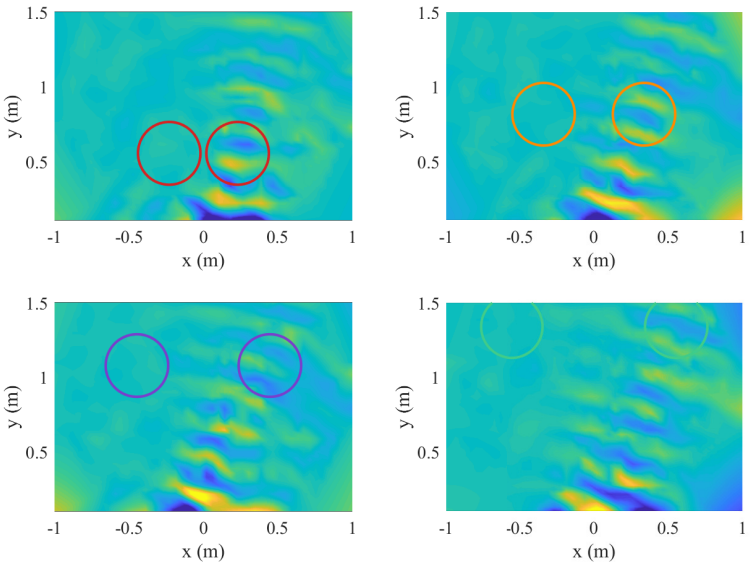


FIGURE 5.14: Left: Bright and dark zone locations with zone centres spanning a constant angle of 45 degrees. Right: Measured acoustic contrast using corresponding zone locations.

As mentioned in Section 5.1.3, the laboratory used for the transfer response measurements and playback has a mid-frequency reverberation time of 0.11 seconds. This short reverberation time allows for control of the sound field at a large distance from the source array. As the distance between the loudspeaker array and zones increases, the direct contribution to the sound field from the loudspeaker array decreases in intensity compared to the diffuse reverberant field. Beyond the critical distance, where the intensity of the reverberant field dominates over the direct sound, the level of acoustic contrast that is achievable by a system decreases, due to the homogeneity of the reverberant sound field. The wider effects of reverberation on acoustic contrast, and their consequent effects on speech privacy are discussed in Chapter 8.



(A) 500 Hz



(B) 1500 Hz

FIGURE 5.15: Pressure fields produced by the 27-channel loudspeaker array at A) 500 Hz and B) 1500 Hz for the four pairs of zones described in Figure 5.14. In each subplot, the zone locations are indicated by circles, with the bright zone to the right of the array, and the dark zone to the left.

5.4 Effects of Array Aperture and Element Spacing

As mentioned in Section 5.1.1, the spatial aliasing frequency for horizontal beamforming is determined by the horizontal spacing of the array elements. The effects of loudspeaker spacing are investigated by selecting two sub-arrays, each with $L = 9$ elements, from the full 27-channel array. Figure 5.16 shows a front-view of the geometry of the full array with the two sub-arrays

highlighted. The alternating vertical offset of odd and even drivers allows the horizontal driver spacing, δ, to be minimised - this small vertical displacement has negligible effect on the vertical directivity. The *narrow* array, indicated by solid outlines in Figure 5.16, has loudspeakers spaced at $\delta_n = 0.035$ m intervals and the *wide* array (dotted outlines) has $\delta_w = 3\delta_n = 0.105$ m, yielding array widths of $D_n = 0.28$ m and $D_w = 0.84$ m.

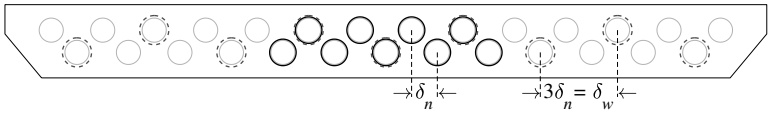


FIGURE 5.16: Front view of loudspeaker array. Groups of 9 elements were selected from a 27-channel array to form two arrays with different horizontal element spacing. Narrow and wide sub-arrays are indicated with solid and dotted lines respectively.

Figure 5.17 shows a plan view of the two 9-channel sub-arrays; the narrow and wide arrays are indicated by orange plusses and purple crosses respectively. The bright and dark zones are specified by the positions of two microphone arrays, the elements of which are shown using red and blue symbols respectively. In both cases, the centres of the bright and dark zones are situated 0.72 m in front of the loudspeaker array in order for all microphones to be within the critical distance , $d_c = 1.04$ metres, of the loudspeaker array in the room. The centres of the two zones are spaced 1.0 m apart, which is comparable to the physical width of the wide array, and is thus consistent with its intended performance limitations, as also investigated in Section 5.3. This symmetrical geometry also corresponds with the de facto standard for evaluating personal audio systems that has emerged from the literature, e.g. [9, 16, 17, 53, 60, 190]. Each zone has a radius of 0.15 m, covering enough space for a human head, with 20 microphones distributed on a grid within each zone. As described in Section 5.3, half of the microphones are used to optimise the zoning filters, while the other half are used to evaluate the reproduced sound field, to reduce bias in the acoustic contrast estimates [60, 188], and are indicated by the filled and empty symbols respectively in Figure 5.17.

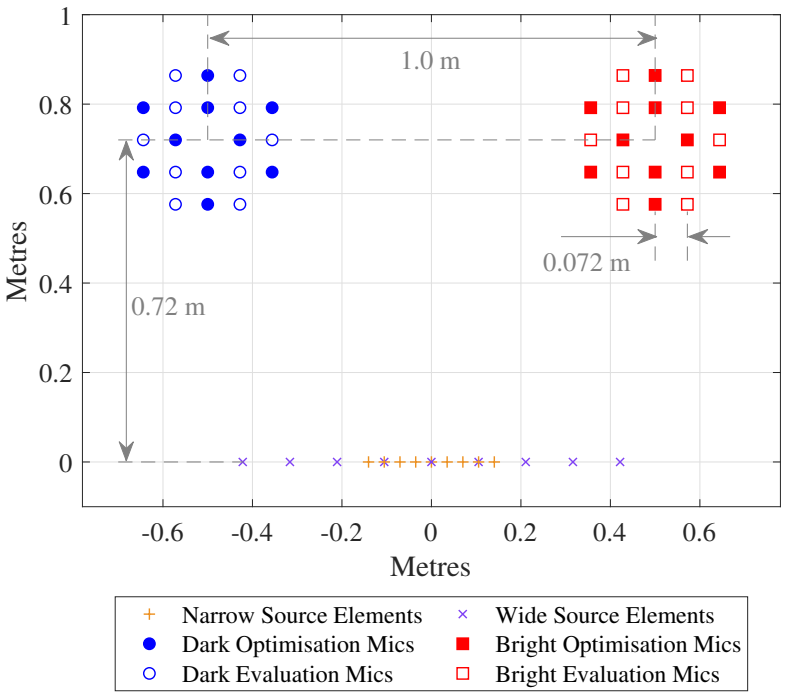


FIGURE 5.17: Plan view of the personal audio system geometry, showing source and microphone locations.

Figure 5.18 shows the acoustic contrast, C, for the narrow and wide arrays. The ACC method is used, and the frequency dependent regularisation parameter is set to be proportional to the condition number κ of $\mathbf{Z}_d^H \mathbf{Z}_d$, as discussed in Section 5.2.1. The proportionality constant, β_0, is set to 10^{-13}. This value was selected to provide a trade-off between robustness to changes in the environment, acceptable acoustic contrast across the speech frequency range, and flatness of the frequency response. A ⅓-octave band equaliser is applied to the input signals to maintain the spectrum of the speech signal in the bright zone and the masker in the dark zone, compensating for any residual colouration to the frequency response caused by the ACC filters. From these results it can be seen that both arrays produce between 15 and 20 dB of contrast across a bandwidth comparable with that of speech. The wide array has a slightly higher contrast at low-mid frequencies due to its wider aperture, but spatial aliasing due to the inter-element spacing causes a substantial reduction in contrast between 2 and 4 kHz. When a fixed number of elements is used, and a uniform inter-element spacing is chosen, a trade-off is always present between these two factors.

To provide further insight into the acoustic contrast results presented in Figure 5.18, room impulse responses were captured using the microphone array grid depicted in Figure 5.17, positioned at multiple locations within the room. Output from the array was simulated using the weights $\mathbf{q}_b$ calculated above to produce maps of the radiated tonal sound fields at 1.2, 2.4 and 4.8 kHz for each array configuration. To more clearly display the beams formed by the arrays at each frequency, the colour scale in Figure 5.19 represents the mean-square pressure, plotted on a decibel scale, as opposed to the real pressure field, which was shown in Figures 5.13 and 5.15 to demonstrate the planarity of the reproduced sound fields. From these results it can be seen that at 1.2 kHz, the aperture of the wide array provides a more focussed beam pattern than the

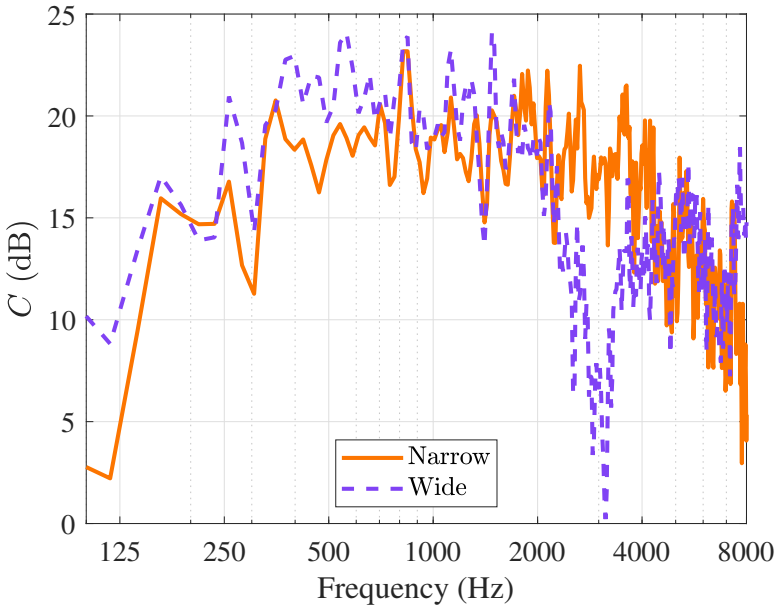


FIGURE 5.18: Acoustic contrast measurements for the narrow and wide loudspeaker array configurations.

narrow array, while at higher frequencies, the aliasing limit of the wide array begins to become evident. With the wide array geometry, at 2.4 kHz, a secondary lobe in the directivity begins to impinge on the dark zone. This measured result is consistent with the computed spatial aliasing frequency of 2367 Hz, using Equation 35 in Ref. [187]. As the frequency increases, this lobe moves through the dark zone, resulting in a pronounced decrease in acoustic contrast around this frequency, with a minimum at 3 kHz, as shown in Figure 5.18. At 4.8 kHz, the narrow array creates a tightly focussed beam in the direction of the bright zone, whereas the sound field in the room with the wide array comprises of multiple side-lobes being radiated in different directions. In rooms with longer reverberation times, the ratio of direct to diffuse sound around the loudspeaker array will decrease, hindering sound field control at a distance from the array. Furthermore, individual reflections may impinge on the dark zone in the same way as is demonstrated with aliased side-lobes in Figure 5.19 [12]. Undertaking a sound field mapping exercise such as the one provided here, or a ray-tracing simulation with important reflections included, gives significantly more information to system designers than predictions of inter-zone contrast alone, at the cost of increased computational or measurement complexity. Sound field mapping can also provide insight into the design of systems that simultaneously emit masking and speech signals, as will be discussed in Chapter 7, as the maps can quantify the effects of aliasing on leakage from one zone into the other. This can then be linked to the speech intelligibility contrast that is achievable using a certain array geometry or when operating in a particular playback environment.

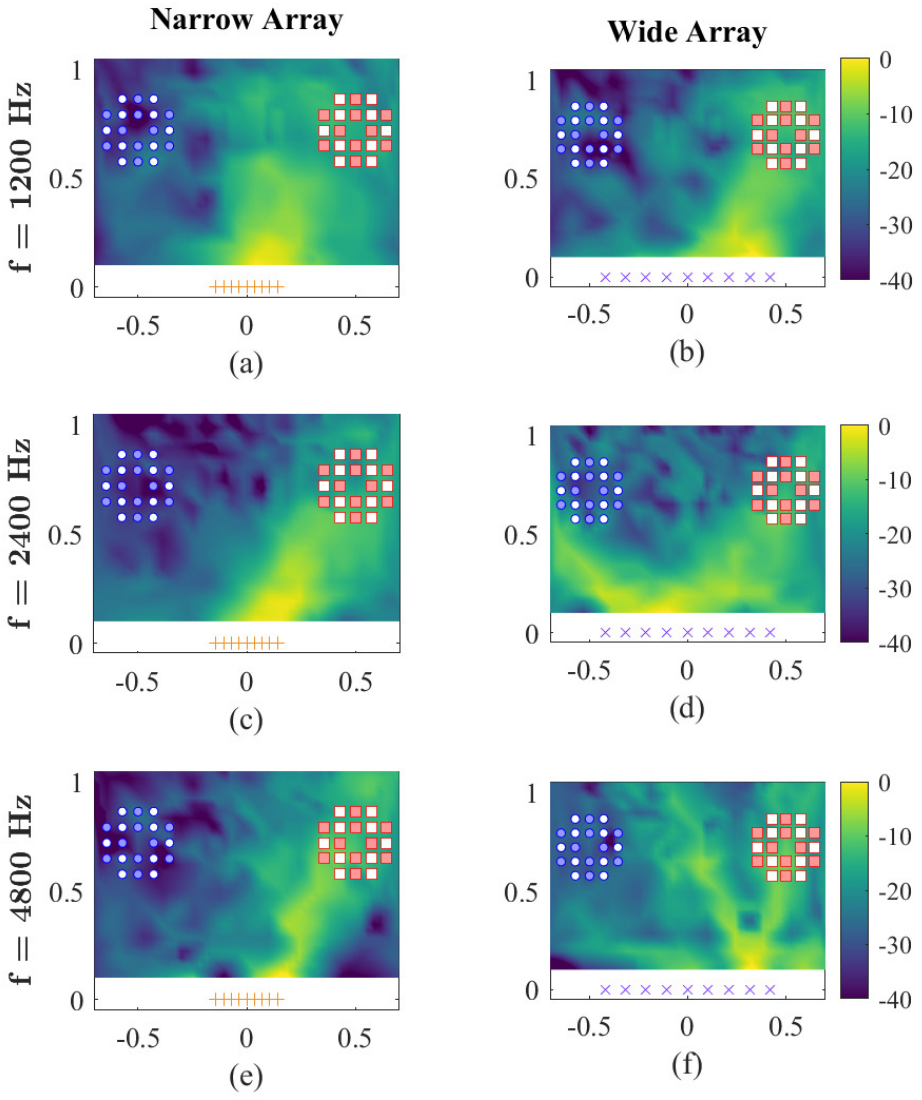


FIGURE 5.19: Relative SPL with tonal signals focussed into the bright zone (square markers) at 1.2, 2.4 and 4.8 kHz for each source array configuration. Dimensions are in metres. Colour scale = dB re. maximum SPL in each map.

5.5 Summary

The loudspeaker array presented in this chapter has been configured to produce a pair of sound zones in a well-damped listening room. This has been achieved using regularised ACC, based on measured transfer responses. The effect of changing the relative positions of the zones has been explored, and higher levels of contrast are demonstrated, particularly at low frequencies, when the zones are spaced further apart from one another. This phenomenon is explained by visualising the beams of sound which are produced from the centre of the array towards each zone, and this beamforming process is effective when the zones are placed within the direct field of the loudspeaker array. The effects of adjusting the geometry of the array have also been

explored by selecting a pair of sub-arrays, each with 9 source elements, with different inter-element spacing and overall aperture. Both array configurations achieve 15-20 dB of acoustic contrast between 300 Hz and 2 kHz, and in the considered symmetrical zonal configuration, this corresponds to a 30-40 dB difference in inter-zonal SNR if two complementary sound zoning processes are used. Comparison of these values with the corresponding speech intelligibility metrics in Figure 3.6 indicates the potential for very high levels of speech intelligibility contrast, but this cursory analysis assumes the provision of acoustic contrast over the full audio bandwidth, which cannot be achieved in practice. With a fixed number of loudspeakers, increasing the array aperture necessitates an increase in the inter-element spacing, resulting in a fundamental design trade-off. Better low-frequency performance is afforded by a wider aperture, but a greater separation between the array elements leads to a reduction in the spatial aliasing frequency and a loss of acoustic contrast at higher frequencies. To investigate the extent to which these limitations impede the provision of inter-zone privacy, the following chapter expands upon the analysis of these two sub-arrays through speech intelligibility evaluation and preference testing. The quantitative results presented in this chapter are used to help explain the results of the perceptual studies.

Chapter 6

Speech Intelligibility and Subjective Preference Testing

In Chapter 5, the physical performance limitations of personal audio systems were described in terms of the acoustic contrast. Two loudspeaker array designs with different spatial aliasing characteristics were tested, and a single ACC process was used to focus signals into the bright zone whilst minimising radiation into the dark zone. Acoustic contrast levels from 10 to 20 dB across the speech frequency range were achieved with a 9-channel loudspeaker array, but the perceptual relevance of this leakage has not yet been described. It is this perceived performance that is most important to consider in systems intended for the control of speech privacy. A target listener's privacy is compromised when speech focussed into the bright zone remains intelligible in the dark zone, so to prevent this, a secondary zoning process that focuses a masking signal into the dark zone has been proposed. The present chapter describes a pair of listening tests that have been designed to investigate the objective and subjective performance of speech privacy control systems which use this method. An objective speech intelligibility test is used to determine the relationship between the bandwidth of the masker and the level at which it must be reproduced in order to provide sufficient privacy, and a subjective test is used to select the masker bandwidth that is preferred in the dark zone. Firstly, a brief review of speech intelligibility and subjective preference testing is provided, followed by details of the experimental designs. Key results from the tests are then presented, and are used in Chapter 7 to produce a series of general design rules for speech privacy control systems.

6.1 Literature Review

While the objective and subjective metrics described in Chapters 3 and 4 can provide valuable indications of performance, a more complete evaluation can be obtained through the use of jury testing. The following subsections provide a brief review of tests used for the objective evaluation of speech intelligibility and the subjective evaluation of listener preferences.

6.1.1 Speech Intelligibility Testing

The body of literature surrounding speech intelligibility testing is unsurprisingly centred around audiology practice and hearing aid research. Understanding speech in noisy environments is frequently cited as the main problem encountered by hearing-impaired people [191]. Accordingly, current guidance from the British Society of Audiology recommends that in addition to pure-tone audiometry, full-sentence speech tests should be included when attending to patients [192]. A number of speech tests have been developed over the last 50 years to capture the intelligibility of words either in isolation [113, 114], in the context of carrier phrases [112, 193–196] or in valid sentences [102, 116, 117, 197, 198]. A common result provided by many of these tests is the Speech Reception Threshold (SRT). This is defined as the SNR at which a 50% score is recorded in the test. In tests that aim to predict the SRT, improvements have been made in the following areas:

- **Efficiency:** Listening test efficiency is improved by reducing the number of stimuli presented, or length of time required, before the SRT can be reliably estimated [199]. In sentence tests, listener responses can be scored based on how many keywords were successfully understood. By increasing the number of statistically independent scorable items in each presented sentence, the test efficiency can be increased [200]. Conversational, predictable sentences, or those with a known context can be thought of as containing redundant information; if part of a sentence is misheard, it may be possible to reconstruct the missing word or words from those that were understood [201] - for example, the word "knife" in the sentence "They slice the sausage thin with a knife" [a] is not statistically independent from the other words in the sentence, as it is eminently guessable. Grammatically correct sentences with unpredictable words do not suffer from this phenomenon, and so can provide faster convergence to the SRT, whilst maintaining face validity, i.e. "whether the test 'looks valid' to the examinees who take it" [202].
- **Validity:** Tests where the speech material consists of single words or short groups of digits do not accurately represent real-world speech, as co-articulation artefacts between words are omitted [203]. Sentence-length stimuli provide a face-valid presentation that is short enough to incorporate into a practical test. Longer passages may result in the appraisal of working memory performance, rather than speech recognition, and subjects may require longer training to be able to repeat running speech accurately [23].
- **Reliability:** A test can be considered reliable if there is a small variation between repeated applications of the test to the same subject. This means that training, priming and fatigue effects of the test must be understood and minimised, and if stimuli are selected randomly from a bank of recordings, that intelligibility is equalised between them [115, 204].

As private personal audio system design is a novel application, a specific test for this particular application has not been developed. The variety of speech stimuli and typical maskers used in the tests referenced above is evidence that a "one size fits all" strategy for intelligibility testing is unachievable - in order to provide a reliable and valid intelligibility estimate for a particular application, tests must be designed specifically to provide this information. For example, the

[a] Taken from the Harvard Sentence Corpus [100], List 45, Sentence 1

Coordinate Response Measure [193] has been used to test auditory fitness for duty in military personnel [205], as the stimuli used in this test resembles the instructions that service personnel might be required to interpret. Similarly, the Automated Toy Test [195] was designed with the accessibility needs of young children in mind. Tests have also been designed to reduce the burden on experimenters or clinicians by allowing multiple choice responses or by including interfaces that enable the rapid scoring of tests. Examples include matrix sentence tests, in which listeners respond by selecting words from a "matrix" of options [102, 206] and the Quick SIN test [198], which can provide clinically useful information on SNR loss in hearing-impaired listeners after 2-3 minutes of testing.

One problem with conventional sentence tests such as the Bamford-Kowal-Bench test [116] is the requirement for a corpus of individual recordings of each sentence. Besides being a time-consuming process to acquire, involuntary changes in the level of a speaker's voice across words in a sentence can cause certain words to be more or less intelligible. Furthermore, a large number of test sentences must be recorded, to avoid listeners being able to remember sentences from previous tests, e.g. when auditioning different hearing aids. Faced with this difficulty, Hagerman [102] developed a method for cutting individual words from recordings of sentences with a fixed grammatical structure, which could be randomly combined to produce new, realistic-sounding sentences. The approach also allowed the level of individual words to be subtly adjusted to reach a similar intelligibility level. In this initial work, and the subsequent translation of the test from Swedish into German, [207], sentences were presented without giving listeners the full matrix of candidate words to choose from. This is referred to as an "open-set" test. The alternative "closed-set" presentation was considered as an option later in the Danish language DANTALE II test [208], and has been found to yield less significant training effects [209]. The closed-set response format allows the advantages of self-pace and self-completion of the test. A project by Hearcom [210] to produce and standardise a range of equivalent tests in multiple languages is ongoing. A review of matrix tests in 14 different languages is provided by Kollmeier et al. [206].

A British English recording of the word matrix is available [23, 209], and software accompanying this corpus allows the generation of tests with arbitrary background noise, and a range of testing formats, such as the ability to present stimuli at a fixed SNR, or allow the test to adapt the SNR based on the participant's responses. This flexibility, along with the general efficiency and reliability of matrix tests make them an ideal candidate for evaluating the provision of privacy by personal audio systems. By varying the level and bandwidth of stationary, random noise maskers, the results from the matrix tests will indicate the relationship between these parameters and the intelligibility of speech. Masking signals which provide an adequate level of privacy, when reproduced at the correct SNR, can then be compared by listeners in a subjective listening test.

6.1.2 Perceptual Testing

A key reference for the design and execution of perceptual testing is Bech and Zacharov's "Perceptual Audio Evaluation - Theory, Method and Application" [211]. This source recommends a process for designing and conducting perceptual tests, the key steps of which will be used to structure this review.

Firstly, it is recognised that perceptual testing can be complex and nuanced, so repeating tests that have already produced conclusive results is wasteful. Therefore, a detailed literature review in the specific area of interest must be carried out. However, speech privacy control is a developing field and to the best of the author's knowledge, there have not been any subjective studies that directly evaluate the privacy in sound zones or the perceptual consequences of its provision. As described in Section 2.3, an implicit link is often assumed between inter-zone acoustic contrast and privacy [15, 64, 92–94], and in other studies, objective intelligibility metrics are used as a proxy for privacy [9, 97]. In both of these referenced studies, additional masking noise is introduced into the environment as an integral part of the privacy control method, but the perceptual effects of this additional noise are ignored. The aim of the perceptual testing presented in this thesis is to take a set of masking signals that have already been verified for their capability to provide privacy, and assess which of these is preferred. This recognises that the primary function of a speech privacy control system is to provide private listening zones.

Widening the focus away from sound zoning systems, significantly more perceptual tests have been carried out to evaluate the acceptability of open plan office sound masking systems. One prolific author in this area is Valtteri Hongisto, who has contributed a number of studies that are pertinent to the design of masking signals for personal audio systems. Apposite examples include the subjective and objective rating of spectrally different pseudorandom noises [90], the generation of new metrics and important attributes for modelling noise annoyance [212], and the appraisal of different types of masking sound, such as water, music or ventilation noise [85, 213]. The body of literature has converged on some rules of thumb that have been adopted as general best practice for open plan office sound masking design, e.g. a spectral slope of -5 dB per octave and a masker level of less than 45 dBA [84, 88, 91, 214, 215]. Less consensus has been reached regarding adjustment of the masking level to a schedule, or adaptively based on the occupation of a space. Proponents of the technology state that in offices, the need for privacy and freedom from distraction varies throughout the working day, with social interaction being beneficial at the start and end of the day, and more focussed time being provided in the middle of the day. With adaptive or programmed masking signals, this can be facilitated more easily than with a fixed sound masking system [216]. However, it has been argued that masker adaptation and scheduling could be self-defeating [217]. At quiet times, when the masking level would be reduced by the system, privacy requirements actually increase as individual voices are more intelligible at this time [218]. Furthermore, the temporal and spatial resolution offered by commercial adaptive masking systems is often inadequate for the types of distractions and privacy concerns that exist in open plan offices, such as when people talk whilst walking across an office [217]. The ASTM Standard Guide for Office Acoustics [101] remains ambivalent on the topic, stating that masking sound generators should include a means to adjust the equalisation and level of the masker, without explicitly recommending continuous adjustment of these parameters during operation.

Despite the wealth of recommendations made in the literature concerning office sound masking systems, these are potentially only of limited use for the speech privacy control problem. Firstly, the requirements of sound masking systems are subtly different to those of speech privacy control. The low level of masking that is considered adequate for open plan offices would in many circumstances not be sufficient to render speech sufficiently unintelligible to claim privacy. Additionally, recent studies of sound masking systems have found that the widely accepted rules of thumb for sound masking design do not necessarily result in significant improvements to the experience

of office workers in the long-term [91, 213]. Given this apparent failure, and the accompanying recommendations to carry out a holistic analysis of the requirements of each space, it is evident that perceptual evaluation is required for the novel application proposed in this thesis.

Bech and Zacharov's next recommendation for the preparation of subjective tests is to ensure that the magnitude of perceptual differences is neither too large to merit interesting conclusions, or too small to be evaluated in a reasonable length of time [211]. As evidenced from the large range of candidate masking signals used in the office sound masking tests described above, signals in the proposed perceptual tests are likely to be sufficiently distinct for testing to be meaningful, even given the restriction of providing similar levels of speech privacy. Initial pilot experiments will be used to confirm the perceptual distinctiveness between the candidate masking signals identified by the intelligibility tests.

Once the necessity of conducting listening tests has been established, attention can be turned to the specifics of the test itself. Two critical factors are the response attribute and response format. The response attribute is the specific element or feature that test participants are asked to evaluate, and the response format describes the way this evaluation is carried out. In the proposed tests, there is potential for the presented masking sounds to elicit a diverse range of responses in each subject. Some response attributes, such as loudness and sharpness, are likely to be experienced similarly by each participant. Where this is the case, established subjective metrics can take the place of listening test evaluations. In other words, asking participants to report on the loudness of each sample in a test is unlikely to provide ground-breaking results. Other impressions are unavoidably specific to each listener; for example, a particular sample may remind a listener of the sound that their car's air-conditioning system makes. These types of responses are unlikely to offer information pertinent to the general problem that the listening test is designed to solve. The response attribute should therefore be carefully specified such that it falls between these two extremes; it must be neither too universal that unsurprising, unanimous results are obtained from participants, nor too particular to each participant that any underlying trends are impossible to extract. This process of selecting an appropriate response attribute must be combined with the choice of an appropriate response format. The response format must transparently, i.e. without bias, guide the test participants into providing the experimenter with useful information. Table 6.1 presents a small selection of possible response formats, alongside some key advantages and disadvantages of each method.

Response Format	Advantages	Disadvantages
Ordinal Rating Scales, e.g. ITU-T P800 [219]	Response format is simple and widely used; familiar to participants.	Subject to several biases, such as avoidance of extremes, coarse granularity of responses
Continuous Quality Scales e.g. ITU-R BS. 1284-1 [220]	Format provides a way to rank the stimuli in order, and assess the perceived magnitude of the differences directly.	If labelling is used, these must be perceived to be equidistant on the scale to avoid bias.
Multiple Stimulus, Hidden Reference with Anchor (MUSHRA) e.g. ITU-R BS. 1534-1 [221]	Hidden reference and anchor stimuli reduce the effect of biases that are found in standard continuous rating scales.	Can be time-consuming to complete as multiple passes through a set of comparisons is often required.
Paired Comparison Testing [222]	Response format is simple and intuitive for participants.	Testing can be time consuming if multiple stimuli must be compared.
Free-elicitation e.g. [3]	Intuitive for participants to respond.	Significant training or expert listeners required for useful results. Post-processing of results is complex.
Non-verbal elicitation, e.g. drawing, pointing [223]	Possible for participants to express their perception of multiple response attributes in a single modality. Particularly suited to spatial audio testing.	Training and familiarisation with the response format is necessary. Post processing is more complex than with numerical rating scales.

Table 6.1: Table describing a selection of response formats for the perceptual evaluation of audio stimuli.

Given that the objective of the perceptual testing here is to ascertain which of the candidate masking signals identified in the speech intelligibility test is preferred, it is counter-productive to request responses on a rating scale, such as ITU-T P800 [219] or BS 1284-1 [220] for a specific attribute such as annoyance. This type of response format would unnecessarily restrict listeners to considering how annoying each signal appears to them. While there is an expected correspondence between annoyance and preference, the impression of annoyance is by no means the only factor that could contribute to the preference of one noise over another. Forcing participants to use an inappropriate or confusing rating scale may amplify the biases that are inherent in this response method, as described in Table 6.1. Some of the effects of response format bias can be countered through the use of hidden references and anchor stimuli [221], but this does not solve the problem of requiring listeners to quantify a complex emotional response such as annoyance on a simple rating scale. Furthermore, each participant may have a subtly different internal definition and threshold of annoyance, which would complicate the interpretation of results. A

paired comparison test allows the attribute of interest, i.e. preference, to be ascertained directly, and randomisation and repetition of paired comparisons can be used to test respondents' repeatability and consistency [222]. One drawback of preference testing is that the underlying reasons for the reported preferences are not gathered. To gather these, a free elicitation exercise can also be carried out, and the results from this test can then be compared against preference test results to extract the features that are most important to control when designing a masking signal.

6.2 Test Design

Given the considerations described in the previous section, the design of the listening tests will now be described in full. Six array-masker configurations are considered in total, corresponding to the two array widths described in Section 5.4 and three cut-off frequencies for a low-pass filter applied to the masker. These conditions will be referred to using the code $[W|N]_f$, for the wide and narrow arrays respectively. The first pair of conditions, W_A and N_A, refers to the case where the cut-off frequency is set to the point where a spatially aliased side-lobe begins to impinge on the opposite sound zone, as illustrated for the wide array in Figure 5.19d. This approach to setting the low-pass filter cut-off frequency was proposed by Donley et al. [9] and is designed to eliminate the leakage of the masking signal into the bright zone above the aliasing frequency, therefore preserving the intelligibility of the speech signal for the target listener. With this approach, the required cut-off frequency is 2.4 kHz for W_A and 8 kHz for N_A. For the second pair of conditions, W_S and N_S, the cut-off frequency is set to 4 kHz for both arrays, in order to filter out frequencies that contribute strongly to the sensation of sharpness [79]. The final pair of conditions, W_∞ and N_∞, refer to the case where the low-pass filter is bypassed, such that the upper frequency is given by the anti-aliasing filter at approximately 20 kHz. A sample test battery containing each of these conditions is presented in Table 6.2. The speech intelligibility and preference tests were interleaved to reduce the impact of listener fatigue, and a break was offered between tests that used the narrow and wide arrays. Perceptual comparisons between the narrow and wide arrays were not carried out as it is expected that for a given use-case, the choice of the overall array size will be principally governed by space limitations and other practical constraints. Furthermore, the primary objective of the paired preference testing was to provide information regarding the design of the masking signal, rather than the loudspeaker array producing it. The positions of the wide and narrow array elements, and the location and size of the zones are as described in Section 5.4. Diagrams of the loudspeaker array and zone geometry are reproduced for convenience in Figures 6.1 and 6.2.

#	Type	Array Width	LPF Cut-off (Hz)	Code	
1	Sentence Test	Wide	Full Bandwidth	Training	
2	Sentence Test	Wide	Full Bandwidth	W_∞	Random Order
3	Sentence Test	Wide	2400	W_A	
4	Sentence Test	Wide	4000	W_S	
5	Preference Test	Wide	All from #2,3,4	PrefWide	
6	Sentence Test	Narrow	Full Bandwidth	N_∞	Random Order
7	Sentence Test	Narrow	4000	N_S	
8	Sentence Test	Narrow	8000	N_A	
9	Preference Test	Narrow	All from #6,7,8	PrefNarrow	

TABLE 6.2: Order of tests carried out by each participant.

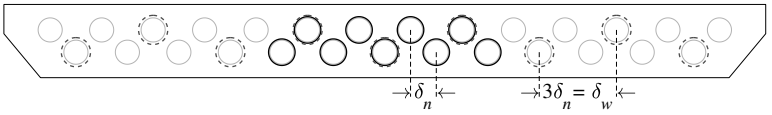


FIGURE 6.1: Diagram of loudspeaker array elements selected from a 27-channel array (described fully in Ref. [180]), to form two 9-channel arrays. The narrow array elements, marked with solid circles, have a horizontal spacing $\delta_n = 35$ mm and the wide array elements, marked with dashed circles, have a horizontal spacing $\delta_w = 3\delta_n = 105$ mm. The vertical spacing between elements is 30.40 mm.

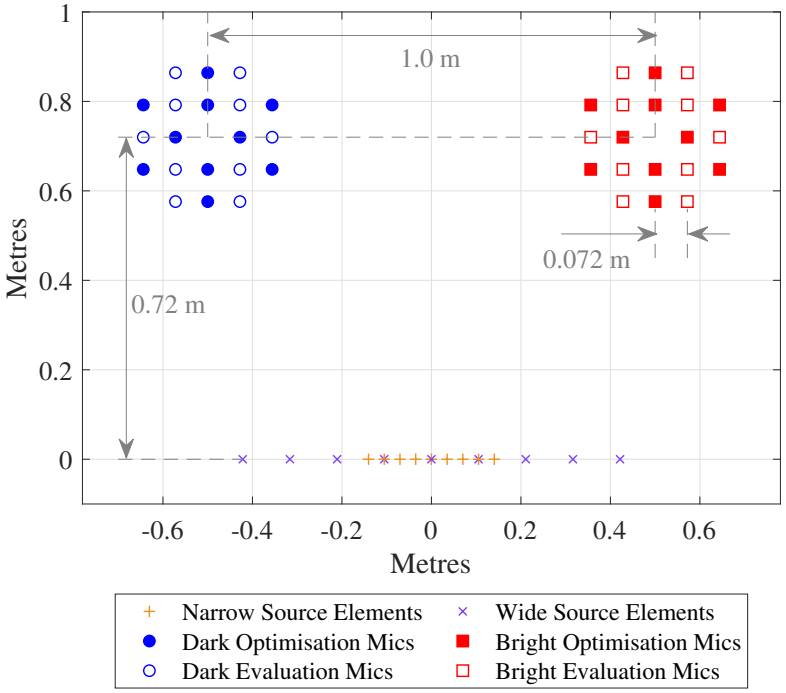
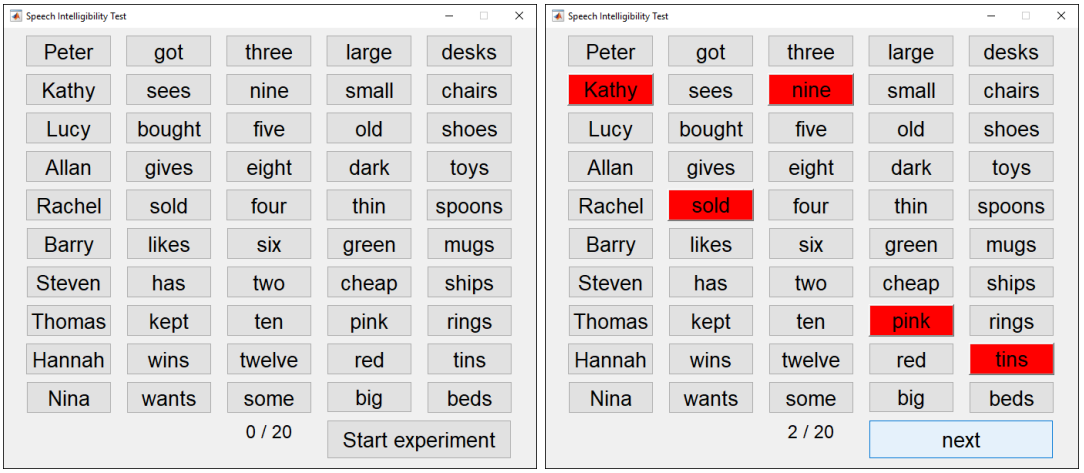


FIGURE 6.2: Plan view of the personal audio system geometry, showing source and microphone locations.

6.2.1 Sentence Test Design

The speech intelligibility test is derived from the English matrix test [23, 209]. Participants listen to five-word sentences in noise with a fixed grammatical structure: "Name verb numeral adjective object", e.g. "Kathy sold nine pink tins", then select the words they heard using a screen-based interface. Figure 6.3 shows the 10×5 matrix of selectable words.



(A) Start of test. (B) During test, after sentence presentation.

FIGURE 6.3: User interface for sentence test. Participants can see the word matrix at all times during the test. After a sentence has been presented, participants select as many words as were heard, before continuing on to the next sentence. No feedback regarding the correctness of responses is given to the participant during the test or afterwards.

Ten options are available for each word in the sentence, forming a matrix of 50 words. The database of individual word recordings was generated by recording all possible pairs of words that could appear in the sentence test e.g. "Peter got", "Peter sees", "Peter bought" ,... "got three", "got nine", and then manually cutting out the first word from these pairs to preserve any coarticulation artefacts with the next word in the sentence. This yields ten copies of each word in the first four columns of the matrix, and one copy of the last column. When randomly generated test sentences are procedurally generated from the database of individual word recordings, the correct version of each word is selected based on the next word. The resulting synthesised sentences have a high degree of realism but are lexically unpredictable, meaning that learning effects are reduced compared to everyday sentences used in other speech tests [208]. The sample rate of the recordings is 44.1 kHz and the bit depth is 16 bits per sample. For more information on the recording process, see Refs. [23, 102]. Speech-shaped masking noise was generated from the database of recordings by concatenating the entire database of words, converting to the frequency domain using the Discrete Fourier Transform, randomising the phase of the Fourier coefficients, then converting back to the time domain using the inverse DFT.

In order to present participants with the experience of the dark zone of a personal audio system, the sentences and noise are processed through a loudspeaker array simulation based on binaural room impulse responses measured in the ISVR audio laboratory, from the 27-channel loudspeaker array to a KEMAR mannequin turned towards the centre of the array, as described in Section 5.1.2. In this simulation, the sentence stimulus is focussed into the bright zone of the system and the speech-shaped masker is focussed into the dark zone. Therefore, the listener hears a

combination of the direct masking signal and the leakage of the speech signal into the dark zone. The SNR begins at 0 dB and is adjusted each time the participant responds, according to the percentage of correct words relative to a target score, usually set at 50%. As the test progresses, the size of the adjustments to the SNR are reduced, so that by the 20^{th} sentence, the SNR has converged to a level representative of 50% words correct [200]. This value is defined as the SRT for the matrix test.

Firstly, to familiarise each participant with the sentence test procedure, a training set of 20 sentences is presented. In other studies with matrix tests of various languages, training effects have been observed from within the first two tests [209] to beyond the sixth test [208]. However, these tests were carried out in an open-set format, i.e. listeners did not have sight the matrix of words, and rather responded by repeating the received sentence aloud. Tests carried out using a closed-set format, similar to that used in this test, found fewer training effects [209], but still recommended two training tests be carried out, one at a high, fixed SNR to familiarise listeners with the speaker's voice, and one adaptive test, to expose participants to a range of sentence difficulties. In this case, these two training tests were combined into a single adaptive test. The adaptive SNR procedure in the training set aims for a sentence intelligibility rate of 70%, so that the participant gains more experience with the voice of the speaker than with the standard aim point of 50%, which is used for the remainder of the sentence tests.

Next, three further sentence tests using the wide array are presented, using the three low-pass filter settings, to acquire an SRT for each. The presentation order of these tests is randomised between participants to reduce the impact of any residual training effects across all participants, as described in Table 6.2. The matrix test format induces less listener fatigue because the closed-set presentation slightly eases word understanding; SRTs are approximately 1 dB higher in open-set presentations of the same test [206]. Pilot testing confirmed that at the SRT of the closed-set test, participants reported that speech could be considered private, and that without access to the word list, understanding would be significantly impeded. This is consistent with the published slope of the reference psychometric function for the test of 13%/dB SNR at the SRT [206], i.e. a 1 dB change in SNR results in a change of 13% in the intelligibility score.

6.2.2 Preference Test

The final stimuli from each of the three sentence tests for each array width are expected to have an intelligibility of approximately 50%, due to the adaptive SNR procedure. These three stimuli are aggregated into a preference test by repeating all three possible pairs of combinations four times, giving 12 trials in total, as illustrated in Figure 6.4. For each array width, the level of each stimulus is adjusted to a standard SNR so that each participant makes comparisons that are comparable to one another. These SNRs are given in Table 6.3, and were established during pilot tests to be representative of SRTs. Participants were instructed to use the on-screen interface shown in Figure 6.5 to play stimuli A and B, then, paying attention only to the noise in each stimulus, select which they prefer. No additional contextual information or alternative definitions of preference were provided to participants, in order to maintain identical conditions between participants and to eliminate bias towards preferences that would be specific to a particular use-case. Participants were able to audition the stimuli any number of times before submitting their decision, but most participants completed all 12 comparisons in under five minutes.

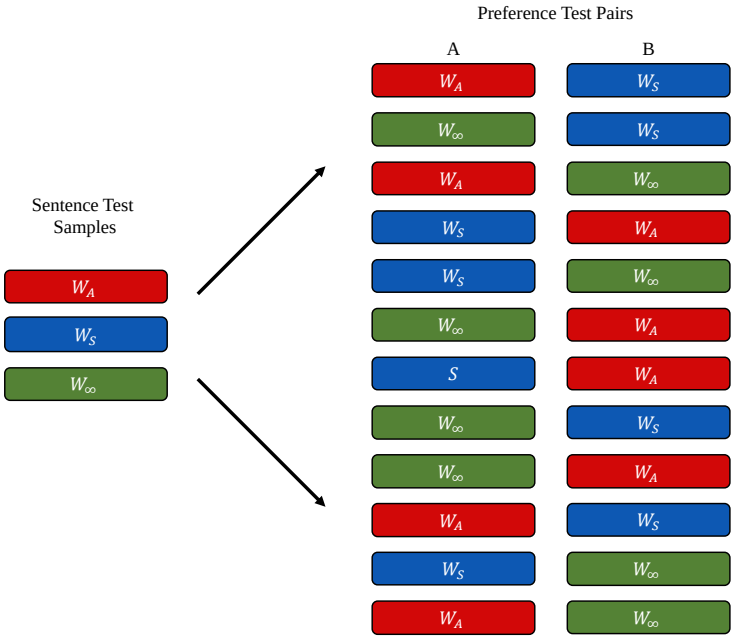
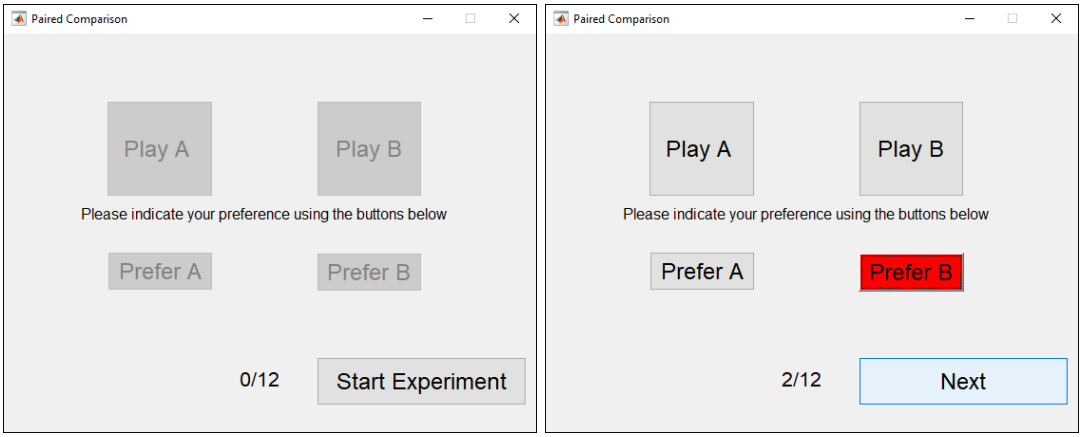


FIGURE 6.4: The last presented stimuli from each sentence test at a given array width are adjusted to a fixed SNR, then presented in pairs to listeners.



(A) Start of test. (B) During test.

FIGURE 6.5: User interface for preference test. Participants are forced to make a choice between stimulus A and B before being allowed to continue.

Code	SNR (dB)
W_S	-21.5
W_A	-27.6
W_∞	-16.1
N_S	-15.4
N_A	-14.2
N_∞	-13.8

TABLE 6.3: SNRs used during the preference test.

6.3 Results

22 students and researchers from across the University of Southampton were invited to participate in the test. The mean age of the participants was 26.1 years (σ = 3.7 years). All were over 18 years of age and reported that they had normal hearing and were fluent in the English language. Some participants had experience of participating in other speech intelligibility tests, but none had any formal technical listening training. Data from native and non-native English speakers is included in the data analysis and a comparison of the SRTs for these two groups is provided in Appendix B. The average test duration, including optional breaks between test sections was 41 minutes (σ = 6.1 minutes). Ethical approval for the research was obtained from the University of Southampton (Reference ERGO/FEPS/50192).

6.3.1 Sentence Test

The results of the sentence test are SRTs in dB, and are presented as a series of box plots for each configuration in Figure 6.6. From these results it can be seen that the SRTs for the wide array are significantly lower than for the narrow array, i.e. the level of the masking signal must be increased to achieve the same level of (un)intelligibility. The wide array has poor high-frequency control due to spatial aliasing, so a significant amount of high frequency speech information is leaked into the dark zone. This necessitates an increase in the masking signal level compared to the narrow array, which has more consistent levels of ACC in the speech frequency range, as shown in Figure 5.18. This is supported by the finding that the median SRT for condition N_∞ is very close to the reference SRT [209] in steady, speech-shaped noise found for the closed-set English matrix test from which this test was derived (see black dashed line in Figure 6.6 at -8.9 dB SNR). Any adjustment to bandwidth, via array processing or low-pass filtering of the masker, reduces the SRT.

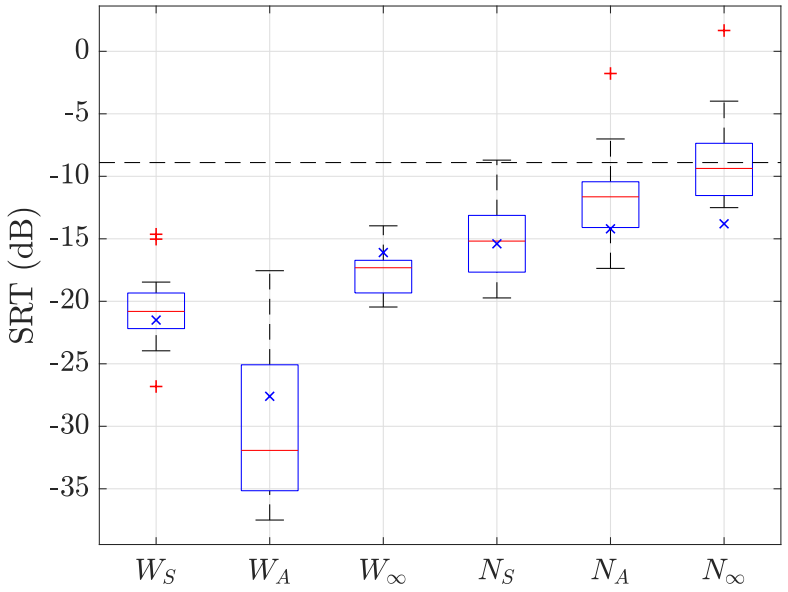


FIGURE 6.6: Distribution of Speech Reception Thresholds (SRTs) achieved for each array configuration. N=21 participants. Blue crosses indicate the SNRs presented to listeners in the preference test, determined during pilot tests. Red plusses are outliers, identified as all those results which lie greater than 1.5 times the box length from the edges of the box (approximately $\pm 2.7\sigma$). The reference SRT for the closed-set English matrix test [209] is represented with a dashed horizontal line.

The variability of the acoustic contrast level across frequency is also responsible for the large range of SRTs recorded for condition W_A, which can also be seen in the results presented in Figure 6.6. The low-pass filtered masker in this condition cannot adequately mask consonant sounds, giving listeners increased opportunity to correctly guess words from the provided matrix of sentences. This enlarges the inter-subject variability beyond that usually expected from the normal-hearing population.

6.3.2 Preference Test

The results of the paired preference tests described in Section 6.2.2 have been analysed using statistical tools developed by Perez-Ortiz and Mantiuk [222]. These tools analyse patterns of preference across participants to enable outliers to be removed, indicating potential misunderstandings of the task. One participant's data was removed from the analysis as their preferences were consistently misaligned with those of the rest of the group. Due to the interaction and sharing of stimuli between the sentence and preference tests, this participant's data was also excluded from the analysis of the sentence tests. The remaining raw paired comparison results are converted into ratings for each condition, using a JND scale. On this rating scale, the distances between conditions are related to the probability of members of the tested population preferring one condition over another. The scale is normalised such that a distance of 1 corresponds to 75% of comparisons favouring one condition over another; a distance of 2 corresponds to 91%. In general, distances in JND-space, δ, can be converted to a percentage likelihood of a preference being indicated using the formula

Condition A	vs.		Condition B
W_S	97.4%	2.6%	W_A
W_S	11.8%	88.2%	W_∞
W_A	0.4%	99.6%	W_∞
N_S	22.3%	77.7%	N_A
N_S	22.7%	77.3%	N_∞
N_A	50.5%	49.5%	N_∞

TABLE 6.4: Percentage likelihood of one condition being preferred over another, using data gathered from the preference tests.

$$\% \text{ Preferred} = \Phi_{\mu=0,\sigma=1.4826}(\delta), \tag{6.1}$$

where Φ is the cumulative normal distribution function, with mean $\mu = 0$ and standard deviation $\sigma = 1.4826$. The results of this conversion are presented in Table 6.4 for each of the comparisons made during the preference tests.

From the results presented in Table 6.4, the preference test results for the wide array were conclusive, with participants demonstrating a clear preference in all three paired comparisons. The least preferred condition was W_A, where the low-pass filter is set at 2.4 kHz to prevent aliasing. Figure 6.6 shows that the median SRT at this condition is -32 dB, at least 10 dB lower than the other two conditions in the test. The corresponding increase in the masker level was clearly perceivable, and was disliked by participants. In the comparison between W_S and W_∞, a significant proportion of participants selected W_∞, the case where the masker has broader bandwidth and a lower overall level.

Preferences were not as well-defined in the case of the narrow array, suggesting that the signals in each condition were perceptually more similar than those with the wide array. Applying the low-pass filter to reduce sharpness (N_S) was disliked when compared with both the unfiltered condition N_∞ and against N_A, where the low-pass filter cut-off was set to 8 kHz to prevent spatial aliasing. This can again be related to the increased masker level required when the cut-off frequency of the low-pass filter is reduced to 4 kHz. No significant preference was shown between the N_∞ and N_A conditions when preferences are aggregated across participants, but more than half (11 out of 21) listeners consistently chose their preferred condition across all four repeats of the N_A vs N_∞ test. Only three participants selected each condition twice (equivalent to chance), indicating that for most listeners, it was possible to distinguish the samples and make a repeatable preference judgement.

Directly after the completion of each preference test, participants were asked to give, in their own words, any reasons for their judgements or features of the noise samples which they were listening for. Despite the apparent correlation between masking signal level and preference judgement, only 8 out of 21 participants mentioned loudness or quietness in their descriptions. 17 participants distinguished between sounds by referring to their spectrum, using words such as “high/low pitched”, “sharpness” or “harshness”. 10 participants used words referring to naturalness or artificiality, e.g. “smooth”, “natural”, “sounds like a jet engine / waterfall / the London Underground”. When such a preference was expressed, participants unanimously preferred sounds they deemed to be “natural” over those which were “artificial”. When correlated

with the participants' individual choices in the preference test, sounds with broader bandwidth were deemed to sound more "natural".

Six participants commented, after the narrow array preference test, that although masking signals with wider bandwidth were preferred in general, having too much high frequency content was detrimental. Of these six, five preferred the condition N_A over N_∞, backing up their comments with their preference decisions. Their comments are transcribed below:

- "Sharper sounds were better, but not too sharp."
- "Highest one was unpleasant."
- "Didn't like higher pitched."
- "The second highest pitch was preferred."
- "I preferred less high frequencies, I would avoid low-frequencies [and prefer a wider bandwidth masker] unless there was noticeable additional hiss, this hiss was worse than the low-frequency sound."
- "My preference was for not too sharp, or too loud."

These results show that for certain listeners, applying modest low-pass filtering can improve the perception of masking signals. Although participants only listened to filtered random noise samples, a surprising breadth of semantic descriptions were attached to these sounds. This encourages future experimentation on the acceptability of different types of masking signals in different contexts.

After participants were asked to provide reasons for their preferences, some also gave other comments about the test which may provide insight into the reasoning behind their judgements or their feelings about the test procedure. These comments were provided without prompting, so it is impossible to know if other participants had the same issues but did not raise them. Four participants, two with English as a first language, and two with English as an additional language, mentioned that they felt their performance was affected by working memory limitations rather than mishearing the sentences. These listeners reported that that were able to focus on a few words in the sentence, but may forget the full sentence whilst searching for words in the matrix. However, the advantages of unpredictable sentences in terms of the reduced training effect, and greater number of statistically independent elements to perceive per sentence [200], outweighs the disadvantage of the sentences being potentially more difficult to remember.

Three participants with English as an additional language reported that the test was tiring, and requested breaks after the first half of the test (W_A, W_S, W_∞ and PrefWide). This fatigue effect may account for the higher speech reception thresholds found in those with English as an additional language, as seen in Figure B.1. An advantage of the matrix test format for consideration in future work is that participants would be able to complete the test in their first language, without the experimenter needing to be familiar in that language [206].

6.4 Summary

A two-stage listening test has been carried out to obtain speech reception thresholds and preference information for six array-masker configurations. Reducing the cut-off frequency of the low-pass filter applied to the masker causes an increase in the required masking signal level to ensure privacy. Quieter masking signals, with wider bandwidth, were in general preferred over louder, low-pass filtered maskers. However, arbitrarily reducing the loudness of a given masking signal also reduces its effectiveness, and hence its ability to provide privacy for the target listener. This trade-off between objectives highlights the necessity of setting intelligibility constraints in each listening zone, as described in Chapter 1.

The listening tests provide useful information that can be used to recommend the characteristics and level of masking signals that should be used, for the two loudspeaker array configurations under test. However, the data from the tests presented in this chapter are specific to these configurations, and are thus limited in scope. This is evidenced by the fact that each array and low-pass filter configuration had a different SRT; the raw results provide no way to generalise these SRTs to alternative system designs. The following chapter provides a means to generalise the results of the listening tests to arbitrary system designs by fitting the listening test results to objective and subjective metrics.

Chapter 7

Masking Signal Design Based on Speech Privacy Constraints and Listener Preferences

Parts of this chapter have been published as "Design and Evaluation of Personal Audio Systems Based on Speech Privacy Constraints", in the Journal of the Acoustical Society of America 147(4):2271-2282

Attention has been paid throughout this thesis to the practicalities of designing personal audio systems to produce private sound zones. This chapter continues with this theme by assembling the elements described in Chapters 3 to 6 into a general methodology for designing the masking signal that is required by such systems. Chapter 3 described various objective metrics for the prediction of speech intelligibility from properties of speech signals, and Chapter 4 contained information on subjective metrics to translate between objective features of signals and various elements of human sound perception. In this chapter, both types of metrics are correlated with the results from the objective and subjective listening tests described in Chapter 6, to provide a set of guidelines for designing the masking signal. The objectives for the design process are:

- The masking signal must provide enough intelligibility reduction that an eavesdropper in the dark zone should not be able to understand the message provided to the target listener in the bright zone.
- Any consequent leakage of the masking signal into the bright zone must not impede the intelligibility of the message for the target listener.
- Whilst satisfying the constraints above, the masking signal must be as acceptable to listeners in the dark zone as possible.

7.1 Comparison Between Listening Test Results and Objective and Subjective Metrics

In the listening tests described in Chapter 6, an auralisation process was used to present binaural stimuli to each participant. Each of these stimuli was saved for further analysis using objective and subjective metrics. The following two subsections describe the comparisons between these metrics and results from each section of the listening tests.

7.1.1 Comparison Between Intelligibility Test Results and SII

In all envisaged applications of speech privacy control, setting the level of the masker correctly is crucial to a system being regarded as acceptable. Adjustment to the masker level affects all three objectives that were detailed in the introduction to this chapter. If the masker level is too low, speech privacy could be compromised, and if it is too high, the intelligibility in the bright zone may be impeded, and nearby listeners could be annoyed by the additional environmental noise. As speech and noise are radiated simultaneously by the proposed system, the required masking signal level in a given system principally depends upon the level of the speech that must be masked. An alternative definition of the problem can remove this dependency on the level of the speech, drawing on the fact that speech intelligibility is largely unaffected by absolute signal levels [107]. Rather, the key problem is finding an appropriate SNR within the dark zone, denoted SNR_d, that corresponds to conditions of sufficient unintelligibility - this removes the dependence on speech level in the following analysis.

The results from the listening test presented in Section 6.3.1 demonstrate that the Speech Reception Threshold, analogous to SNR_d, varies with different arrays and masking signal spectra. The relationship between the broadband SNR_d and the impression of privacy is frequency dependent because the masking ability of a given signal depends on the matching between the reproduced speech spectrum, and that of the masker. This frequency dependence is taken into account by the SII metric. As described in Section 3.1, SII aggregates narrowband SNRs into a single-number value that corresponds to a given level of speech intelligibility. This single-number value is therefore independent of the absolute speech level *and* any frequency dependent effects, such as the choice of masking signal spectrum and the frequency responses of the loudspeaker array and sound zoning filters. The SII metric can therefore be used to set universal limits, or targets, on the speech intelligibility within each zone, provided that the conditions in each zone are within the scope of the metric [124].

The standard that defines the SII states that good communication systems are characterised by SII values greater than 0.75 [124]. This value is selected as a lower SII limit for the reproduced bright zone sound field; $\mathrm{SII}_b > 0.75$. A similar SII limit that corresponds to privacy, or sufficient unintelligibility is not found in the same standard, but this value, denoted SII_d, can be determined from the speech intelligibility tests described in Chapter 6. Once both the bright and dark zone SII limits are found, these can then be applied to new system designs by reversing the process of abstraction that has been described above. Using measurements or simulations of the sound zones produced by the new design with a fixed, representative speech level in the bright zone,

the masking signal level should be adjusted so that the SII in the dark zone is below the limit given by SII_d and the SII in the bright zone is greater than $\text{SII}_b > 0.75$.

The process for determining SII_d is as follows. In the speech intelligibility tests described in Chapter 6, each stimulus that was presented to participants during the matrix test was also passed through the SII algorithm, providing an objective rating of intelligibility that could be compared with the listening test score for that sentence. Each participant listened to a set of 20 sentences in each test condition. For each of these sets, the percentage of correct words identified in each stimulus was fitted to the corresponding SII using a sigmoid function. Each curve of SII values against intelligibility was then interpolated at the 50% words correct level to give a single-number value representative of the SII for 50% words correct, SII_{50}. Four examples from different subjects and test conditions are provided in Figure 7.1, and the corresponding SII_{50} value is marked on the x-axis. When the SII_{50} values are averaged across all participants and conditions, a score of 50% words correct in the matrix test corresponds with an SII value of 0.05. This value is assigned as the dark zone intelligibility limit; $\text{SII}_d < 0.05$. Comparisons with the published privacy indices presented in Table 3.1 indicates that this value exceeds the required level for "Confidential Privacy" in open plan spaces, of SII < 0.10, and provides comparable conditions to the "Minimal Speech Privacy" offered between closed rooms (SII $= 0.03$).

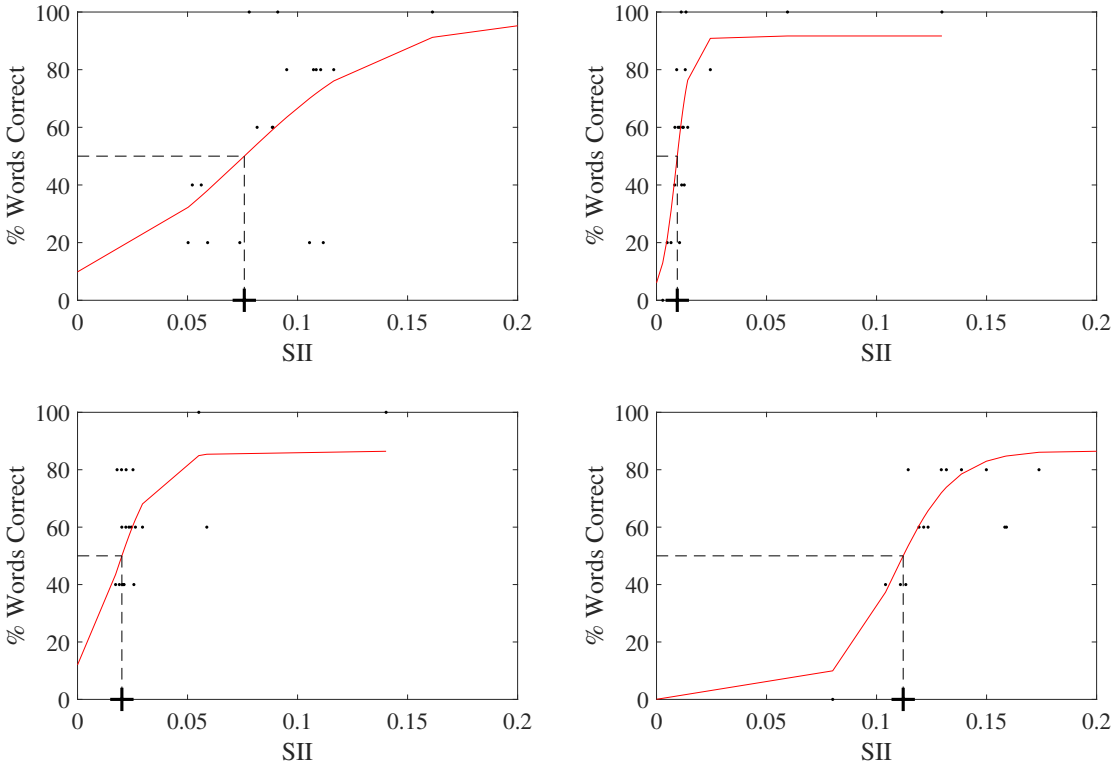


FIGURE 7.1: Example logistic mappings between SII and percentage of words correct in the speech intelligibility test, for a range of subjects and test conditions.

For each new array and masker design, the fixed value of $\text{SII}_d = 0.05$ can be used to select a corresponding value of SNR_d. From a design perspective, this SII-based approach to selecting SNR_d is attractive as it eliminates the need to conduct listening tests for each new loudspeaker array design, but the limitations of the method must also be taken into account. Table 7.1 shows

the SNR required to achieve privacy (SNR_d), according to either the measured SRT values, or using SII_d following the process described above.

Condition	SNR_d (dB)	
	$SII_d = 0.05$	SRT
W_S	-16	-21
W_A	-27	-32
W_∞	-12	-17
N_S	-13	-15
N_A	-12	-12
N_∞	-12	-10

TABLE 7.1: Comparison between dark zone SNRs, SNR_d, required for privacy when estimated using SII simulations and experimental SRTs.

Although $SII_d = 0.05$ represents the average SRT across all tested array geometries and conditions, as described above, the results in Table 7.1 show that there is a 5 dB difference between the required SNR for the two methods in the case of the wide array. This can be attributed to the design of the SII algorithm, which aggregates SNRs from several frequency bands into a single-number rating using a weighted average. If an array aliases within the speech frequency range, and this results in a reduced contrast, for example as shown in Figure 5.18 at 3 kHz for the wide array, then the SNR in this band is greater than in other bands. Consequently, some speech sounds can be more easily understood than others, increasing the intelligibility over that predicted by the weighted average SNR [224]. It is therefore recommended that array designs that exhibit sharp reductions in acoustic contrast within the speech frequency range, due to spatial aliasing or other effects, be avoided. With these designs, additional masking is required to compensate for these limitations, and the true intelligibility cannot be as reliably predicted using standard metrics [124].

7.1.2 Comparison Between Preference Test Results and Subjective Metrics

Subjective metrics are designed to quantify perceptual features from signals, and are useful alternatives to costly, complex jury testing. Table 7.2 shows the values of the psychoacoustic annoyance, loudness, roughness and sharpness metrics that were described in Chapter 4, when applied to the stimuli presented in the preference test that was described in Section 6.2.2. As these metrics are calculated across several sentence presentations, each metric evaluation has an associated uncertainty, and this is displayed as ± 1 standard deviation in Table 7.2. Highlighted cells indicate where the order of the metric evaluations, from lowest to highest, matches the order of preference from the test results displayed in Table 6.4, without an overlap in the stated confidence interval. The preference order from the listening tests is reproduced for convenience in the first column of Table 7.2.

From these results it can be seen that for the wide array, where preferences between conditions were very clearly agreed upon by participants, that the loudness, roughness and psychoacoustic

annoyance metrics all correctly predict the order of preference. In the tests using the narrow array, the N_∞ and N_A conditions were equally preferred over the N_S condition. In this case, the psychoacoustic annoyance evaluation fails to distinguish all three narrow array conditions. Greater variation is expected for the psychoacoustic annoyance, as this metric is formed by combining the other metrics, each of which has its own degree of uncertainty. Although the metrics themselves are deterministic, the uncertainty in the evaluation of each condition stems from the fine structure of the randomly generated speech and noise signals. The loudness metric is less susceptible to these variations, and thus is able to confidently predict that N_∞ is one of the preferred conditions for the narrow array.

Sharpness, which was hypothesised during test development to be an undesirable attribute, was instead found to be inversely related to the preference results; signals with higher sharpness values were preferred on average. However, attention must be paid to the absolute sharpness value. The formulation of the sharpness model [79] states that this attribute only affects psychoacoustic annoyance when it exceeds 1.75 acum (Equation 4.8); none of the tested values exceed this threshold as speech-shaped noise contains relatively little energy above 3 kHz, where sharpness begins to be perceived. Comments from participants who selected the case N_A over N_∞ suggest that the increased sharpness of the N_∞ condition was undesirable.

Condition, pref. order	Annoyance	Loudness	Roughness	Sharpness
W_∞, 1	14.49±0.26	13.68±0.17	0.04±0.003	1.46±0.01
W_S, 2	18.94±0.46	17.50±0.29	0.06±0.003	0.99±0.02
W_A, 3	26.34±0.80	23.68±0.41	0.08±0.009	0.82±0.03
N_∞, =1	16.33±0.25	15.25±0.15	0.04±0.003	1.67±0.01
N_A, =1	16.52±0.25	15.52±0.12	0.04±0.003	1.62±0.01
N_S, 2	16.81±0.39	15.64±0.23	0.05±0.003	1.33±0.02

TABLE 7.2: Metric values of stimuli presented to participants in the paired-preference test. Uncertainties are $\pm 1\sigma$. Highlighted cells indicate where the metric correctly predicts the order of preference within each array width.

Table 7.2 shows that the loudness and roughness metrics perform better than the annoyance and sharpness metrics overall, with the roughness metric correctly predicting the preference order of all conditions and the loudness metric correctly identifying one of the most preferred options in the case of the narrow array. Cursory inspection of Table 7.2 suggests that the roughness metric is the superior predictor, but the absolute value and Just-Noticeable Difference (JND) for roughness perception must also be taken into account. The roughness of unmodulated noise is caused by random amplitude fluctuations and is therefore dependent on its bandwidth [168]. For example, peak roughness between 0.2 and 0.3 asper occurs for noise with a bandwidth of 100 Hz, decreasing thereafter to around 0.05 asper at full audio bandwidth. For amplitude modulated tones, the threshold of roughness perception is 0.07 asper, and the JND limen $\Delta R/R$ is 17%. Therefore, while the roughness values of the stimuli can be distinguished from one another, the absolute roughness level of all the tested stimuli is already very low. Roughness is difficult to explicitly control without also affecting other perceptual features, as no explicit amplitude modulation is included in the masker. Furthermore, the roughness of a given signal increases

slightly with signal level [168], and is thus non-linear. This is exemplified in the case of the wide array where, although the roughness metric correctly predicts the order of preference, this effect is due to the large level difference between stimuli.

The recommendation is, therefore, that a masking signal should be determined based primarily on minimising loudness, with attention also being paid towards minimising the residual roughness of the masker. This further motivates investigation into the context-dependence of masker preference, as the preference for "naturalness" expressed by participants appears to be well-modelled by the roughness metric when applied to stationary random noise. An investigation using a range of natural masking sounds may also shed light on the advantages and disadvantages of masker fluctuation, which is negligible for the stimuli tested here.

7.2 Identification of Feasible Masking Signals

The listening test results described above show the SNR that is required in the dark zone to achieve privacy at three low-pass filter settings, for each array configuration. However, this SNR can also be calculated using SII = 0.05 as a proxy for privacy, using auralisations of the speech and noise emitted from the array, recorded in each zone. Signal auralisations are not strictly necessary for this purpose, as the SII algorithm only uses the spectra of the speech and noise signals to form its intelligibility estimate. These spectra can be synthesised by combining the frequency response of the array with the predicted acoustic contrast and standard speech spectra. The adjustments to the speech-shaped maskers presented in the test can be described using two parameters: the masker level and the cut-off frequency of the low-pass filter. By generating auralisations across a range of values of these two parameters, and analysing the SII in the bright and dark zones at each data point, contours of bright and dark zone SII can be generated. These are shown for the narrow and wide array in Figure 7.2.

Each point on each contour plot shown in Figure 7.2 represents the outcome from a single simulation, where the masker has been low-pass filtered with some cut-off frequency, f_c, and has had its gain adjusted to produce a given SNR in the dark zone, SNR_d. This representation of the masker level, which assumes a constant speech level, is chosen to facilitate comparison with the listening test results presented in Section 6.3.1. The upper row of plots shows the SII in the dark zone, for the narrow array on the left and the wide array on the right. The middle row shows the corresponding SII in the bright zone. The intelligibility constraints $\mathrm{SII}_d < 0.05$ and $\mathrm{SII}_b > 0.75$ are visualised as contour lines of equal intelligibility through the parameter space in the upper and middle plots respectively. All points in the parameter space below the white contour at SII = 0.05 represent situations with sufficiently low intelligibility in the dark zone to claim privacy. Likewise, all points above the bright zone constraint contour, shown by the black line at $\mathrm{SII}_b = 0.75$, exceed the ANSI guideline for "good" speech reproduction in the bright zone [124].

From the results presented in Figure 7.2, it can be seen that in order to provide speech privacy with a masker whose filter cut-off frequency, f_c, is low, the SNR must be reduced significantly as the masker and speech signal do not overlap sufficiently in frequency for the masker to be effective. As f_c is increased, the speech and masker spectra become more similar, so the masker gain can be reduced (i.e. the SNR increased) whilst maintaining the same predicted intelligibility

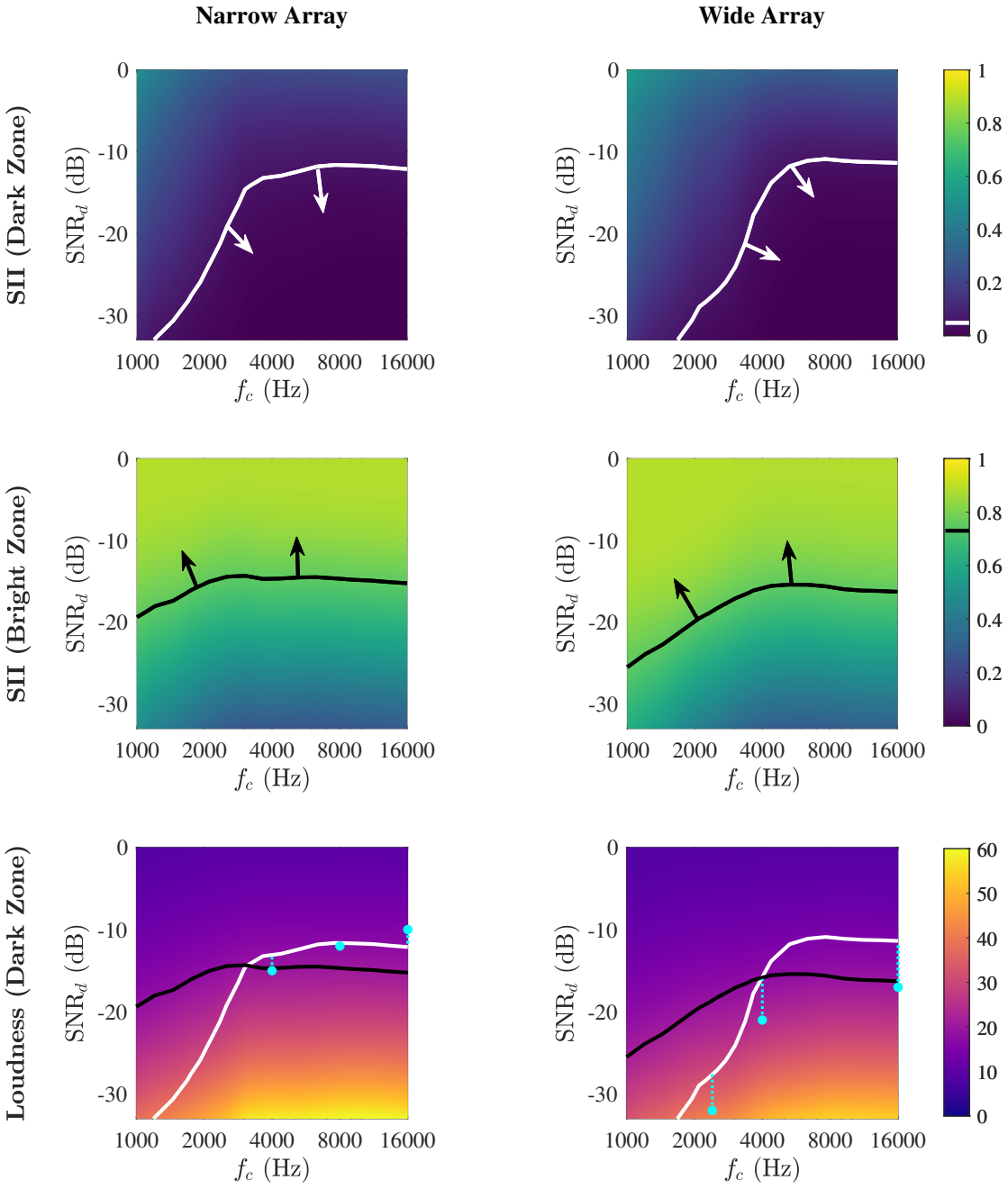


FIGURE 7.2: Contour plots of SII and Loudness with variation in dark zone SNR and masking signal cut-off frequency. Upper Row: SII in dark zone. Middle Row: SII in bright zone. Lower Row: Loudness in dark zone. White line: $\text{SII}_d = 0.05$, Black line: $\text{SII}_b = 0.75$. Arrows indicate regions of the parameter space where intelligibility constraints are met.

level. Above 5 kHz, the contours become approximately constant with an increase in f_c. This is a feature of the SII metric, which assigns each critical band an importance value. Above 4.8 kHz, the relative importance of each critical band to speech intelligibility sharply decreases, so changes in the difference between the speech and noise spectra have little impact on the final value of the SII.

The intelligibility contours from the upper four plots in Figure 7.2 are transferred onto the lower row of plots to provide an enclosed *feasible region* in which both intelligibility constraints are met. The colour scale in the two lower plots represents the loudness evaluated in the dark zone. As SNR_d increases, the loudness evaluated in the dark zone decreases to a constant level that corresponds to the loudness of the leaked speech audible in the dark zone. As f_c increases, so does the perceived loudness, due to the wider bandwidth of the masker, as demonstrated in Figure 4.4. The light blue points in the lower panels of Figure 7.2 represent the experimental SRT values from Table 7.1. The dashed lines between these points and the white contour indicate the difference between the SNR required to achieve the measured SRT and SII = 0.05 at the three filter cut-off frequencies. From this lower set of plots it can be seen that the optimal masking signal parametrisation within the feasible region can be decided by considering the perceptual attributes of the dark zone sound field, a decision that is guided by preference test results and subjective metrics.

It is important to highlight that the contours presented here are specific to the position of the zones, the loudspeaker array used and the surrounding room acoustics. However, the discussion relating to the relative positions of the contour lines and the behaviour with different spatial aliasing characteristics is expected to hold for a range of multi-zone configurations, and therefore provide general insight into the privacy control design problem. Certain situations can cause there to be no intersection between the regions where $\text{SII}_b > 0.75$ and $\text{SII}_d < 0.05$, for example when the size of the array (in terms of the array length, D, or the number of elements, L) prevents sufficient acoustic contrast from being provided, or if room reverberation is too high. When this is the case, either the constraints on the bright and/or dark zone intelligibility must be made less onerous or the system must be redesigned.

7.3 Summary

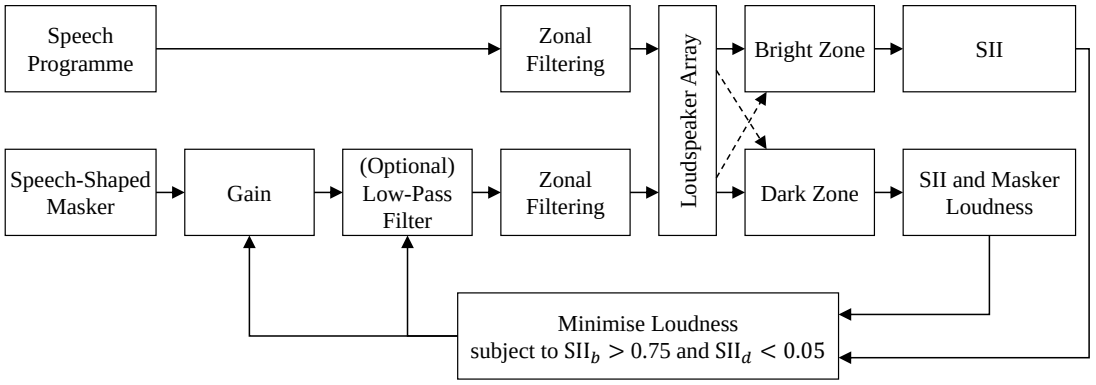


FIGURE 7.3: Block diagram of the proposed personal audio system design method.

The overall structure of the proposed design process is described in block diagram form in Figure 7.3. The method requires an estimate of the transfer responses from a candidate array design to the designated bright and dark zones. These transfer responses, whether derived from measurements or simulations, enable the production of sound zoning filters, which can be used to create auralisations of speech and masker signals in each zone. Speech intelligibility is evaluated in both zones using the SII metric [124], to determine the range of reproduction levels that satisfy the SII constraints of SII < 0.05 in the dark zone and SII > 0.75 in the bright zone. If these constraints can be met simultaneously the masker with the lowest loudness, evaluated in the dark zone, can be selected. Otherwise, the array geometry or the positions of the zones are potentially unsuitable for effective private sound zone reproduction, and measures should be taken that increase the level of acoustic contrast. This can include increasing the aperture of the loudspeaker array, increasing the number of array elements, moving the bright and dark zones further apart from one another and the bright zone closer to the loudspeaker array. Further improvement can be made by adding absorption to the reproduction space, to reduce the reverberation time. The effects of reverberation on sound zoning performance are explored in more detail in the following chapter.

Chapter 8

The Effects of Reverberation on Zonal Speech Privacy

Parts of this chapter are taken from "The Design of Personal Audio Systems for Speech Transmission using Analytical and Measured Responses" which has been published in the proceedings of the 44^{th} IEEE International Conference on Acoustics, Speech, and Signal Processing.

The results in the previous chapter demonstrate that adequate intelligibility difference between listening zones is achievable using a 9-element loudspeaker array in a room with a relatively low reverberation time. However, these types of environments are uncommon in practice and excessive reverberation in a reproduction environment can negatively affect both the measurable and perceived performance of personal audio systems [11]. This chapter will therefore discuss several aspects of private personal audio system design for reverberant spaces. As described in Chapter 5, the process of generating sound zones relies on robust estimates of the transfer responses between the loudspeaker array elements and the zones. In reverberant spaces, appropriate transfer responses are difficult to model accurately, and background noise can impede the acquisition of clean transfer response measurements [225]. The level of acoustic contrast that is achievable between listening zones can also be reduced by reverberation, as dark zones with low SPLs cannot be formed as effectively due to the diffuse, homogeneous nature of the reverberant field. Additionally, reverberation can degrade speech intelligibility by decreasing the effective modulation depth of the speech signal; as described in Chapter 3, short periods of silence provide listeners with information on the boundaries between syllables and words [226], and this information is reduced when both direct and reflected sound arrive at the listening position.

Nevertheless, approaches have been proposed to improve the performance of sound zoning systems in reverberant spaces. Individual room reflections can be taken into account in the sound zoning process [12], either through their implicit inclusion in measured transfer responses, or by modifying an analytical transfer response model to include image sources [227]. In some circumstances, this information about the presence of reflective surfaces in the room could be used to selectively cancel the reflected waves, leaving only the direct sound from the array. Alternatively, reflections deemed to be beneficial to speech intelligibility and clarity in the bright zone, such as those that arrive within 50 ms of the incident sound [36], could potentially be allowed to

remain in the reproduced sound field. The sound field control method proposed by Chaman et al. [97], discussed in Section 2.3, actively requires the presence of reflections in the reproduction environment to provide sufficient mixing between the signals emitted by each loudspeaker, but the reliability of this approach is suspect as it relies on the integrity of many separate reflection paths, long into the reverberant decay. This observation highlights that room reverberation includes both early reflections and late reverberation, and each of these features must be treated somewhat separately. Early reflections are localisable and coherent with the direct sound, making them amenable to inclusion in sound zoning processes, whereas late reverberation can be modelled as an incoherent, diffuse phenomenon. Consequently, a standard approach is to consider late reverberation as a source of random error that sound zoning systems must be robustly designed to overcome [58].

In order to investigate the effects and practical challenges of operating private personal audio systems in reverberant spaces, the chapter opens by describing the potential effects of incorporating room reverberation into the transfer responses that are used to generate sound zoning filters. Methods for simulating anechoic and reverberant transfer responses are provided, and each of these approaches is used in a study of the spatial decay of sound pressure from a personal audio system situated in a free-field environment and in a reverberant room. This study is extended in subsequent sections to include several simulated reverberant spaces in order to quantify the effect of reverberation on acoustic contrast, speech intelligibility and the masking signal levels that are required for privacy. The chapter concludes by comparing these simulations against results measured in a lightly reverberant room, to determine the situations in which measured or modelled transfer responses are most suitable.

8.1 Transfer Response Modelling

Information about the electroacoustical transfer responses from a loudspeaker array to sensors within each zone are necessary for the optimisation of the zoning filters in a personal audio system. These transfer responses can either be analytically calculated using the geometry of the sources and zones, or measured using the process described in Section 5.1.4. Figure 8.1 shows a block diagram of a personal audio system in a reverberant space, and illustrates a key difference between the use of an analytical transfer response model and measured transfer responses, namely, the inclusion of room reflections. The simplest analytical transfer response is that which assumes a free-field, point monopole sources and omnidirectional sensors, though other methods which attempt to model source directivity and room reverberation exist, and are discussed later in this section. A common heuristic, which is supported by the derivation for ACC in Section 5.2.1, is that the greater the mismatch between the transfer response used in the optimisation of the filters and the physical room responses, the poorer the contrast performance. This mismatch could be attributed to imprecision in manufacturing tolerances of the source array, errors in geometrical measurement of the arrays and zones or variation in loudspeaker or microphone sensitivity. However, even if these factors are perfectly measured or exactly modelled by the designer of a system, assumptions regarding the acoustic conditions of the space where the system is installed also have significant effects, as the results presented in this chapter will demonstrate.

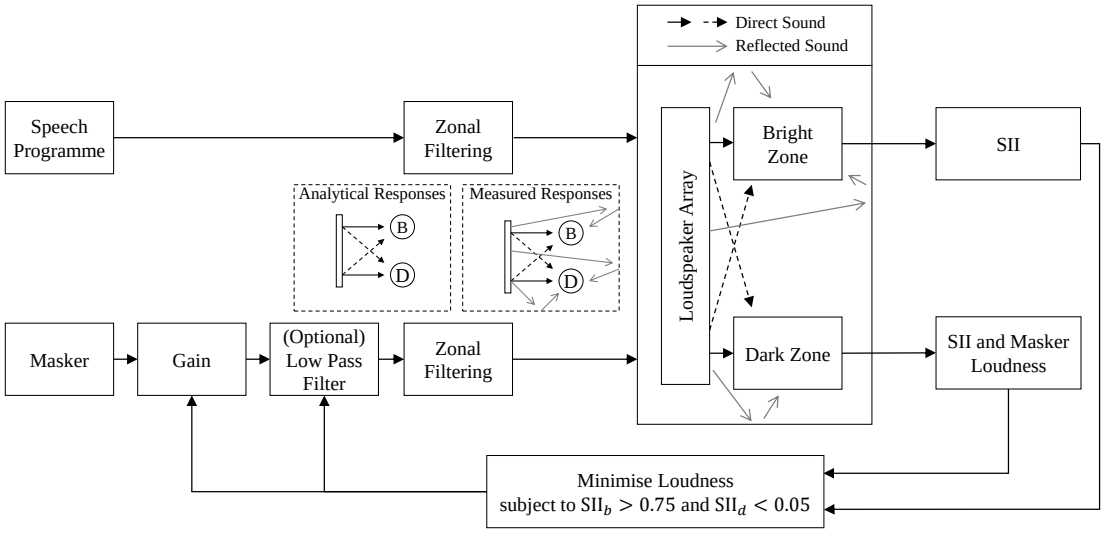


FIGURE 8.1: Block diagram of personal audio system in reverberant space. The zonal filtering process can make use of either analytical or measured transfer responses.

Implementations and analyses of sound zoning systems have been conducted in anechoic environments e.g. [2, 5, 58], with the inclusion of reflections from the head [228], individual room reflections [12] and general reverberation, e.g. [4, 11, 46, 60]. Of these, the performance study carried out by Olik et al. [60] is of particular relevance to the results presented in this work as it considers the leakage between programmes in adjacent bright zones in terms of a perceptual index, the perceived distraction, as well as acoustic contrast. Overall, correlation was found between the perceptual and physical metrics, with higher levels of acoustic contrast resulting in lower levels of distraction, though the strength of this relation varied with different combinations of programme material. In the study, interfering speech was found to be the most distracting. Olivieri et al. [46] present results from informal listening tests that suggest that zoning filters created using free-field responses, as opposed to measurements in anechoic or reverberant conditions, provide subjectively higher audio quality. Although filters generated using the measured responses produced greater directivity than when using a free-field assumption, the perceived channel separation was similar. The question of whether the expense and time involved in measuring transfer responses from a loudspeaker array, potentially in-situ, is justified by a practical increase in performance therefore remains open. The following subsections describe techniques that can be used to model electroacoustical transfer responses, and provide comparisons with measured responses.

8.1.1 Analytical Modelling

One advantage of using modelled transfer responses to generate sound zoning filters is that the responses can be generated based on the geometry of the loudspeaker array and zones. This means that prototype array systems can be built and rapidly tested, without having to conduct acoustical transfer response measurements first. An additional strength is that the positions of the zones can be easily changed by specifying the coordinates of new virtual microphone

positions. Figure 8.2 shows the notation used in the remainder of this section to describe these arrays of coordinate points in 3D space.

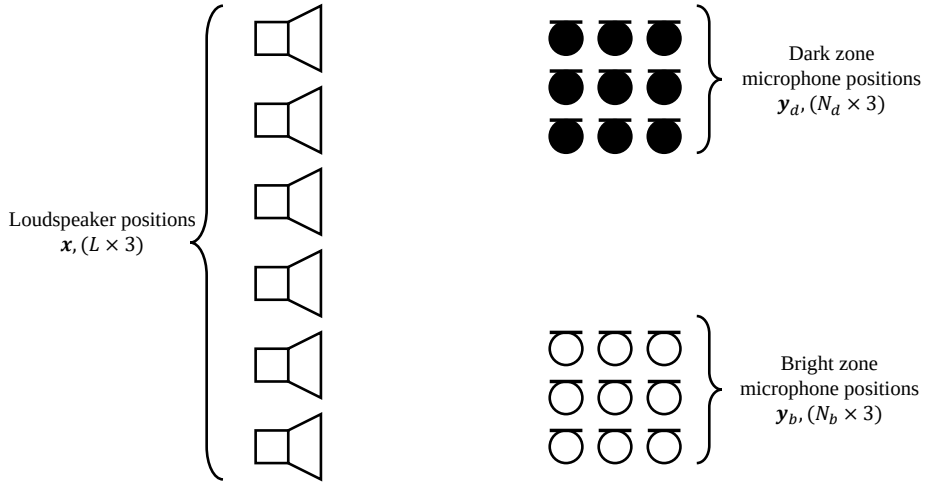


FIGURE 8.2: Notation used to describe the geometrical positions of loudspeaker and microphone array elements, for analytical transfer response modelling.

The simplest type of geometrical modelling of electroacoustical transfer responses assumes a free field containing identical, omnidirectional point sources and sensors. In this case, the $(n,l)^{th}$ elements of the transfer response matrices $\mathbf{Z}_b$ and $\mathbf{Z}_d$ are given by the free-field Green's function between the bright or dark zone microphone n and loudspeaker array element l, which can be expressed at each frequency as

$$Z_{nl,free-field} = \frac{e^{-jkr_{nl}}}{4\pi r_{nl}}, \tag{8.1}$$

where

$$r_{nl} = \sqrt{(\mathbf{y}_{(n,1)} - \mathbf{x}_{(l,1)})^2 + (\mathbf{y}_{(n,2)} - \mathbf{x}_{(l,2)})^2 + (\mathbf{y}_{(n,3)} - \mathbf{x}_{(l,3)})^2}, \tag{8.2}$$

and k is the wavenumber. The ACC method derived in Section 5.2.1 is general with respect to the form of transfer response used to model the source-sensor relationship, and more advanced source terms can also be used. For example, a slightly more complex model could be obtained by modelling the loudspeakers as piston-like sources mounted in an infinite baffle, which is a commonly used loudspeaker model [38]. In the far field, where $r_{nl}/a \gg 1$, the transfer impedance in this case can be expressed at each frequency as

$$Z_{nl,piston} = \left[\frac{J_1(ka\sin\theta)}{ka\sin\theta}\right] \frac{e^{-jkr_{nl}}}{4\pi r_{nl}}, \tag{8.3}$$

where J_1 is the first order Bessel function of the first kind, θ is the angle that the sensor makes with the source axis and a is the radius of the piston source.

A further alternative to monopole source modelling, which allows the directivity to be specified in both the horizontal and vertical directions separately, is to incorporate directivity terms into the transfer impedance, converting the omnidirectional monopole source to a given first order polar pattern, such as a cardioid or dipole response [229]. The far-field directivity, D of such a source is given by

$$D(\Psi, \theta, \phi) = 1 - \Psi + \Psi \cos\theta \cos\phi, \tag{8.4}$$

where θ and ϕ are the azimuth and elevation angles respectively, and Ψ is a directivity parameter that can be continuously varied to obtain different directivity indices. A value of $\Psi = 0.75$ has previously been used to model an array of phase-shift sources, and a similar effect can be accomplished by modelling each source as a pair of closely-spaced point monopole sources, with a fixed delay relationship [229]. A modification of this approach can also be used to simulate the presence of reflections in the reproduction environment [12, 227]; reflections are modelled as direct sound from additional virtual sources, specified with a delay and gain consistent with the path-length difference between direct and reflected sound from the original source. In closed spaces with multiple reflections, this process can be repeated to form an estimate of the room impulse response, and this technique is discussed in the following section.

8.1.2 Image Source Modelling

In order to test personal audio systems in a range of reverberant environments, the image source model of reverberation was used to synthesise impulse responses from loudspeaker array elements to sensors in each zone. This model uses a high-frequency assumption, which represents the set of points on an expanding wavefront directly between a source and a sensor as straight rays. This mitigates the computational demand of computing the entire sound-field, but also removes wave-effects such as diffraction and refraction. In closed, simple volumes such as the cuboid rooms referenced in this chapter, these effects are expected to be negligible.

Snell's law states that the angle of incidence of a ray with a reflecting surface is equal to the angle of reflection. Using this relation, an image source can be synthesised on the opposite side of the reflecting surface, with its strength appropriately attenuated to match the absorption coefficient of the reflecting surface, meaning that the surface can be removed from the simulation. The travel time and direction of arrival of the ray from this image source is equal to that of the reflected ray, so from the perspective of the sensor, there is no difference between the impulse response received using a single source and the reflecting surface and that with two sources and no reflecting surface. This geometrical arrangement is illustrated in Figure 8.3.

As mentioned in the previous section, the effect of a single reflection can be modelled analytically by directly adding an appropriately attenuated and delayed image source term into an analytical transfer response model, such as those described in Equations 8.1 and 8.3 [12]. However, when multiple reflective surfaces are present, such as is the case in a closed room, it becomes more convenient to model image sources numerically [230]. Additional higher order reflections, that is, rays from the source which reflect off more than one surface before being intercepted by the sensor, can be incorporated by adding further image sources. However, this process can quickly

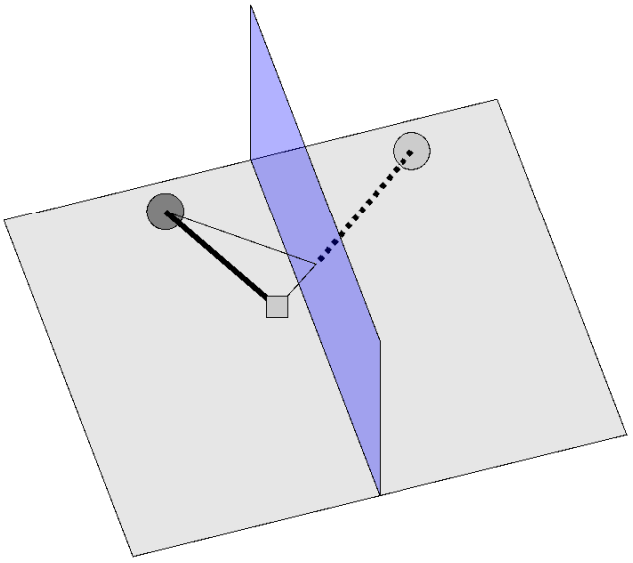


FIGURE 8.3: Diagram of a single reflection, modelled using the image-source method. The heavy, solid line indicates the direct path from the source (dark circle) to the sensor (grey square). The reflection in the blue wall is modelled by tracing a ray from a virtual image source (grey circle) to the sensor.

become computationally intractable as the number of image sources rises combinatorially with the computed reflection order. Typically, prominent reflections either heard as distinct echoes or perceived as a broadening of the sound source occur early in the reverberant decay, and later reflections are perceived as diffuse reverberation. This perceptual effect is harnessed by Lehmann and Johansson in their diffuse reverberation model, which transitions between a low-order image source model for early reflections and a decaying random noise model for the late reverberation [230]. The decay rate of the random noise is informed by a prediction of the energy decay curve, which is exponential in rooms where the absorption is well-distributed across all surfaces. A MATLAB implementation of this model [231] provided by Lehmann is used in the remainder of this chapter. This implementation assumes that all sources within each modelled room are point monopoles. For consistency when comparing various aspects of system performance later in this chapter, where an analytical model is referenced, this implies that the associated transfer responses are also those of monopole sources, radiating into a free field, as described in Equation 8.1. The loudspeaker drivers used in the 27-channel array each have a radius, a, of 32 mm, and are mounted in separate sealed enclosures so this monopole assumption is valid up to a frequency of approximately 3.4 kHz ($ka = 1$) [232].

8.1.3 Comparisons with Measurements

To illustrate the differences between the transfer response models described in the previous two subsections, each model is used to generate an impulse response from one element in the 27-channel array described in Section 5.1.1 to a microphone within the ISVR audio laboratory, described in Section 5.1.3. These are presented in Figure 8.4 and are compared against a measured impulse response corresponding to the same source-sensor combination, in the same space.

The ideal impulse from the monopole source lies between samples in the digital representation shown in the upper panel of Figure 8.4, so this impulse response exhibits a minor ringing artefact, which is also visible in the frequency domain as a small ripple at high frequencies in the lower frequency response plot. When compared to the measured impulse response, shown with a blue trace, the arrival time of the direct sound is accurately represented.

Comparison of the output from the image source model against the measured impulse response shows that the arrival time of the direct sound and first three dominant reflections, indicated by red asterisks, are well matched by the simulation, though the relative magnitudes are less well reconstructed. A number of possible reasons can be given for this discrepancy; there may be a mismatch between the estimated absorption coefficients of the room surfaces that are input into the model and the physical absorption coefficients of the associated surfaces in the room. Alternatively, the stronger reflections in the modelled impulse response could be attributed to the perfectly specular reflection which is assumed by the model, when in practice, the reflection from the wall will include some acoustic scattering. However, the main cause is likely to be due to the assumed omnidirectional nature of the modelled source - this will increase the off axis radiation compared to the physical source, and hence result in more significant reflections. This feature also explains the spurious impulses that are present in the simulated impulse response but are not evident from measurements, indicated with black asterisks in Figure 8.4. The omnidirectional source used in the model causes a reflection from the wall behind the array to be received at the virtual microphone position. In practice, the directivity of the individual sources in the array means that this reflected path makes very little contribution to the measured impulse response.

In order to more closely approximate the measured frequency response of the system, the response of the monopole sources in the image source model have been equalised to match the average frequency response of all the sources in the array. The small random variation between the measured and simulated frequency responses in the example given in Figure 8.4 is due to the specific interactions of the individual reflections at the selected sensor location, and the random diffuse tail of the impulse response. The image source model will be used in the following section to illustrate the effect of locating the listening zones of a personal audio system close to the array, where the direct sound dominates, versus further away, where the diffuse reverberation forms a significant component of the sound field.

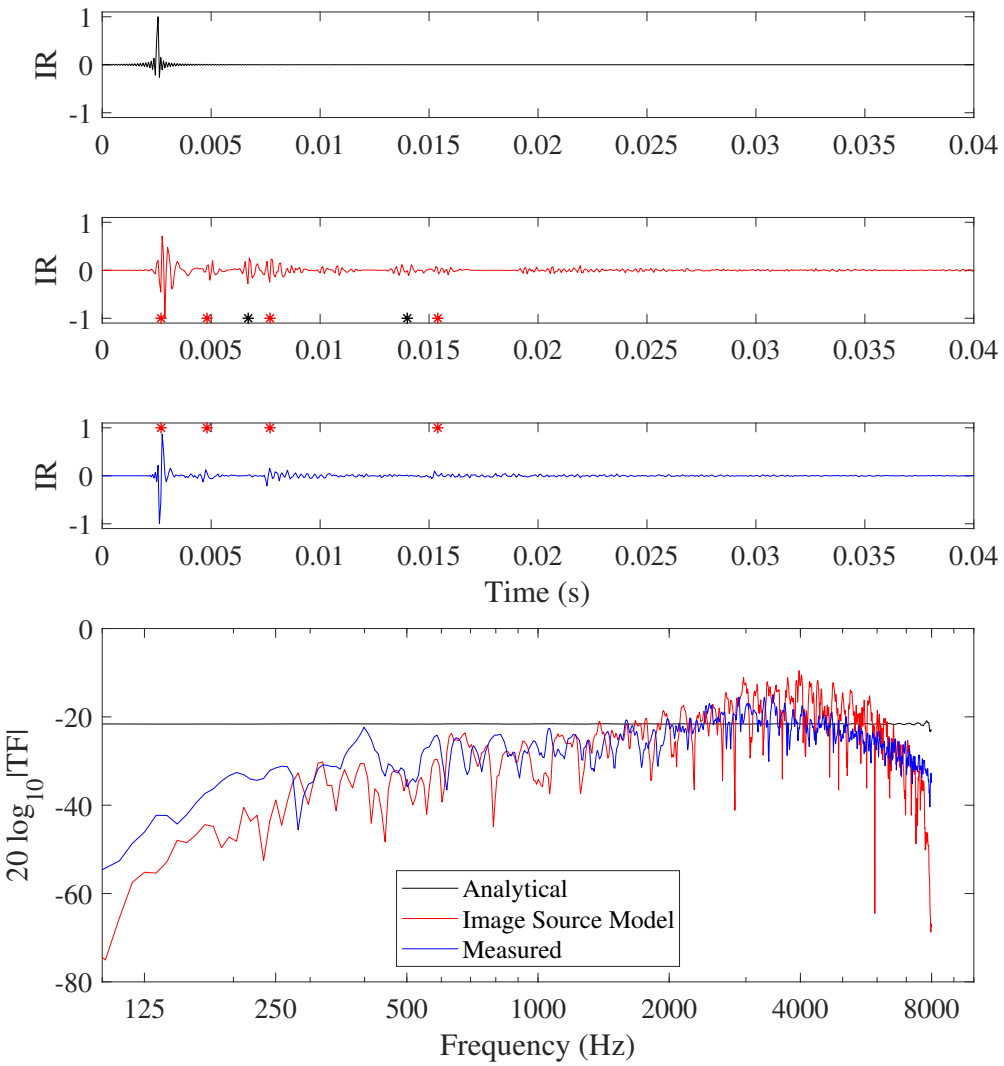


FIGURE 8.4: Impulse responses and associated transfer response magnitudes between a single source and sensor location in the ISVR audio laboratory (see Section 5.1.3). Black traces show simulated responses based on the free-field propagation model, red traces are simulated with the image source model, and blue traces represent measured responses.

8.2 The Effects of Array-to-Zone Distance in Reverberant Spaces

The image source model is useful for predicting individual early reflections in a reproduction environment, but the repeated reflections that characterise the late reverberation have a secondary effect that must be considered when designing a personal audio system for a reverberant space. When a source emits continuously into a closed room, a steady-state sound field is built up. At each point within the room, the sound field consists of a direct component, which would exist in the absence of reflective surfaces, and a diffuse component which builds up from multiple reflections. The energy in the direct sound field decays with distance away from the source, consistent with the source directivity, but the diffuse component has, to a first approximation, equal energy at all points in the room [183]. Accordingly, at different points in the room, the sound field can be dominated by either the direct or diffuse component. The critical distance,

d_c, associated with a particular source-room combination is defined as the distance away from the source at which the energy of the diffuse sound field in the space equals the direct energy from the source and is given by [183]

$$d_c = 0.1\sqrt{\frac{GV}{\pi T_{60}}}, \tag{8.5}$$

where V is the room volume in m^3 and G is the directivity of the source. G is defined in this case as the ratio of the maximum intensity in a given direction, usually on-axis with the source, to the average intensity over a sphere surrounding the source. For a monopole source, by definition, $G = 1$.

Figure 8.5 shows the predicted monopole critical distance value according to Equation 8.5 for a range of room volumes and reverberation times as a contour plot. A constant ceiling height of 3.0 metres and floor aspect ratio of 1.6 are used. The inclusion of the source directivity factor in Equation 8.5 scales up the critical distance values by a factor of $\sqrt{G}$, meaning that the monopole critical distance, with $G = 1$, provides a lower bound to the true critical distance. As a conservative first approximation, and in the absence of source directivity information which may not be known a-priori, zones should be established within the critical distance predicted using monopole sources, by interpolating the contour plot in Figure 8.5. This guarantees that the zones will be located in a region where the direct sound from the loudspeaker array is dominant over the uncontrolled diffuse field.

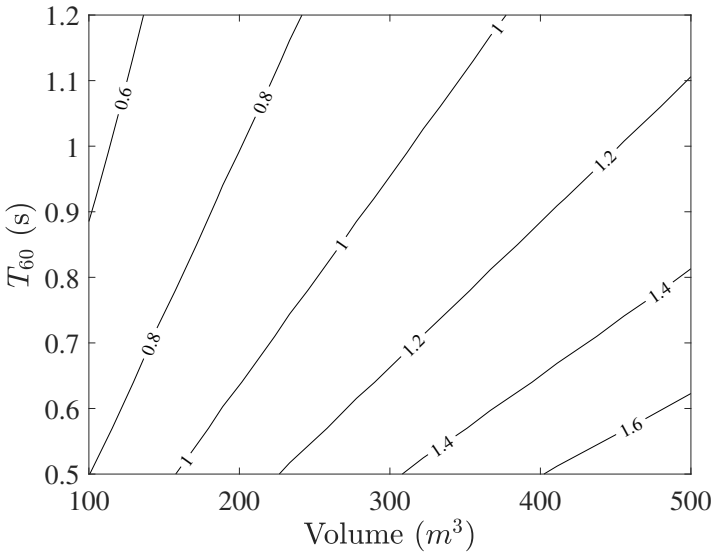


FIGURE 8.5: Contours of critical distance from a monopole source with variation in room volume and reverberation time.

The corresponding average absorption coefficient, $\bar{\alpha}$ for the surfaces of these simulated rooms varies between 0.10 and 0.33. This was calculated using the Sabine reverberation model [233], in which the reverberation time, T_{60} is defined as

$$T_{60} = 0.16\left(\frac{V}{S\bar{\alpha}}\right), \tag{8.6}$$

where S is the surface area of the room boundaries. Substitution of Equation 8.6 into Equation 8.5 allows a second interpretation of the contours in Figure 8.5, based on the effect of the surface area of the room boundaries. For rooms with a given average absorption coefficient, $\bar{\alpha}$, the performance of a personal audio system with zones set a fixed distance from the array, in terms of acoustic contrast, is predicted to be better in larger rooms, with a larger surface area, as the critical distance in these rooms will be further from the array.

Whilst the critical distance gives a convenient single-number value, which can help to inform the feasibility of sound zone positions relative to an array, further information regarding the performance of an array in a space is encapsulated by the spatial decay of SPL. To illustrate this a room representative of the ISVR audio laboratory, with $T_{60,mf} = 0.11$ seconds, is simulated using the image source model. Inputting this reverberation time and the room dimensions into Equation 8.5 yields a critical distance of 1.05 metres. The positions of the 27-channel loudspeaker array and microphones used to demarcate the bright and dark zones are shown relative to the room boundaries in Figure 8.6. Two lines of microphones centred at the origin, passing through the centres of each zone are also shown. These lines of sensors are used to assess the spatial decay of sound levels from the array within the room.

The SPL at a range of source-sensor distances are presented in Figure 8.7. Solid lines indicate the SPL predictions for the array when optimised for, and operating in a free-field environment, and dashed lines indicate the sound levels predicted in the reverberant environment when a separate set of reverberant transfer responses are used to optimise the zoning filters. For comparison with the 27-channel array, the sound field from a single monopole at the coordinate origin, coincident with the centre of the 27-channel array, is also simulated (black lines). The monopole source strength is set such that the level in the bright zone in the room is equal to that provided by the array.

The solid black line in Figure 8.7 shows the expected spatial decay of sound level from a monopole source of -6 dB per doubling of distance. On the logarithmic distance scale used in the figure, this relationship is shown by a straight line. When the same source is assessed in the reverberant room (dashed black line) the SPL begins to plateau as the distance from the source increases because the contribution to SPL from the reverberant field increases. The vertical dash-dot line indicates the predicted critical distance from the monopole source according to Equation 8.5. As the monopole source is omnidirectional, i.e. $G = 1$ in Equation 8.5, this value represents a worst-case, minimum, critical distance. The spatial decay profiles of the more directional 27-channel array shows that the associated critical distance from this source is greater.

Between the 27-channel array and the bright zone, whose position is indicated with grey shading in Figure 8.7, the level along the bright measurement line (red) decays at a lower rate to that predicted for a single monopole. The level difference between 0.2 and 0.8 metres from the array is 4 dB, compared to 12 dB from a monopole. Beyond the bright zone, the free-field decay rate matches that from the monopole source. Along the entire extent of the bright measurement line, there is little difference between the SPL predicted in the free-field and in the room, indicated by the solid and dashed red lines. This suggests that the direct field of the array extends well beyond the location of the bright zone in this example. This is further supported by the slight increase in SPL beyond the location of the dark zone, indicated by blue lines, which would not be possible if this region was dominated by diffuse sound. The blue and red dashed lines, representing the

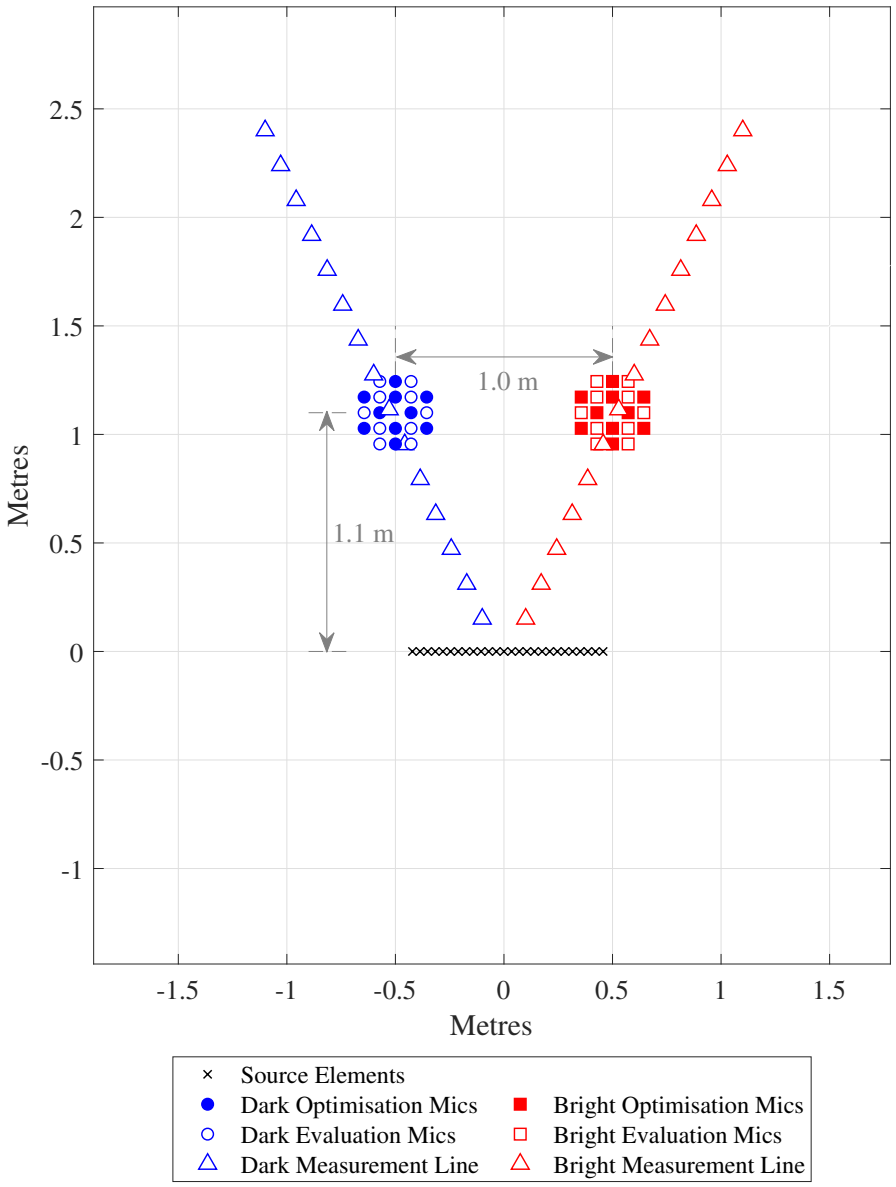


FIGURE 8.6: Plan view of system geometry. The extents of the plot represent the boundaries of the room. Measurement lines correspond to distances in Figure 8.7.

level decay along the dark and bright measurement lines in the reverberant room respectively, converge towards the end of the measurement lines. The measurement lines are only extended to a length of 2.65 metres as any sensors further along this trajectory would be placed too close to the room walls for the image source model to accurately model their output. In larger or more reverberant rooms, the sound field far from the array would be dominated by the diffuse field and the levels at the ends of the measurement lines would be equal. This places a limit on the distance from the source at which the zones can be reliably formed.

Despite the bright and dark zones pictured in Figure 8.6 being clearly situated within the direct field of the array, the broadband level difference in the free field between the centres of the bright and dark zones is predicted to be around 30 dB, whereas in the reverberant room, this reduces to around 15 dB. This shows that even modest levels of reverberation can have a substantial negative impact on the performance of a zonal audio system, compared to predictions in anechoic

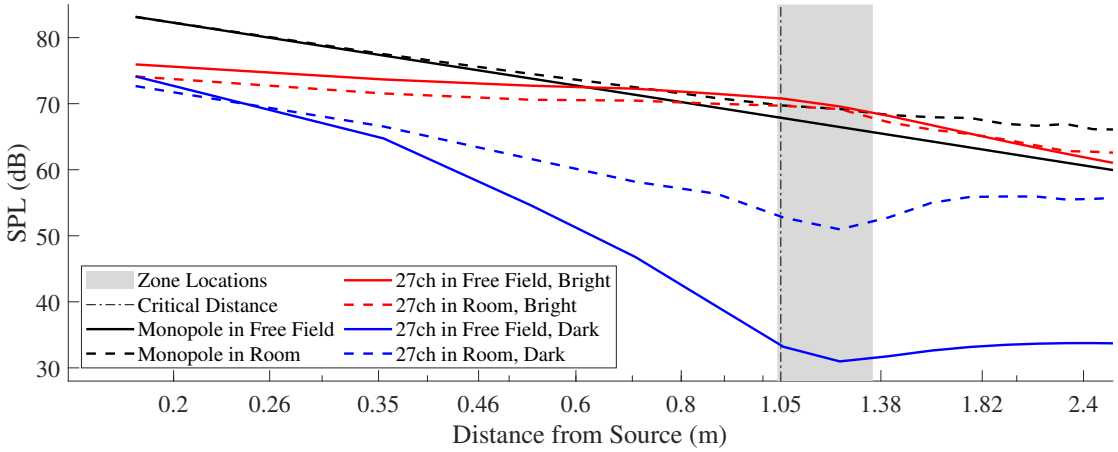


FIGURE 8.7: SPL from a monopole source (black lines) and 27-channel array of monopoles in a simulated free-field and reverberant room at positions along the bright (red) and dark (blue) measurement lines displayed in Figure 8.6. A logarithmic distance scale is used to show the -6 dB SPL decay per doubling of distance from a monopole source as a straight line.

environments. The next section explores a corollary of this conclusion, that speech intelligibility is also adversely affected by reverberation, by considering the effect of changing the surface materials in the same simulated room in order to increase its reverberation time.

8.3 Simulations of Reverberant Spaces

The derivation of ACC in Section 5.2.1 indicates that the transfer responses between source elements and sensors are essential for optimising the zoning filters for the two ACC processes that are necessary for speech privacy control, i.e. focussing the speech into the bright zone and the masker into the dark zone. As referenced previously in this chapter, these transfer responses may be measured in-situ or synthesized based on the geometry and a mathematical model of sound propagation from the sources to the sensors, such as that expressed in Equation 8.1. These analytical responses are free from noise, but may still contain errors due to inaccuracies in the measurement of physical distances. Furthermore, the effect of reflections from the room, which have been shown in the previous section to have an increasingly significant effect as the measurement point moves away from the source array, are clearly not incorporated into these responses.

The effects of reverberation on sound zoning performance are quantified by posing an array processing problem in five spaces, one free of reflections, and the others with various reverberation times, shown in octave bands in Figure 8.8. In each space, the 27-channel array and the microphones that define the zones are arranged as shown in Figure 8.6. The transfer responses from the loudspeaker array to the sensors within each of these spaces are modelled using the image source method described in Section 8.1.2, facilitating a comparison between the use of free-field transfer responses and reverberant transfer responses in personal audio system design. In each case, the systems are evaluated using an independent set of modelled reverberant transfer responses, corresponding to the bright and dark zone evaluation positions shown in Figure 8.6. The simulated rooms are all of equal size to the ISVR audio laboratory, 3.7 $\times$ 4.4 $\times$ 2.3 metres,

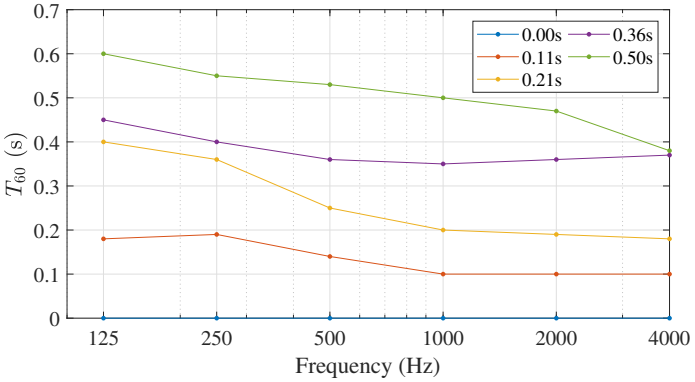


FIGURE 8.8: Octave band reverberation time T_{60} of five simulated spaces. The orange line represents the listening room described in Section 5.1.3. Legend entries show the corresponding mid-frequency reverberation time $T_{60,mf}$.

and the reverberation time is modified by changing the octave band absorption coefficients associated with the the walls, floor and ceiling. In each of the spaces, absorption is evenly distributed across all surfaces, to ensure a diffuse reverberant decay [234]. The orange line in Figure 8.8 matches the measured octave band reverberation time in the ISVR audio laboratory. In the legend of Figure 8.8 and in further plots in this section, each of these reverberation conditions is referred to using the mid-frequency reverberation time $T_{60,mf}$, which is the arithmetic average of reverberation times in the 500 Hz, 1 kHz and 2 kHz octave bands. This index is commonly used in architectural acoustics (e.g. [182]).

Firstly, simulated playback in each room is restricted to a single ACC process, i.e. without additional masking to provide speech privacy. In each space, a 30-second spoken passage from the VCTK corpus [22] is reproduced by the array and measured in the bright zone. The resulting intelligibility evaluations are plotted against $T_{60,mf}$ in Figure 8.9 for the cases where the zonal filters are generated based on monopole, free-field responses and reverberant responses from the image source model. At each data point, the regularisation parameter for the ACC process is selected to maximise the intelligibility in the bright zone, the effects of system regularisation are described in more detail in Sections 8.4.1 and 8.4.2.

Due to the presence of high levels of reverberation in some of the simulated examples, the intelligibility in the bright zone shown in Figure 8.10 is described using the ESTOI metric [121], rather than the SII used elsewhere in this thesis, as the ESTOI metric captures the reduction in intelligibility caused by the loss of modulation information in reverberant spaces. In each space, Figure 8.9 shows that the averaged ESTOI evaluation across the bright zone microphones remains in excess of 0.6. In the following section, this value will be shown using the comparison between Figures 8.14 and 8.15 to correspond to a high level of word intelligibility. The negative correlation between T_{60} and ESTOI is not unexpected, given that T_{60} affects the balance of direct to reverberant sound within the bright zone; direct sound contributes positively to speech intelligibility, and reverberant sound is detrimental to intelligibility. In each simulated reverberant space, the achievable levels of intelligibility are similar between cases where free-field and reverberant responses are used in the ACC process.

To investigate the effect of additional masking on speech intelligibility, a similar experimental procedure to that described in Chapter 7 can be carried out. For a range of reverberation times,

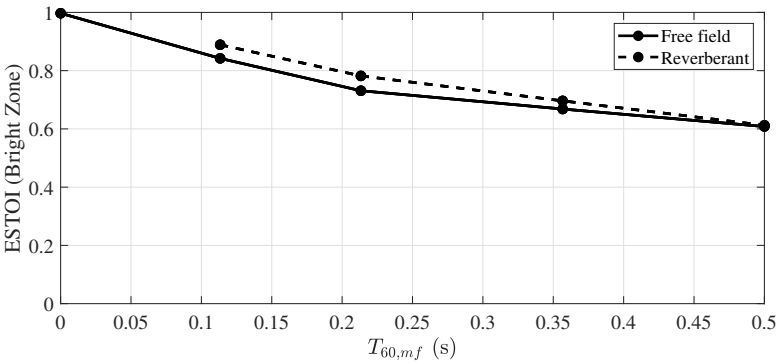


FIGURE 8.9: ESTOI evaluated in the bright zone with variation in room reverberation time with no additional masking, for two systems using sound zoning filters based on simulated free-field transfer responses (solid line) and simulated reverberant transfer responses (dashed line).

the masking signal level is adjusted iteratively such that an ESTOI level of 0.1 is achieved in the dark zone. Comparison of the ESTOI index with the SII in low-reverberation conditions where both metrics are applicable shows that this ESTOI value is approximately equivalent to the dark zone intelligibility constraint of $\text{SII}_d = 0.05$ determined in Chapter 7. Figure 8.10 shows the resulting intelligibility in the bright zone when this constraint is met. When compared to the results presented in Figure 8.9, the bright zone intelligibility falls significantly, due to leakage of the newly introduced masker into the bright zone. This is particularly evident for the case where filters are based on free-field transfer responses; in these trials, the ESTOI value measured in the dark zone is consistently and significantly lower than when reverberant responses are used to design the ACC filters.

The fall in the ESTOI evaluation as the reverberation time increases may appear gradual in Figure 8.10, but the AITF for ESTOI, which converts between metric evaluations and true intelligibility scores, is particularly sensitive in this region. At ESTOI values in excess of 0.5, word intelligibility scores tend to saturate at 100%, whereas privacy can be claimed at ESTOI values less than 0.1. Accordingly, these results predict that high levels of intelligibility can be achieved when reverberation is taken into account in the sound zoning process, up to a mid-frequency reverberation time between 0.25 and 0.30 seconds, whereas when free-field responses are used, this limit is reached at much lower reverberation times. To explain this effect, the masker level in the dark zone and the acoustic contrast, shown as a frequency average between 100 Hz and 8 kHz, are shown in Figures 8.11 and 8.12 respectively. In general, as the reverberation time increases, the achievable acoustic contrast decreases, but the use of reverberant responses consistently provides greater acoustic contrast than when free-field responses are used. For the tested array configuration, the difference in acoustic contrast between the methods increases by 1.3 dB for each 100 ms increase in the mid-frequency reverberation time $T_{60,mf}$, from a 2.3 dB difference at $T_{60,mf} = 0.11$ seconds to a 7.1 dB difference at $T_{60,mf} = 0.50$ seconds. Olik et al. found a similar degree of contrast improvement (0-4 dB) when the effect of a single specular reflection was incorporated into the ACC process [72]. In the context of speech privacy control, this additional contrast reduces the leakage of the programme into the dark zone, meaning that the masker SPL can be reduced. This effect is doubled as the consequent leakage of the masker into the bright zone is also decreased, so higher levels of bright zone intelligibility are possible when using the filters designed using the reverberant responses.

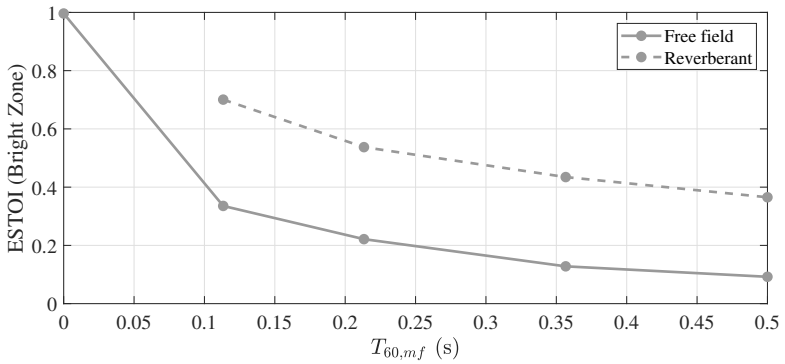


FIGURE 8.10: ESTOI evaluated in the bright zone with variation in room reverberation time, with masking signal adjusted to give $\text{ESTOI}_d = 0.1$, for two systems using sound zoning filters based on simulated free-field transfer responses (solid line) and simulated reverberant transfer responses (dashed line).

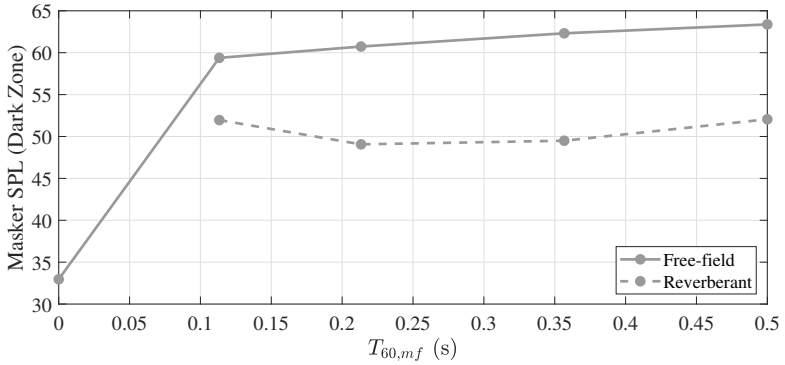


FIGURE 8.11: SPL of masker in dark zone required to maintain the dark zone intelligibility constraint $\text{ESTOI}_d = 0.1$ with changes in reverberation time, for two systems using sound zoning filters based on simulated free-field transfer responses (solid line) and simulated reverberant transfer responses (dashed line).

The results from the simulations presented in this section show that room reverberation can severely degrade both the acoustic contrast and speech intelligibility contrast that can be provided by a speech privacy control system. When reverberant responses that are matched to the reproduction environment are used, the simulated experiments suggest that performance is significantly improved compared to the case where free-field responses are assumed in the sound zoning process. To appraise the validity of this simulated study, results based on experimental measurements are provided in the following section. In this case, incorporation of the room reverberation into the sound zoning process is achieved through the use of measured transfer responses from the ISVR audio laboratory. These are compared against using free-field responses based on the geometry shown in Figure 8.6 in terms of the acoustic contrast, speech intelligibility and the masking signal levels required for privacy.

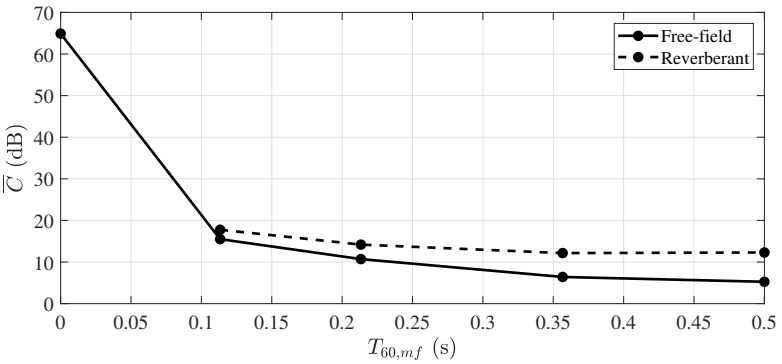


FIGURE 8.12: Acoustic contrast, frequency averaged from 100 Hz to 8 kHz in each simulated reverberant room, for two systems using sound zoning filters based on simulated free-field transfer responses (solid line) and simulated reverberant transfer responses (dashed line).

8.4 Experimental Validation

The process of measuring electroacoustical transfer functions, as described in Section 5.1.4, is time-consuming and requires large quantities of measurement equipment, so if simpler solutions based on the system geometry are shown to result in adequate performance, this will substantially increase the practicality of implementing speech privacy control systems. Results from the previous section indicated that there was a tangible benefit to including the reverberant room response in the sound zoning process, both in terms of acoustic contrast and the desired speech intelligibility contrast between listening zones, but these were based on simulations, which contain none of the uncertainty associated with transfer response measurements [225].

In Section 8.1, three methods were described for estimating the true transfer responses between the sources and sensors in the personal audio system. Transfer response measurements represent the gold-standard, as these fully characterise the acoustical path from source to sensor within the room, but two alternative approaches were also described; simulating the sources and sensors using a free-field approximation, and modelling the room using the image source method. The former approach has been used successfully in a number of publications [5, 61, 65, 229], but a substantial body of evidence in the literature suggests that producing simulated room impulse responses of sufficient fidelity for sound field control is impractical. Firstly, compared to the simplicity of free-field simulations, including reflections into a transfer response model requires significantly more geometrical information, leading Jacobsen et al. [61] to dismiss the incorporation of reflections as an unnecessary complication in their comparison of sound field control strategies. Beyond this increase in complexity, the benefits of including modelled reflections are also limited. In the previously mentioned study on the effect of discrete reflections on sound zone performance by Olik et al. [12], numerical predictions showed that a single reflection could be accounted for in the ACC process, yielding higher contrast to when this reflection was ignored, but even in this idealised numerical study, including the effect of higher order reflections did not significantly affect the contrast performance or array effort [12]. In another study by the same authors, in this case using measured reverberant responses, it was necessary to truncate and process the late reverberation present in the measurements in order to improve robustness to the mismatch between setup and playback conditions [60]. Both of these results suggest that including room reflections beyond the first order can be detrimental to performance and reliability.

This is corroborated by Poletti et al., who state that reducing the influence of room reflections is possible, but active compensation for these reflections requires an in-situ measurement [235].

Finally, visual inspection of the modelled and measured impulse responses shown in Figure 8.4 reveals some significant differences in individual source directivity and frequency response. This indicates that if modelled reverberant transfer responses are used to design zoning filters, and performance is evaluated using measured responses, in effect simulating playback in a real room, the resulting mismatch between the setup and playback conditions will lead to a reduction in acoustic contrast. Therefore, in this section, results are reported for only two different approaches to calculating the zoning filters. In the first case, the geometric positions of the microphones and loudspeakers are used to simulate the analytical transfer responses. A monopole model is used for the loudspeakers, and the microphones are assumed to be compact omnidirectional sensors in a free-field. In the second instance, measured transfer responses are used to design the filters, and these are truncated after 0.1 seconds to exclude low level measurement noise after the end of the reverberant decay.

As described in Equation 5.11, in the ACC filter design process, the matrix $[\mathbf{Z}_d^H \mathbf{Z}_d]$ must be inverted. The close proximity of the loudspeakers and microphones within the source and sensor arrays can cause $[\mathbf{Z}_d^H \mathbf{Z}_d]$ to be ill-conditioned. Regularisation reduces the condition number of this matrix to improve numerical stability and reduce the array effort, but has additional audible effects, such as flattening the frequency response in the bright zone and reducing the achievable level of acoustic contrast. To selectively increase the regularisation at frequencies where the solution is poorly conditioned, the regularisation parameter at each frequency in the ACC process is given by $\beta_0 \kappa$, where κ is the condition number of $[\mathbf{Z}_d^H \mathbf{Z}_d]$. The proportionality constant β_0 is varied to provide a range of levels of regularisation, and the effects of varying this constant on acoustic contrast and speech intelligibility are discussed in Sections 8.4.1 and 8.4.2.

In order to maintain consistency between bright zone signals under different regularisation conditions, a ⅓-octave band equaliser is applied to the programme signal to equalise the programme in the bright zone to the transfer response magnitude of a single loudspeaker, shown by the blue line in the lower panel of Figure 8.4. The use of a ⅓-octave band equaliser is a simplistic means to avoid over-equalisation of narrow bands, which may lead to poor robustness and long-ringing resonances [236]. More advanced strategies for providing system equalisation exist, such as complex smoothing [237], which better preserves time-domain information such as transients compared to fractional-octave smoothing, and frequency-warped filters [238] which reduce the emphasis on equalisation of perceptually irrelevant high frequency resonances and antiresonances. The perceptual effects of the choice of equalisation carried out here may be considered in future work, but is expected to be less significant to speech intelligibility and reproduction quality than other features of the system, such as the introduction of masking noise.

Following equalisation, the input gain is set such that the spatially averaged SPL of the programme in the bright zone reaches 60 dB SPL. As described in the previous section, and following the process described in Chapter 7, when a masking signal is included to provide privacy for the target listener, the level of the masking signal is adjusted iteratively to give an SII level of 0.05, averaged across microphones in the dark zone. In the previous section, the regularisation parameter in the ACC process was selected to maximise the speech intelligibility in the bright zone,

but in the following two sections, results are presented for a range of regularisation parameters to illustrate the sensitivity to setting this parameter correctly.

8.4.1 Effect of Regularisation on Acoustic Contrast

As the level of regularisation is varied, the ACC process is affected in several ways. As can be seen in Equation 5.11, the effect of adding regularisation weights the diagonal of the squared transfer response matrix, $[\mathbf{Z}_d^H \mathbf{Z}_d]$, and this introduces a mismatch between the transfer responses used to design the ACC filters and the true transfer responses between the loudspeaker array and the zones. This mismatch can affect the achievable level of acoustic contrast. As previously shown in Equation 5.5, the acoustic contrast is defined as the ratio of the mean squared pressure in the bright and dark zones, and is calculated at each frequency as

$$C = 10\log_{10}\left(\frac{N_d \mathbf{q}^H \mathbf{Z}_b^H \mathbf{Z}_b \mathbf{q}}{N_b \mathbf{q}^H \mathbf{Z}_d^H \mathbf{Z}_d \mathbf{q}}\right). \tag{8.7}$$

where N_d and N_b are the number of microphones in the dark and bright zones respectively. In the examples in this chapter, $N_b = N_d = 10$.

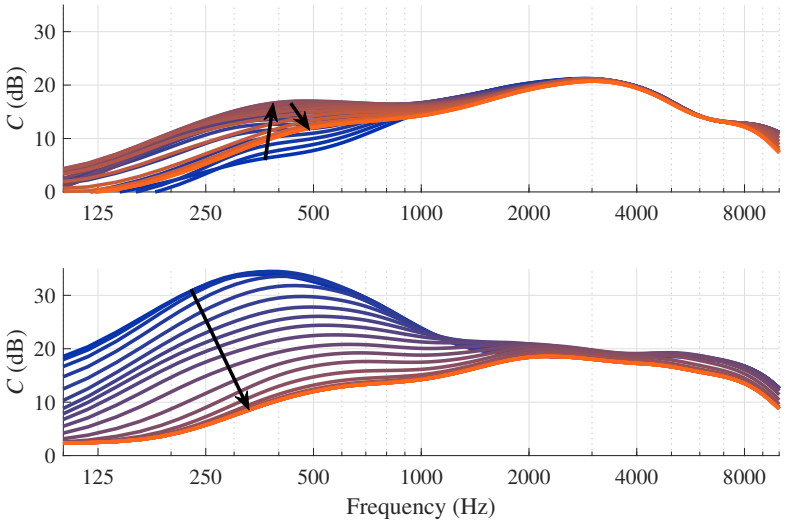


FIGURE 8.13: Acoustic contrast for the ACC process that focuses speech into the bright zone, with an increase in the regularisation parameter β_0 from 10^{-28} to 10^{-16} indicated by colour change from blue to orange. Arrows indicate increasing regularisation. Acoustic contrast results presented in the upper plot use ACC filters based on analytical transfer responses, and for the lower plot, these are based on measured responses.

To demonstrate the effect of adding regularisation to the system, whose geometry is shown in Figure 8.6, Figure 8.13 shows the variation in measured acoustic contrast when analytical (upper panel) and measured (lower panel) transfer responses are used in the filter design process for different levels of regularisation. In both cases, performance is evaluated in the measured reverberant environment of the ISVR audio laboratory. The change in the regularisation proportionality constant β_0 from a value of 10^{-28} to 10^{-16} is indicated using a gradual change in the line colour from blue to orange.

The lower panel in Figure 8.13 shows the variation in acoustic contrast with different levels of regularisation when the ACC filters are calculated using measured transfer responses. At regularisation levels below $\beta_0 = 10^{-28}$, the transfer responses for which the filters are optimised match the reproduction environment closely, so a high level of acoustic contrast is achieved. Exact matching is prevented through the use of separate optimisation and evaluation microphones, as shown in Figure 8.6. As regularisation increases, a mismatch between the responses assumed in the filter calculation and the actual response in the room is introduced, so the acoustic contrast decreases, particularly at frequencies below 1 kHz where the acoustic wavelength is comparable with the size of the zones. Above $\beta_0 = 10^{-16}$, this regularisation term dominates $[\mathbf{Z}_d^H \mathbf{Z}_d]$, reducing the ACC to the simpler brightness control method [14], that is, maximising the level in the bright zone alone. This can be confirmed by noting the similarity between the orange lines in the upper and lower panels of Figure 8.13.

From the upper panel in Figure 8.13, which shows the performance achieved when using analytical transfer responses to calculate the ACC filters, it can be seen that the maximum level of contrast is lower than that achieved by the filters designed using the measured responses, regardless of the regularisation level. This difference in the upper performance limit is because, while the analytical responses only contain the time of arrival of the direct sound from the array, the measured transfer responses contain information about the time of arrival of both direct sound and early reflections in the impulse response, as shown in Figure 8.4, as well as a component due to diffuse reverberation. Additionally, the measured transfer responses inherently contain the specific directivity information, frequency response and sensitivity of each driver in the array, thereby compensating for any differences between the drivers that is not encoded by the monopole model.

The upper panel of Figure 8.13 also shows that when β_0 is low, the acoustic contrast is low compared to the corresponding results presented in the lower panel that were calculated using the measured responses, particularly in low- to mid-frequencies. This is due to the more significant mismatch between the monopole responses used in the filter design process and the physical room responses. As the system is regularised, the effect is equivalent to adding a random component to the analytical transfer responses [11, 58], which approximates the diffuse reverberation component in the measured responses. The acoustic contrast achieved by the filters designed using the analytical responses is maximised when the level of regularisation is sufficient to ensure robustness to the difference between the analytical and physical responses. For this example, this occurs at a regularisation level corresponding to $\beta_0 = 10^{-22}$. Continuing to increase regularisation beyond this point results in a transition to brightness control, which offers less acoustic contrast, as seen in the case where measured responses are employed in the ACC process.

8.4.2 Effect of Regularisation on Speech Intelligibility

It has been shown in the previous section that the system optimised using the measured responses is capable of achieving a higher level of acoustic contrast than when the analytical responses are used, provided that the regularisation parameter is chosen correctly. However, of primary importance to the design of personal audio systems for the reproduction of private speech content is the level of speech intelligibility contrast between the bright and dark zones. As described in Section 8.3, the simulated playback from the array comprises of speech and speech-shaped

noise, reproduced at a level that achieves SII = 0.05 in the dark zone. Figure 8.14 shows the corresponding SII in the bright zone, averaged across the ten evaluation microphones in the zone, with different levels of regularisation used in the ACC process. To maximise the bright zone intelligibility according to this metric, the zonal filters must effectively reduce the cross-talk between zones, so that the average SNR in each speech frequency band is maximised. To supplement this intelligibility evaluation, the ESTOI metric [121], described in Section 3.1.3, is also used to assess how the temporal structure of the speech programme material is affected by the sound zoning process with different levels of regularisation, as the SII metric is unable to provide this information. These corresponding ESTOI evaluations in the bright zone are presented in Figure 8.15.

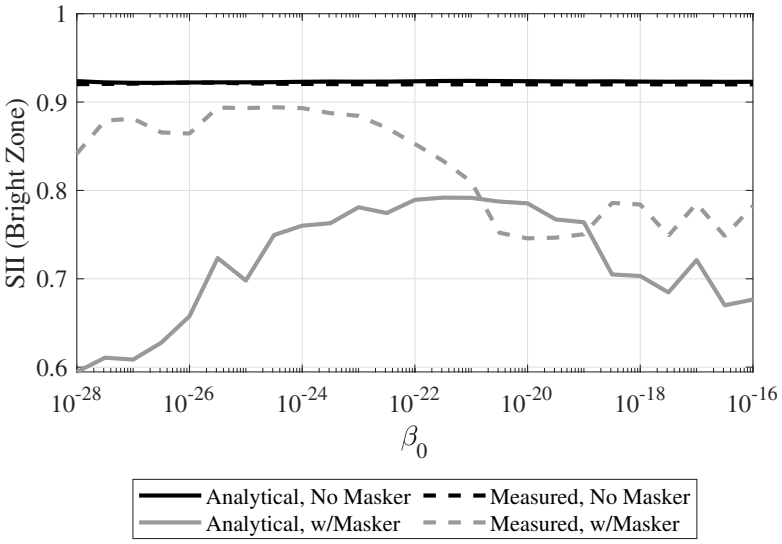


FIGURE 8.14: SII measured in the bright zone for different levels of the regularisation proportionality constant β_0, for systems using analytical and measured transfer responses in the ACC process, both with and without an additional masking signal.

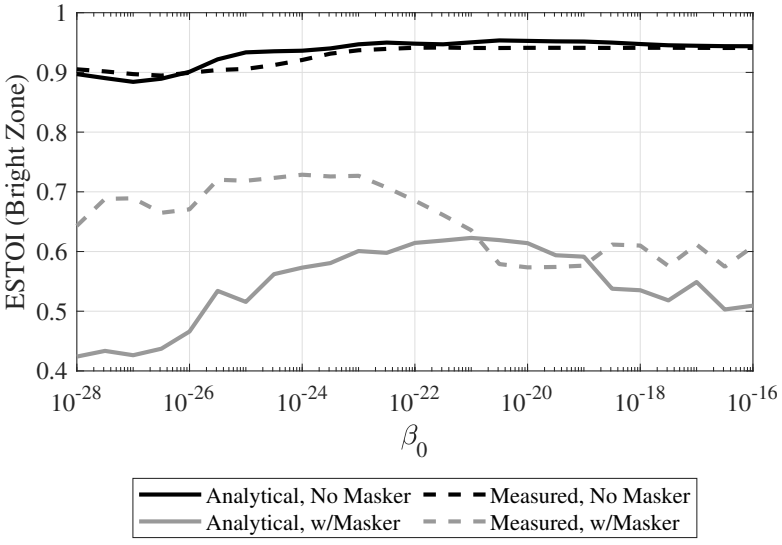


FIGURE 8.15: ESTOI measured in the bright zone for different levels of the regularisation proportionality constant β_0, for systems using analytical and measured transfer responses in the ACC process, both with and without an additional masking signal.

Firstly, the performance of the system is considered while operating in quiet, with no masking signal. These ideal conditions give the upper limit for the bright zone intelligibility when each type of response is used in the filter design process. The black traces in Figure 8.14 show that when using both analytical responses (solid line) and measured responses (dashed line), the SII remains essentially constant at a level of 0.95, which corresponds with excellent intelligibility across many types of speech material, as shown in Figure 3.7. This high SII value verifies that the equalisation applied to the programme signal, to account for the non-flat frequency response of the loudspeaker array and the ACC filters, is effective. When the same conditions are evaluated using the ESTOI metric, shown with the black traces in Figure 8.15, the trend is similar with an ESTOI value in excess of 0.9, though there is a gradual increase in ESTOI with β_0. The ESTOI value predicted in the modelled listening room, shown in Figure 8.9, also approaches this value. With no additional noise in the environment to degrade intelligibility, the distortion to the speech signal observed when low levels of regularisation are used must be attributed to the temporal structure of the filters themselves.

To illustrate how regularisation affects the time-domain response of the ACC filters, Figure 8.16 shows the impulse response of four ACC filters; the left column shows a case with a low level of regularisation, $\beta_0 = 10^{-28}$, for each filter type, and the right column shows the filter responses when a higher level of regularisation, $\beta_0 = 10^{-16}$ is used. From this figure it can be seen that the length and complexity of the filters based on measured responses is significantly higher than those made using analytical responses, due to the additional information regarding the room reverberation that is contained within the measurements. Comparisons within each filter type, between the different levels of regularisation, show that when either analytical or measured transfer responses are used, additional regularisation results in simpler, effectively shorter filters with lower levels of noise away from the main impulse. This in turn results in less degradation to the fidelity of the input speech signal, thus improving the ESTOI rating.

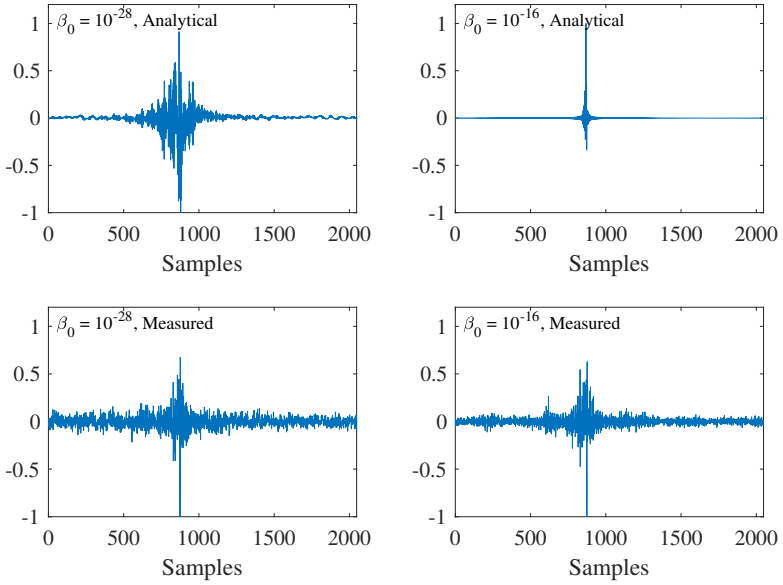


FIGURE 8.16: Impulse responses of filters from high and low regularisation conditions, using analytical and measured transfer responses.

An additional cause of the small degradation to the intelligibility at low levels of regularisation may also be attributed to the additional equalisation required when low levels of regularisation

are used. Figure 8.17 shows the ⅓-octave band gain that is pre-applied to the input programme and masker signals to account for the non-flat frequency response of the zoning filters. A colour change from blue to orange indicates an increase in β_0, using the same colour scheme used in Figure 8.13. This indicates that the equalisation required is most significant when regularisation is low. The minor loss of intelligibility at low regularisation levels visible in Figure 8.15 could be caused by the amplification of low-level noise in the input speech recordings. This would reduce the effective modulation depth of the input speech signals, and hence the ESTOI evaluation.

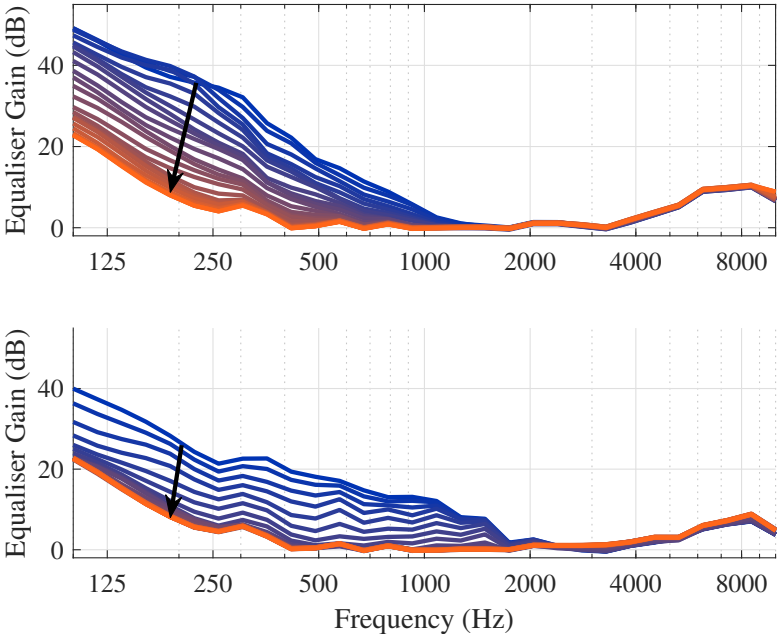


FIGURE 8.17: ⅓-octave band gain setting required to equalise the programme material in the bright zone at different levels of regularisation. Upper panel: Filters generated using analytical transfer responses, Lower panel: Filters generated using measured transfer responses. Colour change from blue to orange indicates an increase in the regularisation proportionality constant, β_0, from 10^{-28} to 10^{-16}.

It is important to note that the intelligibility loss due to the effects of temporal degradation and loss of headroom is marginal, certainly when compared to the case where a masking signal is introduced into the dark zone. This case is shown using grey lines in Figures 8.14 and 8.15 for the SII and ESTOI intelligibility metrics respectively. The grey traces in Figure 8.14 show how the SII rating in the bright zone depends on β_0 when a masking signal is directed into the dark zone, and adjusted in level to provide SII = 0.05, spatially averaged across the evaluation microphones. Figure 8.18 verifies that the SII target of 0.05 is achieved to within ± 0.02 units - this small random variation is due to the adaptive level adjustment of the masking signal in 1 dB, i.e. just noticeable, steps [79]. The lower panel of Figure 8.18 also shows the corresponding ESTOI values recorded in the dark zone, which average to a value of 0.1. This value was used in Section 8.3 as the dark zone intelligibility constraint for more reverberant spaces, due to the better suitability of the ESTOI metric for predicting the degradation in intelligibility in these acoustical conditions, compared to SII.

Leakage of the masker into the bright zone decreases the intelligibility level compared to the case where no masker is present. This difference is most significant for the filters designed using the analytical transfer responses, due to the lower overall level of acoustic contrast achievable

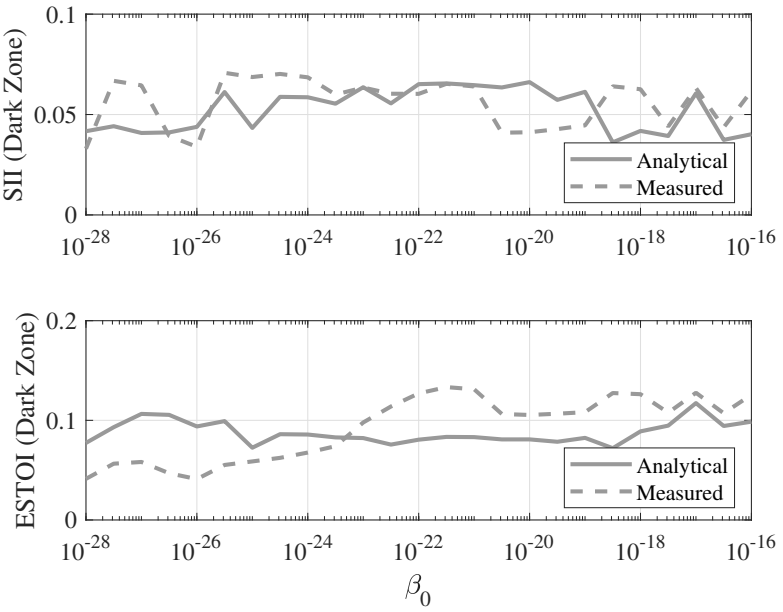


FIGURE 8.18: Speech intelligibility in the dark zone of the system according to the SII and ESTOI metrics, with variation in the regularisation parameter β_0.

using this approach, as shown in Figure 8.13. The lower panel of Figure 8.13 shows that when measured transfer responses are used to design the ACC filters, an increase in the regularisation corresponds to a reduction in the acoustic contrast. The dashed grey trace in Figure 8.14 shows that this trend is mirrored in the bright zone SII. As acoustic contrast decreases, the SII in the bright zone also decreases as the masking signal level must be increased to maintain an SII level of 0.05 in the dark zone. The corresponding increased leakage of the masker into the bright zone results in a reduction in the intelligibility at the target listener position. The same is true when analytical transfer responses are used to generate the ACC filters; in this case, a maximum level of acoustic contrast is reached around $\beta_0 = 10^{-22}$. At this value, the regularisation is sufficient to provide robustness to the inherent mismatch between the analytical transfer responses and the true transfer responses in the room. Levels of regularisation both higher and lower than this value result in poorer intelligibility in the bright zone. These trends are also visible when the ESTOI metric is used to assess intelligibility in Figure 8.15, confirming that the effect of additional masking noise makes the most significant detriment to intelligibility, compared to the other effects described above.

For the example system described in this section, the SII value in the bright zone exceeds 0.6 across the entire tested range of regularisation parameters, regardless of the use of analytical or measured transfer responses in the sound zoning process. When this SII value is matched to an intelligibility score using the AITFs presented in Figure 3.7, this suggests that most types of speech material will be 100% intelligible in the bright zone. A score around 75% correct is predicted for the significantly more challenging task of identifying nonsense syllables under the same conditions. Furthermore, when analytical responses are used, SII values greater than 0.75, the benchmark level for "good communication systems" as described in the SII standard [124], can be achieved over a range of β_0 that spans five orders of magnitude, indicating relative insensitivity to the choice of regularisation parameter.

8.4.3 Masking Signal Levels

Throughout this thesis, one of the key elements of practical personal audio system design has been to consider the experience of both the target listener and others nearby. Therefore, it is important to assess the masking signal levels which are required to meet the zonal intelligibility constraints placed on a given system. Figure 8.19 shows the SPL of the masker in the dark zone at a range of system regularisation levels. In this figure, the values of β_0 that require the minimal masking signal level in the dark zone are marked with arrows. These regularisation conditions correspond to the respective maxima in the bright zone SII that are recorded in Figure 8.14, further highlighting the strong correspondence between masking signal levels, acoustic contrast and speech intelligibility that is discussed elsewhere in this thesis.

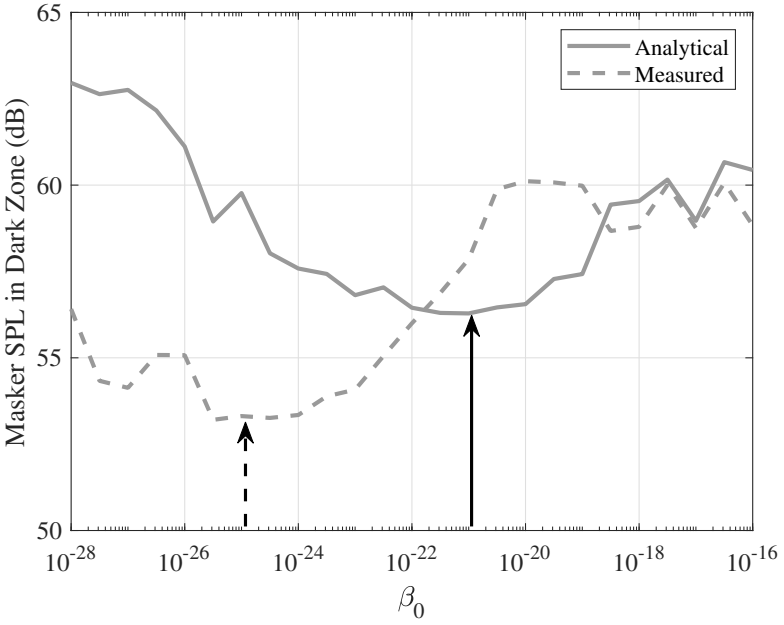


FIGURE 8.19: SPL of the masking signal measured in the dark zone, with variation in system regularisation, for systems using analytical and measured transfer responses to produce sound zoning filters. Arrows indicate the regularisation levels that correspond with the maximum bright zone intelligibility, as shown in Figure 8.14.

The minimum masker level required when measured responses are used is 3 dB lower than the corresponding masker level required with analytical responses. This corresponds to a 9% decrease in the psychoacoustic annoyance value measured in the dark zone. This decreased dependency on the masking signal level to provide adequate privacy demonstrates the main advantage of using measured responses over analytical responses, particularly in system designs where high masker levels are anticipated due to low levels of acoustic contrast. Systems using measured transfer responses may also be preferred in installations where ambient noise levels are low, as these conditions may emphasise the intrusiveness of the artificial masker. The wider effects of the ambient noise in the reproduction environment are discussed further in Chapter 9.

For the simulated room where the mid-frequency reverberation time, $T_{60,mf}$, is 0.11 seconds, matching that of the ISVR audio laboratory, the difference in the required masker level between the cases where free-field and reverberant responses are used is 7 dB. The equivalent comparison between analytical and measured responses, as shown in Figure 8.19, revealed only a 3 dB difference between approaches at the regularisation levels that maximise the intelligibility in the

bright zone. Therefore, in the simulated room, a greater benefit to including reflections in the transfer responses was predicted, compared to in the measured room. This can be attributed to the assumption made in the image source model that the loudspeaker array is composed of omnidirectional sources. The linear array configuration only offers control in the horizontal plane, with omnidirectional directivity in the vertical direction, so this maximally excites the reverberant field in the room [11]. Therefore, incorporating the reflections into the sound zoning process carries a greater benefit in the simulated room, compared to the case where measured responses are used to optimise and evaluate the system performance. An additional consideration to explain the differences between measured and modelled results is that the measured transfer responses will naturally contain low levels of measurement noise, which will decrease their performance when used in sound zoning filters, compared to the reverberant transfer responses calculated using the image source method. Accordingly, the intelligibility results presented in Figure 8.10 for more reverberant spaces can be regarded as a lower bound. Higher bright zone intelligibility and lower required masker levels can be expected if source directivity is included in the loudspeaker array model used to calculate the signals received in each zone.

8.5 Further Considerations

In the studies described in this chapter, the degree of reverberation in a given reproduction environment has been characterised by the reverberation time, either in octave bands or as a single, frequency-averaged value. However, it is potentially misleading to reduce the complex effect of reverberation to a single number as the arrival time and relative strength of early and late reflections is also dependent on the size of the room, the distribution of absorption and the directivity of the source providing excitation. This is evidenced by the difference between measured and modelled room impulse responses in Figure 8.4 which have identical reverberation times. This discrepancy makes comparisons between simulations and measurements of the same space difficult, and points towards the necessity of integrating source directivity into the room acoustic simulation, as described in Section 8.4.3. The simple image source model used in this chapter is ideally suited to rooms with simple surface geometries, due to its simplistic handling of specular reflections. Analysis of larger, more complex spaces, integration of source directivity and finer control of frequency dependent absorption and scattering may be facilitated through the use of more advanced ray-tracing algorithms such as those provided by *CATT Acoustic* [239].

An increase in the fidelity of reverberant simulations would also allow for additional experimentation into the benefit of early reflections for intelligibility. Room acoustics indices used for the assessment of speech transmission in rooms such as Clarity, C_{50}, and Definition, D_{50}, are increased by the presence of early reflections, defined as energy arriving no more than 50 ms after the direct sound [36]. The ESTOI intelligibility metric is principally designed for assessing the degradation of processed or noisy speech, and as such has no explicit provision for recognising the potential enhancement of intelligibility caused by early reflections. As described in the block diagram in Figure 3.4, the ESTOI algorithm divides reference and degraded signals into 384 ms windows, which are then converted into spectrograms with frame length equal to 25.6 ms before normalisation and comparison. The effect of the 25.6 ms frame length means that early reflections arriving before 25 ms are essentially counted as direct sound, and reflections arriving from 25-50 ms only adversely affect a single spectrogram frame, and thus any negative effect that the

ESTOI algorithm may otherwise attribute to these early reflections is limited. Nevertheless, the explicit inclusion of metrics that fully account for the benefits of early reflections may provide a more complete assessment of personal audio system performance in reverberant spaces.

In addition to these temporal effects of reverberation, spatial effects must also be considered. For the system geometries discussed in this thesis, zones are specified on a horizontal plane coincident with the long axis of the linear loudspeaker array elements. This linear arrangement allows control over the horizontal directivity, but the directivity of the array in the vertical orientation is only limited by the directivity of the individual loudspeaker array elements. As the diffuse, reverberant field in the reproduction environment is built up from multiple reflections from all the surfaces that bound the room, the uncontrolled vertical radiation from the array can result in significant energisation of the reverberant field [11], which in turn can reduce acoustic contrast, and potentially increase the leakage of speech from the bright zone into the dark zone, necessitating an increase in the masking signal level. An open-source design for a loudspeaker array with enhanced vertical directivity [240], intended to counteract this effect, is presented in Appendix C. The diffuse, three-dimensional nature of reverberation can also affect how the direct and reflected sound from the array is perceived; speech signals are not as effectively masked when the interfering noise originates from a different direction, in a phenomenon known as the spatial release from masking [241, 242]. This effect is described in the following chapter, alongside a broader discussion of the effects of diffuse and directional background noise on personal audio system performance.

8.6 Summary

In this chapter, a range of physical and simulated experiments have been documented, with the aim of characterising the performance of private personal audio systems in reverberant spaces. These experiments have compared two approaches to selecting the format of the transfer responses that are used in the sound zoning process; analytical transfer responses are simple to calculate but neglect the effects of reflected sound in the reproduction environment, whereas measured transfer responses do account for room reflections, but are more complex to acquire. The results have demonstrated that sound zoning performance can be significantly impeded by room reverberation.

When evaluated in terms of the maximum achievable acoustic contrast alone, sound zoning filters derived from measured data perform better due to the close matching between the optimisation and the playback environments. This mirrors previous findings in the literature [72], but for the context investigated in this thesis, any gains in performance are compounded as the leakage of both the speech and masking signals are reduced, meaning that the masker can be doubly attenuated.

Measurements and simulations of personal audio systems designed to produce a private speech zone have been carried out in reverberant spaces, in order to quantify the deterioration to acoustic contrast and speech intelligibility caused by reverberation. In general, the higher the reverberation time, the worse the performance, as the desired difference in sound level between the bright and dark zones becomes compromised by the diffuse reverberation. The relative energy of the direct and diffuse components of the sound field can be expressed as a function of the distance

from the array, and analysis of the critical distance within the reproduction environment can be used to place limits on where listening zones may be formed. As the reproduction environment becomes more strongly reverberant, the critical distance decreases and the advantage of incorporating knowledge of early reflections into the sound zoning process becomes more significant.

When a masking signal is radiated into the dark zone to preserve the privacy of the target listener, the level of the masker can be adjusted using the process documented in Chapter 7. The results in Section 8.4 show that systems which attain higher levels of acoustic contrast require lower masker levels to achieve the same level of privacy, compared to when analytical transfer responses are used. This implies that in installations that are particularly sensitive to noise annoyance, the use of measured transfer responses are recommended. However, in situations where a slight increase in masker level can be tolerated, for example where background noise levels are already comparable to the masking signal level, zoning filters based on regularised free-field responses are preferred. This approach can offer equivalent privacy, and comparable levels of bright zone intelligibility, alongside the obvious advantages of simplicity in implementation, robustness to changes in room reverberation and cost-effectiveness in production when compared to measured responses. The results presented in Section 8.3 caveat this conclusion by showing that in more reverberant spaces, optimising a personal audio system using free-field transfer responses can result in significantly lower levels of bright zone intelligibility to when responses that include the reverberation are used in the sound zoning process. However, these simulations represent a worst-case scenario in terms of the assumed source directivity, as the reverberant field in the room is maximally excited by the omnidirectional sources used in the loudspeaker array model. Methods to control the vertical directivity of a loudspeaker array, and hence reduce the excitation of the reverberant field, have been discussed and are presented in Appendix C.

It is clear from the results presented in this chapter that reverberation can negatively affect the performance of personal audio systems, but this is not the only acoustic feature that can place limitations on the acoustic contrast and speech intelligibility contrast between zones. Background noise in the reproduction environment can also impede intelligibility in both the bright and dark zones, and unlike reverberation, can vary significantly depending on the occupancy of the space and the time of day. In the following chapter, a discussion of how background noise can affect personal audio system performance is presented, including how the ambient background noise in a space may potentially be harnessed to reduce the masking requirements of personal audio systems.

Chapter 9

Combining Artificial Masking and Ambient Noise in Personal Audio System Design

Parts of this chapter are based on "Combining Artificial and Natural Background Noise in Personal Audio Systems" which has been published in the proceedings of the 10^{th} IEEE Sensor Array and Multichannel Signal Processing Workshop.

The results from the previous chapter show that reverberation can be detrimental to the performance of personal audio systems used for private speech transmission. The maximum speech intelligibility difference between zones was reduced in more reverberant spaces, when the programme level was kept constant, due to the reduction in acoustic contrast associated with the reverberation. However, reverberation is not the only anticipated source of performance degradation in practical systems. The ambient noise in the environment where a system is installed will decrease the speech intelligibility in both the bright zone and the dark zone. In the bright zone, this additional noise is detrimental to performance as the programme level, or directivity of the array, must be increased to maintain a high degree of intelligibility in the bright zone. In the dark zone, however, the presence of ambient noise can be advantageous, as it may be possible to decrease the level of the additional masking noise provided by the system, thus reducing the potential for the system to be perceived as annoying by nearby listeners. A block diagram of a speech privacy control system installed in a noisy environment is presented in Figure 9.1. In the previous investigations presented in this thesis, systems were assumed to be operating in quiet, meaning that the speech level could be held constant. In this chapter, the inclusion of ambient noise in system simulations means that the levels of both the speech programme and masking signal must be adjustable.

In this chapter, a distinction is made between the *ambient noise* and the *background noise* in a given environment, using the convention described in British Standard BS 4142:2014+A1:2019 [243]. This standard concerns the measurement and rating of industrial and commercial sound and usually requires measurements of a specific sound source to be corrected based on the level of other sound sources in the environment. The ambient sound level, $L_{Aeq,T}$, is described as the

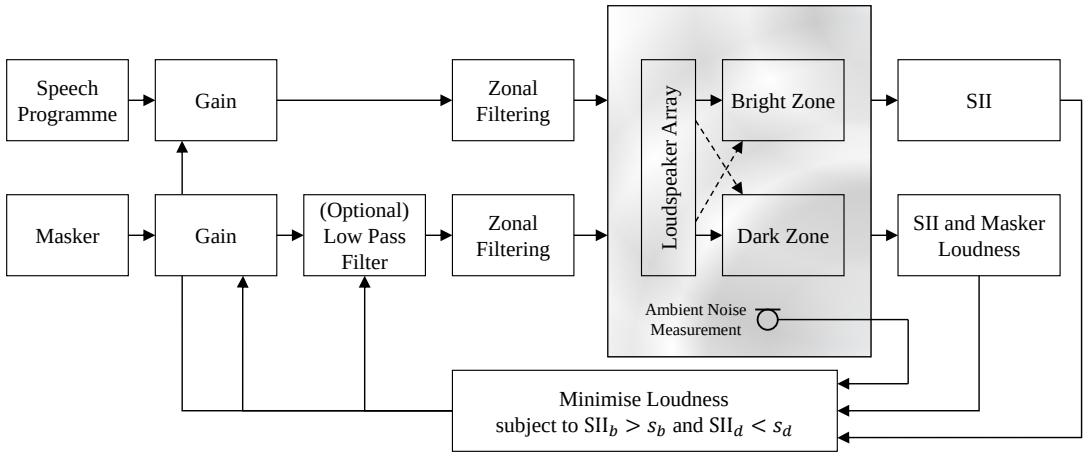


FIGURE 9.1: Block diagram of a personal audio system operating in a noisy environment. In this chapter, a range of bright and dark zone speech intelligibility constraints, s_b and s_d, are considered.

"equivalent continuous A-weighted sound pressure level of the totally encompassing sound in a given situation at a given time, usually from many sources near and far, at the assessment location over a given time interval, T" [243]. The background sound level, $L_{A90,T}$, is the "A-weighted sound pressure level that is exceeded by the residual [ambient] sound at the assessment location for 90% of a given time interval, T" [243]. The background noise level measured in a space therefore excludes contributions to the ambient sound that are intermittent, and thus are less effective at masking continuous speech.

The present chapter begins with a simplified investigation into the effects of ambient environmental noise on the performance of speech privacy control systems. In these initial simulations, a constant level of speech-shaped ambient noise is assumed. Firstly, the optimal combination of speech and masking noise levels that achieve a given pair of intelligibility constraints in the bright and dark zones are found, as shown in Figure 9.1. This combination of signal levels is specific to the amount of acoustic contrast provided by each system, so to generalise the results, the process is repeated for a range of different levels of acoustic contrast performance, simulating different array sizes and geometries. This provides information on the design requirements for speech privacy control systems, ultimately showing that systems that use a combination of artificial masking and ambient noise have lower acoustic contrast requirements than those that rely on the ambient noise alone.

These conclusions are developed in the sections that follow, with a particular focus on the range of applicability of the simplifying assumptions made in Section 9.1. Various practical issues that are anticipated when implementing a personal audio system for speech transmission in a noisy environment are highlighted. These include the perceptual effect of Spatial Release from Masking (SRM), and both short- and long-term temporal variation in the composition of the ambient noise.

9.1 The Effects of Steady Ambient Noise

The investigations in this section quantify the advantages of radiating an artificial masker into the dark zone of a speech privacy control system, compared to solely relying on a single sound zoning process and the masking effect of the ambient noise in a space. In order to provide flexibility and increase computational efficiency in simulating systems with a range of acoustic contrast profiles, a surrogate model is introduced. In this model, the signals received in the bright and dark zones are simulated by applying equalisation and gain reduction to the input speech and noise signals, depending on their intended location. Once an optimal combination of signal levels is found, the accuracy of the surrogate model can be tested using a single full array simulation.

Figure 9.2 shows the simulated acoustic contrast characteristic that is applied to the input signals, in order to form the signals within each zone, for the initial investigations in this section. This curve models the acoustic contrast measured when symmetrical sound zones are placed 0.42 metres apart, and 0.5 metres away from the 27-channel loudspeaker array described in Section 5.1.1, in the ISVR audio laboratory. The frequency-dependent trends shown in this figure match those achieved with a range of linear loudspeaker arrays designed for speech privacy control (see e.g. Figures 5.12, 5.14, 5.18), with a gradual rise in acoustic contrast from low frequencies, to a maximum around 2 kHz, and a decrease at high frequencies corresponding to spatial aliasing limitations.

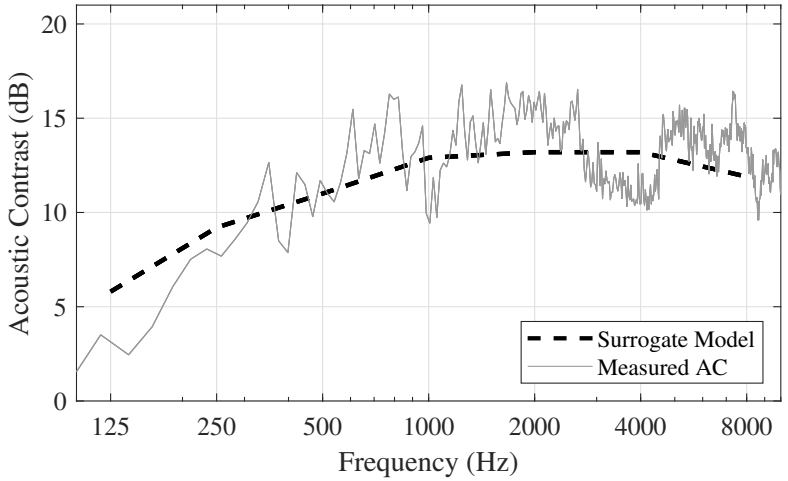


FIGURE 9.2: Acoustic contrast simulated using the surrogate equalisation model, and measured using the 27-channel array in the ISVR audio laboratory.

Specifying the acoustic contrast directly using the surrogate model allows for more precise control of the performance of the simulated systems, so that the effects of changing the acoustic contrast can be explored in detail. The alternative approach, of using full array simulations, offers less direct control over the acoustic contrast as changes to the system must be made in discrete steps, such as by adding or removing loudspeaker drivers, or by iteratively adjusting the regularisation. Furthermore, using a particular loudspeaker array geometry or reproduction environment necessarily narrows the applicability of the conclusions that can be drawn from the simulations. Perhaps the main benefit to the described approach is that there is a significant computational advantage to simulating zonal signals using equalisation. Generating ACC filters

and computing zonal signals using multichannel convolution, i.e. a full array simulation, takes an average of 3.03 seconds to complete on a desktop workstation, whereas the surrogate model simulates the equivalent set of signals in 0.012 seconds. Throughout this chapter, an optimisation algorithm is used to determine the optimal programme and masker signal levels for a given combination of ambient noise, acoustic contrast performance and speech intelligibility constraints. This optimisation can potentially require a large number of trials, so computational efficiency is paramount.

Signals in the bright and dark zones are simulated by summing the contributions of the programme signal, the masking signal and the ambient noise. In the simulations detailed in this chapter, the programme material consists of a concatenation of multiple sentences from the VCTK corpus [22], spoken by various male and female talkers. The masking signal is speech-shaped noise, formed based on the long-term spectral shape of the programme, and in this section the ambient noise is formed of 8-talker babble noise, also generated from the VCTK corpus with a balance of male and female talkers. The long-term average spectrum of the ambient noise is therefore matched to that of the programme and artificial masking signals - the effect of this will be discussed in Section 9.2.

9.1.1 Selection of Optimal Programme and Masker Signal Levels

To gain an understanding of how the levels of the programme and masking signals affect the intelligibility of speech in the bright and dark zones, contour plots of the SII in each zone have been constructed, and are presented in Figure 9.3. Each point on each contour plot in Figure 9.3 represents the results of a single simulation from the surrogate model. This is similar to the process used to construct the contour plots presented earlier in Figure 7.2, except these plots were constructed using full array simulations. The upper panel of Figure 9.3 shows how the SII measured in the bright zone is affected by variation in the programme and masker levels, with a constant ambient noise level of 60 dBA. In the bright zone, the contours show that the most dominant signal is the speech programme, as intelligibility increases most significantly as the programme material increases in level. For a given programme level, increasing the masker level causes a slight reduction in the intelligibility, due to increased leakage of this signal into the bright zone. This effect is most significant when the level of the masker in the dark zone is significantly greater than the ambient noise level. The marked contours can be regarded as minimal speech intelligibility constraints for the bright zone - any combination of programme and masker levels above the marked contours results in an acceptable bright zone sound field, according to the particular constraint used. The middle panel of Figure 9.3 shows similar trends to the upper panel, with regard to the effects of increasing the programme and masker levels, although in this case, the contours and plot colouring refer to the SII in the dark zone. In order to provide privacy for the target listener, the intelligibility of the leaked speech in the dark zone must be minimised, so all combinations of programme and masker levels *below* a given contour in this panel provide a level of unintelligibility, i.e. privacy, in excess of the selected constraint value.

The lower panel of Figure 9.3 shows that when the contours from the upper and middle panels of Figure 9.3 are overlaid, regions can be constructed that satisfy both sets of constraints. At the intersection of a given pair of constraints, indicated with red circles, the signal levels in each

zone are minimised. This combination results in the lowest overall signal levels that can achieve the given constraints - most significantly, this combination minimises the level of the required masking signal in the dark zone, which has been shown in Chapter 7 to have a significant bearing on the acceptability of a speech privacy control system. The positions of the intersection points in the parameter space show how the choice of intelligibility constraints has a significant effect on the optimal programme and masker signals used by a system. The least onerous constraint presented in Figure 9.3 corresponds to the situation where the intelligibility in the bright zone, SII_b, must exceed 0.5, and the intelligibility in the dark zone, SII_d, must be evaluated as less than 0.1. In this case, the programme must be reproduced at 64 dBA to sufficiently overcome the masking effect of the ambient noise, and the masking signal level should be set to 57.5 dBA in the dark zone. This level is 2.5 dB less than the assumed ambient noise level, so while the additional, artificial masking signal will be perceivable, the ambient noise remains the most significant masker of the leaked speech in this case.

Further insight into the programme and masking signal requirements in the presence of ambient noise can be gained by observing the optimal signal levels when the dark zone constraint is reduced to $\mathrm{SII}_d < 0.05$, the level found in Section 7.1.1 to correspond to 50% words correct in the matrix sentence test. In the bright zone, the required programme level remains unchanged at 64 dBA, demonstrating that even with an increase in the masking signal level, the ambient noise remains the dominant source of masking within the bright zone. Meanwhile in the dark zone, the heightened level of privacy demanded by the SII_d constraint requires that the masking signal must be increased in level by 4.5 dB to 62 dBA. In this case, where the incoherent masking signal and ambient noise have similar spectra, their respective broadband SPLs can be summed to determine the overall level. The tightening of the dark zone intelligibility constraint from $\mathrm{SII}_d < 0.1$ to $\mathrm{SII}_d < 0.05$ causes the overall A-weighted SPL in the dark zone to increase from 61.9 dBA to 64.1 dBA, and this shifts the balance of the signals received in the dark zone to be dominated by the artificial masking, rendering this signal potentially more noticeable [244]. Additionally, the simple monophonic evaluation presented here does not account for the perceptual effects of the dominant source of masking originating from the direction of the array, rather than being diffuse - these effects are discussed in Section 9.3.

When the SII constraint in the bright zone is increased to a value of 0.6, and the dark zone constraint is maintained at $\mathrm{SII}_d = 0.05$, the required signal levels in the bright and dark zones are both significantly in excess of the ambient noise level, at 68.5 and 69.5 dBA respectively. In this operating regime, the system makes full use of the ability to selectively raise the speech and noise levels in a confined spatial region, and relies very little on the effects of the ambient noise present in each listening zone. It will be shown in the following section that certain combinations of intelligibility constraints are impossible to achieve for a given loudspeaker array configuration without the inclusion of artificial masking.

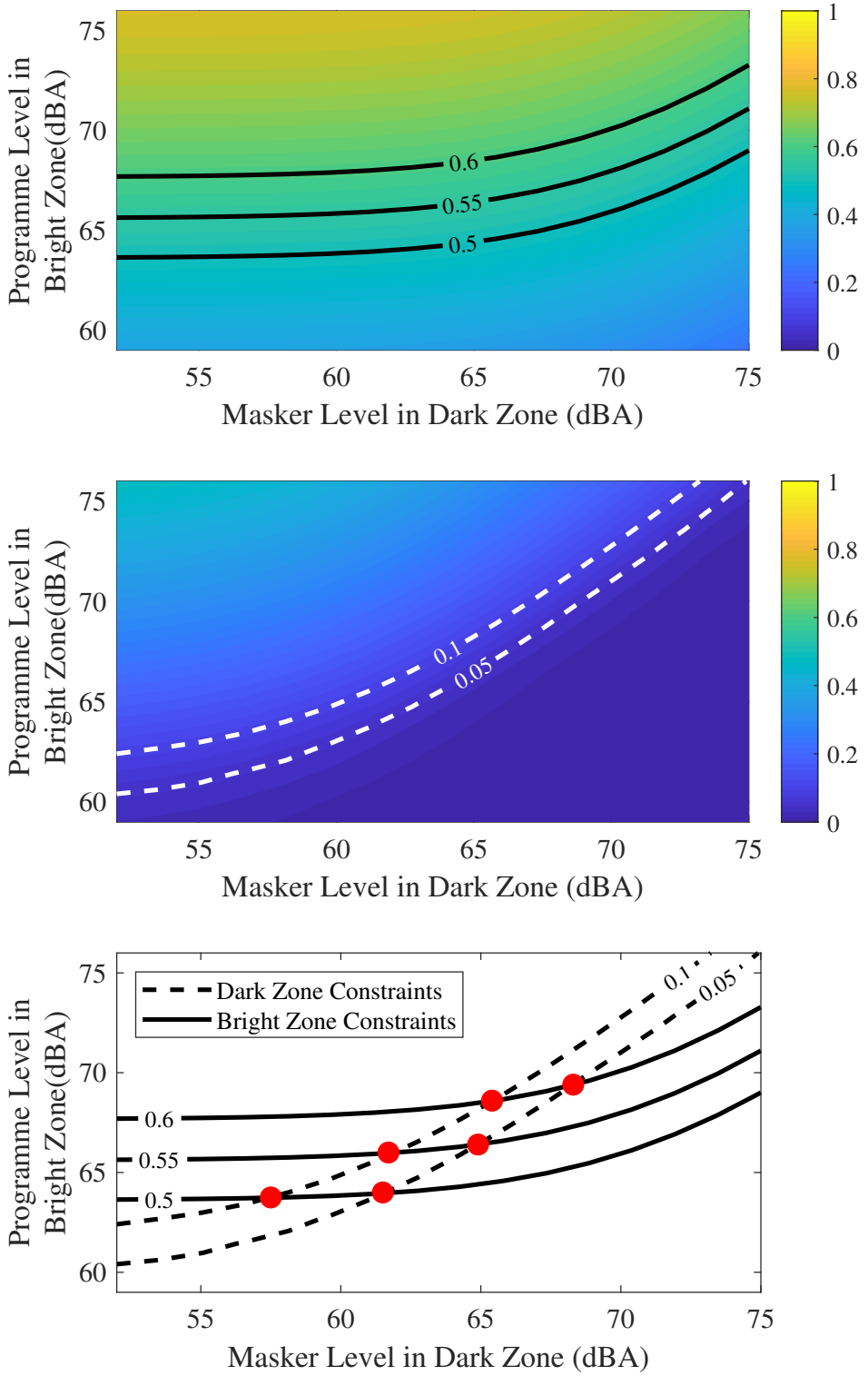


FIGURE 9.3: Predicted SII in bright zone (upper panel) and dark zone (middle panel) with variation in programme and masker levels, with respect to an ambient noise level of 60 dBA. The lower panel shows a superimposition of the contours from the top two panels. The programme and masker levels at the intersections between constraints, indicated with red circles, are the optimal values for minimising the level of the masker.

9.1.2 Specification of Acoustic Contrast Requirements

The results in the previous subsection determined the optimal programme and masker signals for a single simulated speech privacy control system by visualising the intelligibility constraints in the bright and dark zone as contours over a two dimensional parameter space. Analysis of the results showed that these optimal signal levels could be found at the intersection of the constraint contours, as at this point in the parameter space, the signals in each zone have just enough energy to meet each constraint. The acoustic contrast levels provided by the simulated system used in Section 9.1.1, shown in Figure 9.2, were sufficient for a range of combinations of intelligibility constraints. With relatively relaxed constraints of $\mathrm{SII}_d < 0.1$ and $\mathrm{SII}_b > 0.5$, the ambient noise played a significant role in masking the speech leaked into the dark zone. With more onerous intelligibility constraints, higher programme and masking signal levels were required to achieve the specified level of performance. This sequence of simulations demonstrates how the capabilities of a given system can be estimated, but the opposite design problem is likely to be encountered more often in practice; that is, given a certain minimum level of intelligibility in the bright zone, and a corresponding privacy constraint in the dark zone, what level of acoustic contrast is required from a sound zoning system to achieve these targets?

To assess this design problem, a range of systems with different levels of acoustic contrast are simulated, and these acoustic contrast curves are displayed in Figure 9.4, with the single system referenced in Section 9.1.1 marked using a dashed line. The chosen frequency-dependent levels of acoustic contrast follow the trends that have been observed when steps are taken to improve the performance of physical linear arrays. Smaller arrays with zones situated close to one another tend to lack low-frequency contrast, and as the zones are moved apart, the largest gain in performance is in this low-frequency region, as demonstrated with measured results in Figure 5.12. Conversely, the increase in acoustic contrast at high frequencies is less significant, so this is reflected in the choice of acoustic contrast curves presented in Figure 9.4. The methodology presented in this section remains general with regard to different system geometries, as alternative series of acoustic contrast curves that match the frequency-dependent effects of changing a particular variable can be generated. For example, altering the inter-element spacing would adjust the level of high-frequency contrast available due to spatial aliasing effects, as described in Section 5.4. Acoustic contrast curves representative of this geometric change could be generated and used in the optimisation process described later in this section. Alternatively, frequency-independent levels of acoustic contrast may be imposed [27].

The contours displayed in Figure 9.3 were calculated based on a dense grid of evaluations on the parameter space formed by the programme and masker levels. This requires a large number of function evaluations and is therefore an inefficient way to determine the desired optimal point where a given pair of constraint contours meet. The advantage of this expensive calculation is that the contour plots provide general information regarding the trends in SII with changes to the programme and masker levels, and once the SII contours are constructed, all possible intersections of constraint contours within the explored bounds can be determined.

For the determination of the optimal combination of the programme and masker signal levels, a more efficient alternative method is used. The parameter space is explored using an automatic optimisation algorithm, with the goal of minimising the overall level in the dark zone, subject to intelligibility constraints in the bright and dark zones. The algorithm chosen for this purpose is

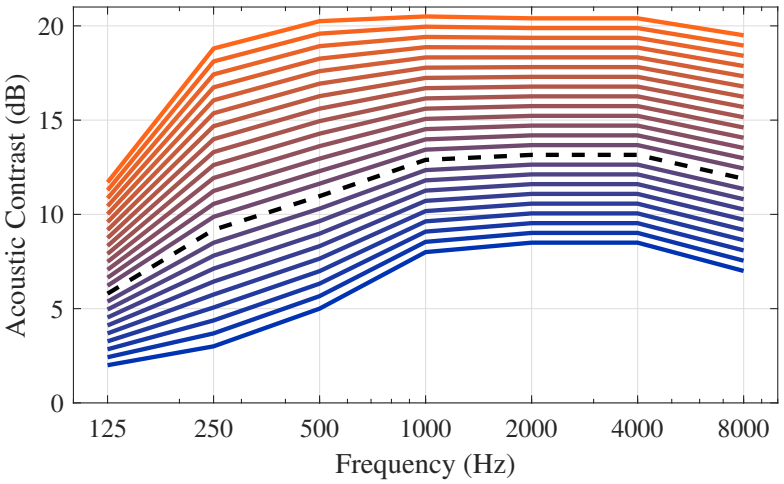


FIGURE 9.4: Series of acoustic contrast levels used to simulate different levels of personal audio system performance.

Pattern Search [245], as this algorithm does not rely on the calculation of gradients in the cost function, which are shallow in regions where the dark zone signal is dominated by the ambient noise, potentially leading to gradient-based solvers halting prematurely. A diagram showing the convergence of the pattern search algorithm is shown in Figure 9.5. The algorithm begins at an initial point and explores the parameter space by assessing the cost and constraint functions at points that are a large distance away in each coordinate direction. In the case of the 2D optimisation problem discussed here, and with reference to the contour plots in Figure 9.3 that the pattern search algorithm is attempting to construct, the solver can be understood to evaluate points that are vertically and horizontally displaced from the initial point, making a plus-shaped pattern. In the simulations presented here, the initial point is set so that both the programme and masker signals are the same level as the ambient noise in their respective zones. Once the cost and constraint functions have been evaluated at the four nodes of the pattern, the location with the minimum cost becomes the new starting location. The size of the pattern is increased each time a new minimum is found, to ensure the algorithm explores the full extent of the parameter space. If all the cost function evaluations at the pattern nodes are greater than the starting point or violate a constraint, the pattern size is decreased, to converge inwards towards the optimal point. The algorithm halts when the size of the pattern is smaller than a pre-defined size, which in this case corresponds to 0.1 dB changes in the signal levels. The output of the algorithm is the optimal pair of signal levels for the given system configuration and noise conditions, i.e. the programme level that must be reproduced in the bright zone and the masker level that must be reproduced in the dark zone.

The constraints selected for the present investigation into acoustic contrast requirements are $\mathrm{SII}_d < 0.05$ and $\mathrm{SII}_b > 0.60$. Figure 9.6 shows the results of optimising systems with the levels of acoustic contrast shown in Figure 9.4, where the acoustic contrast value achieved at 1 kHz is used to identify each system - this will be referred to as the mid-frequency acoustic contrast. At the leftmost edge of the figure, the red shaded region indicates the levels of acoustic contrast where no valid solution to the optimisation problem can be found, as the acoustic contrast is too low to provide the required speech intelligibility contrast. Any increase in the programme level would unacceptably raise dark zone intelligibility, and any increase in the masking signal level would result in excessive degradation to the programme signal in the bright zone. At a

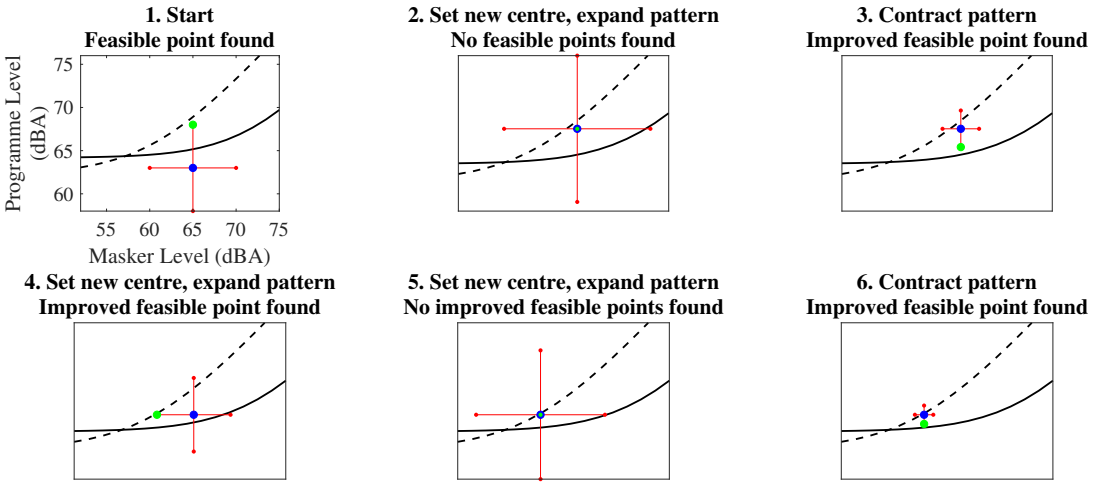


FIGURE 9.5: Schematic diagram of the first six iterations of the pattern search algorithm, with the objective of minimising the overall SPL in the dark zone. Blue and green points indicate the origin of the pattern at the current iteration and the next iteration respectively.

mid-frequency acoustic contrast level of 11.8 dB, a feasible pair of signals is found. With this simulated system, the required energy of the programme and masker signals are 13 dB greater than the ambient noise level, potentially raising dark zone annoyance compared to designs with more acoustic contrast and lower required signal levels. The optimal signal levels for the case considered in Section 9.1.1, with a mid-frequency acoustic contrast of 12.9 dB, are marked with filled circles in Figure 9.6. These match the predicted optimal signal levels indicated by the intersection of the $\mathrm{SII}_d < 0.05$ and $\mathrm{SII}_b > 0.60$ contours in Figure 9.3 to within 0.5 dB. When these optimal signal levels are input into the full array model, whose acoustic contrast profile is shown with a grey line in Figure 9.2, the predicted SII value averaged across the dark zone microphones is 0.048, and in the bright zone, SII = 0.57. This close correspondence between the predicted SII scores demonstrates that the surrogate model is an appropriate substitute for full array simulations.

At higher levels of acoustic contrast, the optimal programme level plateaus at 8 dB above the ambient noise level and the optimal masking signal level decreases at a rate of 2 dB for each increase of 1 dB in the mid-frequency acoustic contrast. This can be attributed to the dual effect of the increase in acoustic contrast: the increased separation between zones reduces the leakage in both directions, i.e. both the leakage of the speech into the dark zone and the leakage of the masker into the bright zone are reduced, so both signal levels can then also be reduced. This gradient continues until the required masking signal level falls below the ambient noise level, at a mid-frequency acoustic contrast level of 17.5 dB. After this transition point, the masking signal level falls sharply whilst the programme level remains constant, in order to overcome the constant ambient noise. The green shaded region denotes the range of systems that provide sufficient separation between zones for the masking signal to be omitted entirely. Systems with this level of acoustic contrast performance earn an additional degree of freedom with regard to the intelligibility constraints in each zone. Maintaining the programme level will result in improved privacy in the dark zone as the mid-frequency acoustic contrast is increased, or alternatively, the programme level can be allowed to increase in order to improve bright zone intelligibility, whilst maintaining the previously set intelligibility limit in the dark zone. In the plots of optimal signal levels presented throughout this chapter, i.e. Figures 9.6, 9.7 and 9.16, the latter approach of

maximising the bright zone intelligibility whilst maintaining the dark zone privacy constraint is presented.

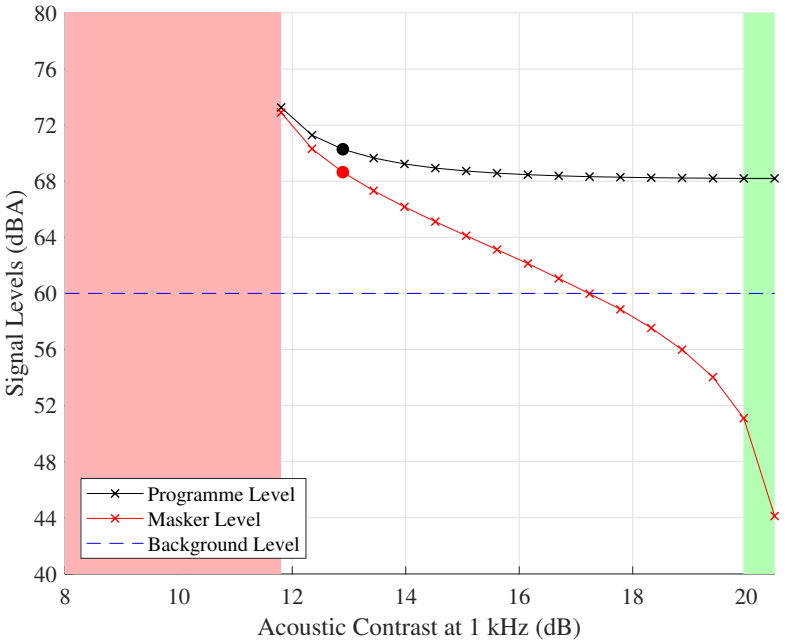
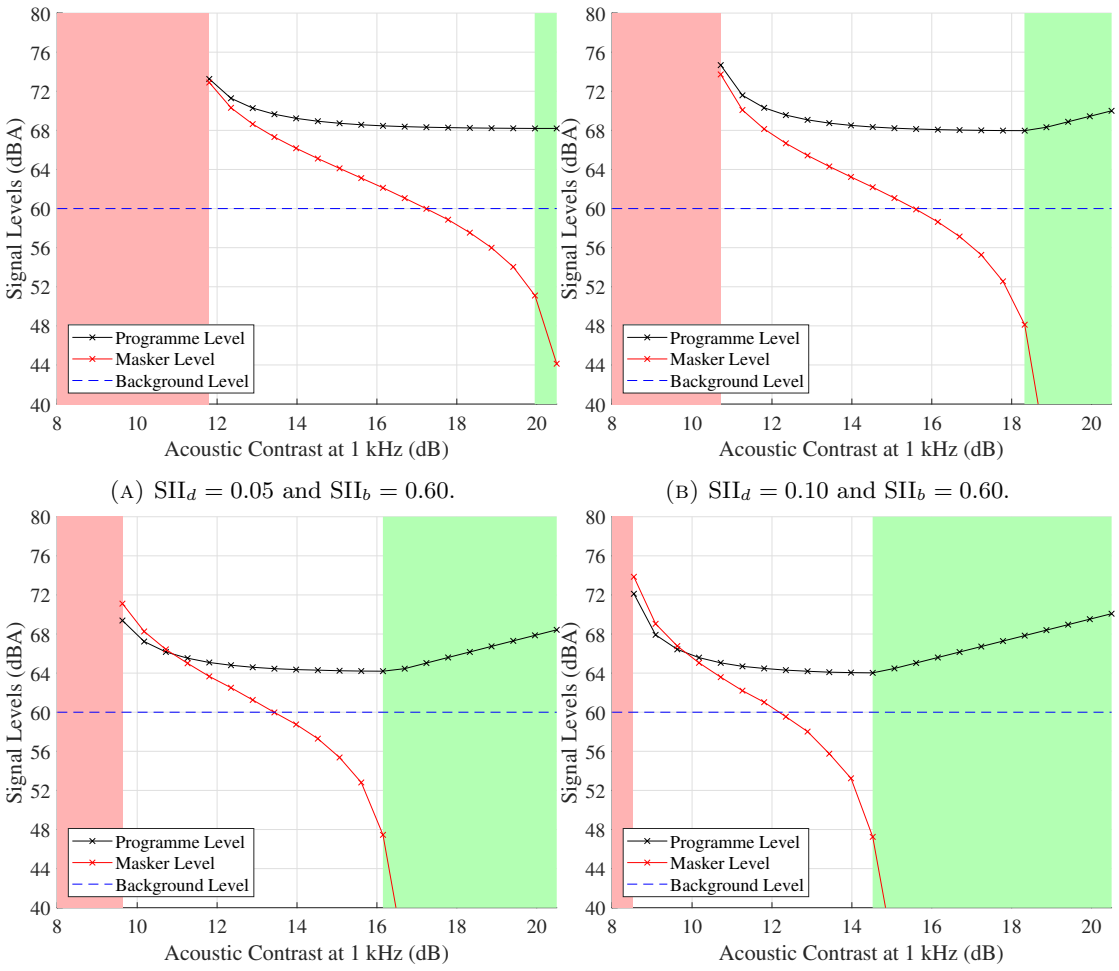


FIGURE 9.6: Optimal programme and masker levels required to achieve the intelligibility constraints of $\mathrm{SII}_d = 0.05$ and $\mathrm{SII}_b = 0.60$ for a range of mid-frequency acoustic contrast levels described in Figure 9.4. In the red shaded region, the prescribed speech intelligibility targets cannot be met, regardless of the programme and masker levels. In the green region, the intelligibility constraints can be met without requiring a masking signal, i.e. the ambient noise alone is sufficient to provide privacy.

The difference in mid-frequency acoustic contrast between the edges of the red and green boundaries in Figure 9.6 quantifies the benefit of incorporating additional masking into a private personal audio system, in terms of the level of acoustic contrast that must be provided. In order to achieve the intelligibility constraints of SII = 0.05 in the dark zone and SII = 0.60 in the bright zone without using any additional masking, i.e. relying on the ambient noise alone, the system must provide a mid-frequency acoustic contrast of 20 dB, potentially requiring a large loudspeaker array. When artificial masking is included, the minimum acoustic contrast requirement is 11.8 dB, reducing the technical requirements of the loudspeaker array system, but it must be noted that at this extreme, the necessary programme and masker levels significantly exceed the ambient noise, potentially resulting in an unacceptable solution. Nevertheless, Figure 9.6 shows that there is a continuous trade-off between acoustic contrast requirements and signal levels, meaning that other target points on the curves could be selected, based on operational requirements. One example would be to set the target acoustic contrast value to the point where the required masking signal level equals the ambient noise level, which for this example is at a mid-frequency acoustic contrast value of 17.5 dB.

When alternative intelligibility constraints are placed on the zonal sound fields, different threshold values for the minimum required acoustic contrast levels are found. Figure 9.7 displays the optimal programme and masker signal levels for four different combinations of intelligibility constraints. The upper left panel of the figure is identical to Figure 9.6, and shows the acoustic contrast requirements when the intelligibility constraints are $\mathrm{SII}_d = 0.05$ and $\mathrm{SII}_b = 0.60$. Relaxing the constraints by reducing the bright zone intelligibility requirement or raising the dark



(A) $\mathrm{SII}_d = 0.05$ and $\mathrm{SII}_b = 0.60$.

(B) $\mathrm{SII}_d = 0.10$ and $\mathrm{SII}_b = 0.60$.

(C) $\mathrm{SII}_d = 0.05$ and $\mathrm{SII}_b = 0.50$.

(D) $\mathrm{SII}_d = 0.10$ and $\mathrm{SII}_b = 0.50$.

FIGURE 9.7: Optimal programme and masker signal levels associated with different levels of mid-frequency acoustic contrast, presented for four combinations of speech intelligibility constraints. In the red shaded regions the speech intelligibility targets in each sub-figure caption cannot be met, regardless of the programme and masker levels. In the green regions, the intelligibility constraints can be met without requiring a masking signal, i.e. the ambient noise alone is sufficient to provide privacy. In this region, the programme signal can be increased in level as the acoustic contrast increases without compromising privacy.

zone intelligibility requirement results in a reduction in the minimum acoustic contrast levels required both with and without the provision of a masking signal, indicated by the edges of the red and green regions. Additionally, the range of allowable acoustic contrast levels between these two boundaries decreases in size as the constraints are relaxed. This indicates that the more onerous the intelligibility constraints, the greater the advantage of providing additional, artificial masking. These conclusions are re-visited in the following section where the effects of the spectral shape and overall level of the ambient noise is investigated.

9.2 Spectral Effects of Typical Ambient Noise Samples

The previous section discussed the relationship between acoustic contrast, signal levels and intelligibility constraints in simulated environments where the ambient noise had the same long-term average spectrum as speech - this corresponds with scenarios where the ambient noise comprises of many people talking at once, a reasonable assumption for the public spaces in which private audio systems might find utility. However, speech is not the only contributor to the overall noise in these environments - other examples include the break-in of traffic noise from outside, industrial noise from mechanical ventilation or air conditioning systems, and the noise associated with the movement of people in the reproduction environment. These additional noise sources do not necessarily match the energy spectrum of speech, and therefore are not necessarily optimal maskers - this reduced masking ability of the ambient noise makes it less useful for providing privacy by masking the speech in the dark zone, but correspondingly, the ambient noise will not obscure the speech programme in the bright zone as efficiently either.

To explore these effects more fully, four ambient noise samples from public places have been selected from the Ambisonic Recordings of Typical Environments (ARTE) database [246]. These recordings were made with an ambisonic measurement system, facilitating analysis of the full 3D sound field, however for the initial investigation presented in this section, only the zeroth order (omnidirectional) ambisonic component was extracted for input into the SII metric. Details of the four selected signals are presented in Table 9.1 and the power spectral densities (PSDs) of the four signals are presented in Figure 9.8. Although they each contain different combinations of noise sources, the four recordings all have similar spectra to the speech-shaped noise sample used in the previous section, except below 125 Hz. The spectral shape of pure speech-shaped noise exhibits a reduction in energy in this frequency region, whereas the four ambient noise samples have a higher level of energy. This difference is most apparent in the library scene, where the PSD is dominated by energy below 125 Hz, associated with the noise from air handling units. The upward spread of masking from low to high frequencies allows this low-level noise to mask portions of speech, despite there being little overlap in their respective spectra.

ARTE ID	Location	Description	dB SPL	dBA
1	Library	University study area in the main library, off-peak hours, quiet.	53.0	46.1
2	Office	Open plan office, people typing, chatting and talking on the phone.	56.7	51.4
5	Church	Small church space, people entering, busy and loud conversations.	65.9	60.9
11	Train Station	Central Station, main concourse – large space, open to the platforms with people walking and talking at peak hour; loud announcements and train sounds.	77.1	73.6

Table 9.1: Ambient noise samples from the ARTE Database [246]. ID numbers correspond with those from the database.

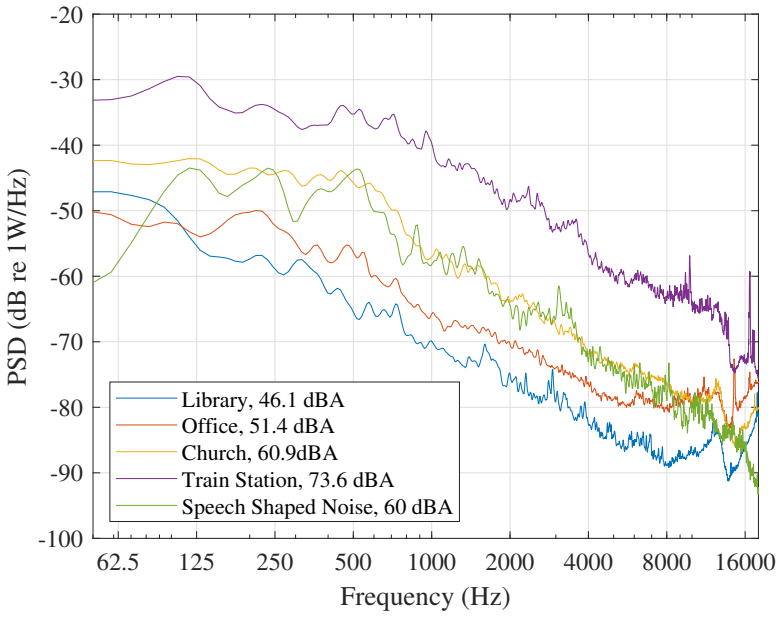
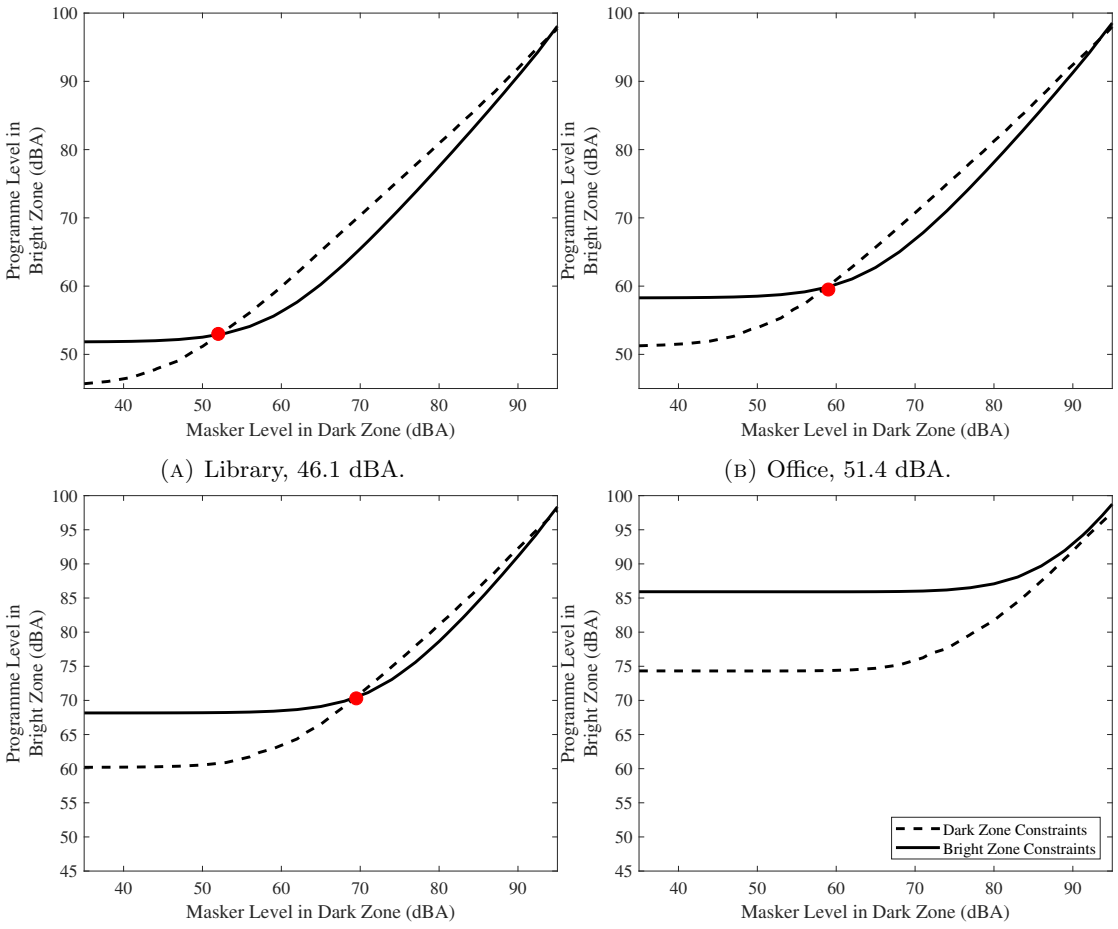


FIGURE 9.8: Power Spectral Density estimates of 60 dBA speech-shaped noise from the VCTK corpus, and four alternate ambient noise signals described in Table 9.1.

A similar analysis to that carried out in Section 9.1.1 shows that for three of the four ambient noise conditions, with intelligibility constraints set at $\text{SII}_d = 0.05$ and $\text{SII}_b = 0.60$, an optimal combination of programme and masker signals exists. The pairs of intelligibility contours for each scene are shown in Figure 9.9. In the library scene, shown in Figure 9.9a, the optimal combination of programme and masker signal levels is low, compared to usual conversation levels, so in this instance, the programme and masker levels may need to be increased, following the dark zone contour to reach a more practical speech level. Should this action be taken, privacy will be predominantly provided by the artificial masker. This removes the reliance on the ambient noise in the space, thus potentially removing the need to provide hardware to monitor this level.

The contours presented in Figure 9.9 all have a similar characteristic shape to the contours presented in Figure 9.3, as the PSDs of the recorded ambient noise samples, shown in Figure 9.8 all have a similar spectrum to that of speech. The main difference between the plots in Figure 9.9 is due to the wide range of ambient noise levels represented; as the ambient noise level increases from scene to scene, the programme and masker signal levels must also be increased to compensate. However, this cannot be continued to arbitrarily high reproduction levels, due to non-linearity in the evaluation of intelligibility at different speech levels. Even without the degrading effect of ambient noise, speech reproduced at an unnaturally high level is judged to be less intelligible than speech produced at normal conversation levels [124]. This is accounted for in the SII metric through the inclusion of the self-masking spectrum level, as described in Figure 3.2. In the loudest scenario, recorded in a train station concourse, the signal levels required to overcome the ambient noise are high enough that the self-masking term in the SII metric becomes significant.

This conclusion is best exemplified in Figure 9.10, which shows the SII evaluation in the bright zone of a system operating in 46.1 dBA ambient noise from the library scene when the programme level is varied. Beyond a speech level of 65 dBA, the SII evaluation begins to decrease, despite



(A) Library, 46.1 dBA.

(B) Office, 51.4 dBA.

(C) Church, 60.9 dBA.

(D) Train station, 73.6 dBA.

FIGURE 9.9: Contours of SII constraints in bright and dark zones with variation in the programme and masker signals, at $SII_d = 0.05$ and $SII_b = 0.60$. Each subplot represents a different background noise condition from Table 9.1. Red circles indicate the optimal, i.e. quietest programme and masker signals that satisfy both constraints, where an optimal point exists.

the programme level significantly exceeding that of the ambient noise and masking signal combined. Consequently, in situations with high levels of ambient noise, for the assumed personal audio system with an acoustic contrast characteristic described by Figure 9.2, there is no valid combination of speech and noise signals that simultaneously satisfy both of the SII constraints - a higher level of acoustic contrast is required. This reveals a fundamental limitation of personal audio systems designed for use in noisy environments - although the ambient noise in a space affects both the bright and dark zone intelligibility, the signal levels in each zone cannot simply be arbitrarily increased. The results presented in this section can therefore be used as a guide for specifying a personal audio system based on the intelligibility constraints and expected ambient noise levels. In locations where the ambient noise level varies, the system must be designed based on the maximum and minimum expected levels. Further detail regarding adaptation of the signal levels according to temporal variation in the ambient noise level is presented in Section 9.4.

Thus far in this chapter, predictions of speech intelligibility have been based on monaural SII evaluations. This ignores the spatial distribution of the ambient sound field, fundamentally

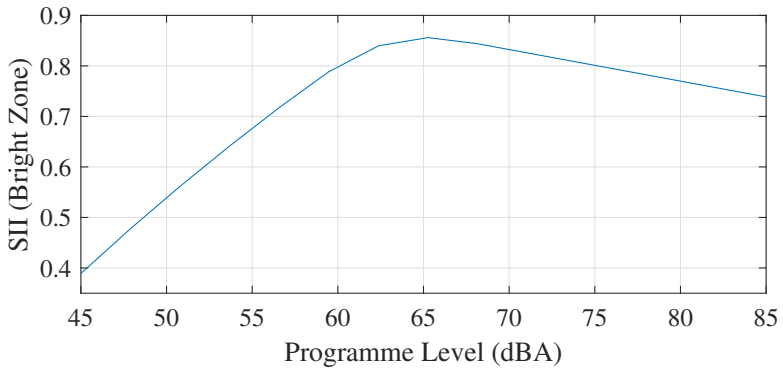


FIGURE 9.10: Variation in the SII evaluated in the bright zone of a personal audio system, with change in programme level. Background noise level = 46.1 dBA (Library scene from the ARTE database [246]) with a constant speech-shaped masking signal level of 47 dBA in the dark zone.

assuming that the effectiveness of a masker only depends on its spectral level. However, the human auditory system can distinguish speech from noise more effectively when the apparent sources of speech and noise are spatially separated [241, 242]. In the following section, the consequences and limitations of this spatial release from masking are discussed, alongside an analysis of the diffuseness of the typical ambient sound fields described in Table 9.1. This provides information on how ambient noise measurements can be processed to isolate the components that can be relied upon for effective masking.

9.3 Spatial Release from Masking

When a target speech signal and an interferer originate from different azimuths, the intelligibility of the target speech is improved compared to the case where the target and interferer are co-located [241, 242], due to a phenomenon termed the Spatial Release from Masking (SRM). In the previous two sections, it has been assumed that the effectiveness of the masking provided by ambient noise can be predicted based on its time-averaged level and spectrum, but in practice, the ambient noise in a room may originate from a small number of discrete sources that surround the listener. In this case, SRM has the potential to reduce the effectiveness of the masking provided by this ambient noise. Therefore, to assess the feasibility of incorporating the ambient noise into the masking predictions made by a speech privacy control system, it is necessary to understand the behaviour of SRM in the presence of multiple sources of masking, and the spatial nature of typical ambient noise fields. These two aspects will be discussed in the remainder of this section.

SRM is attributed to two effects in the auditory system, both related to the signals received at each ear [241]. The first is the better-ear effect, so called as the shadowing effect of the head causes the SNR at each ear to be different when sources of speech and noise are spatially separated. However, in experiments where maskers of equal power are placed symmetrically with respect to the head, negating the better-ear effect, SRM is still observed [247, 248]. These experiments demonstrate that the auditory system also performs binaural processing, i.e. amalgamation of the signals from both ears into a single percept [249]. Structured investigations have been carried out into the binaural processing carried out by the auditory system by presenting

specially constructed signals into each ear to isolate different effects [242, 250]. These experiments have revealed that both interaural time and level differences are significant in complex listening scenarios. In particular, when the signals entering each ear are correlated, for example if they originate from a single point source, the differences in the temporal fine structure between the signals in each ear are important in providing SRM. On the other hand, when the interaural signals are uncorrelated, such as in the case when the noise field is diffuse, the effect of SRM is reduced. This characteristic has potentially significant impacts for personal audio system performance in noisy, reverberant environments.

Several investigations have been carried out to quantify the degree of SRM from a single source of masking placed on the horizontal plane, and a review of research into this area is provided by Bronkhorst [241]. In the same review, attention is also paid to the case where multiple maskers are spatially distributed around the listener. This paradigm is of particular relevance to the present problem of quantifying the relative effects of natural and artificial noise on personal audio system performance, because as the number of discrete masking sources surrounding the listener increases, the masking environment becomes increasingly diffuse.

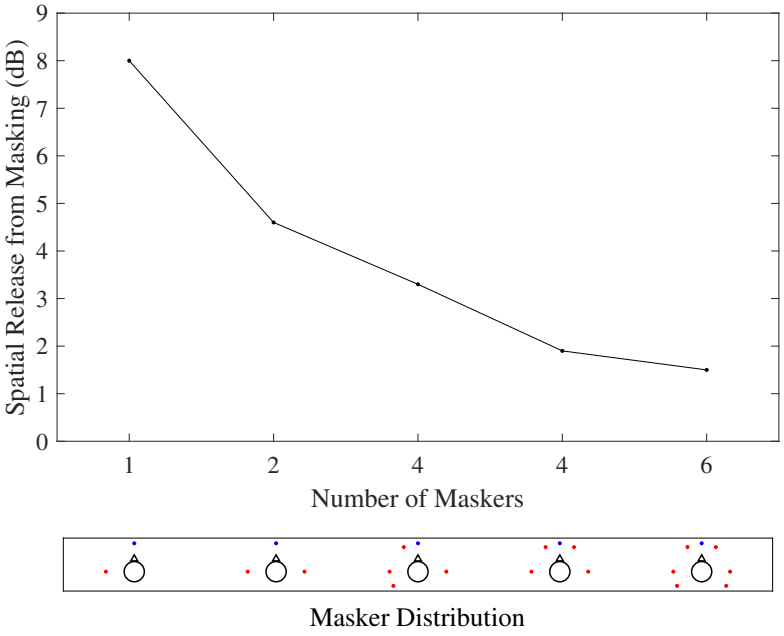


FIGURE 9.11: Spatial Release from Masking (SRM) from multiple maskers (red points) distributed around the listener with respect to a frontal talker (blue points). The abscissa increases with increasing diffuseness of the masking conditions, due to an increasing number or a widened spatial distribution of maskers. Data from Ref. [184].

To illustrate the effect of source positioning on SRM, Figure 9.11 presents data from a study by Bronkhorst and Plomp [184], in which the SRT of meaningful sentences was measured with sources of modulated speech-shaped noise either co-located with the frontal talker, shown by a blue point in Figure 9.11, and spatially distributed around the listener. The difference in SRT between each of these conditions quantifies the SRM, and the results show that as the number of maskers increases and their spatial distribution becomes more homogeneous, the SRM decreases. This trend has been observed in several other studies, using a range of speech tests, masker locations and masking signals [247, 248, 251]. Based on these results, it has been proposed that the binaural processing centre, which would otherwise be able to provide SRM,

can become "overloaded" [252] in complex acoustical scenes where several sources of masking operate simultaneously. The limit to this capacity has been estimated at between three and six individual sources of masking [247, 252]. The study in Ref. [247] showed that when six continuous noise maskers were distributed in front of the listener, the measured SRM was 0 dB. In the context of personal audio system development, this suggests that the level of SRM in diffuse, continuous noise is likely to be low, as this condition represents the mathematical limit of adding many discrete sources of masking to an acoustical scene.

All of the studies referenced above provide evidence that the effect of SRM is reduced, or negligible in diffuse acoustic environments, compared to those with multiple localisable sources of noise, such as nearby conversations, noise from footfall or equipment. However, these types of noise sources may also be strongly time-varying, and therefore are extremely unreliable as maskers for a private audio system, particularly given the practical challenge of monitoring the noise levels within each zone in real-time. On the other hand, the background noise in the reproduction environment, $L_{A90,T}$ [243], is more likely to be spatially diffuse, and be caused by less strongly time-varying, distributed sources such as distant traffic noise, or noise from a ventilation system. Whilst statistical stationarity of the background noise cannot be guaranteed, its ability to mask speech over the time-span of a spoken sentence is easier to predict than the impact of the aforementioned foreground noise sources.

The L_{Aeq} and L_{A90} of the four ambient noise samples detailed in Table 9.1 are displayed in Figure 9.12. The difference in level between the equivalent A-weighted SPL measured in each space and the L_{A90} is also presented, and ranges from 4.5 to 7.0 dB. These L_{A90} values may be used in place of the L_{Aeq} values presented in the results shown in Section 9.2 to better represent the background A-weighted SPL in each space that can be relied upon to provide consistent masking, and that can be more accurately assessed using the SII metric.

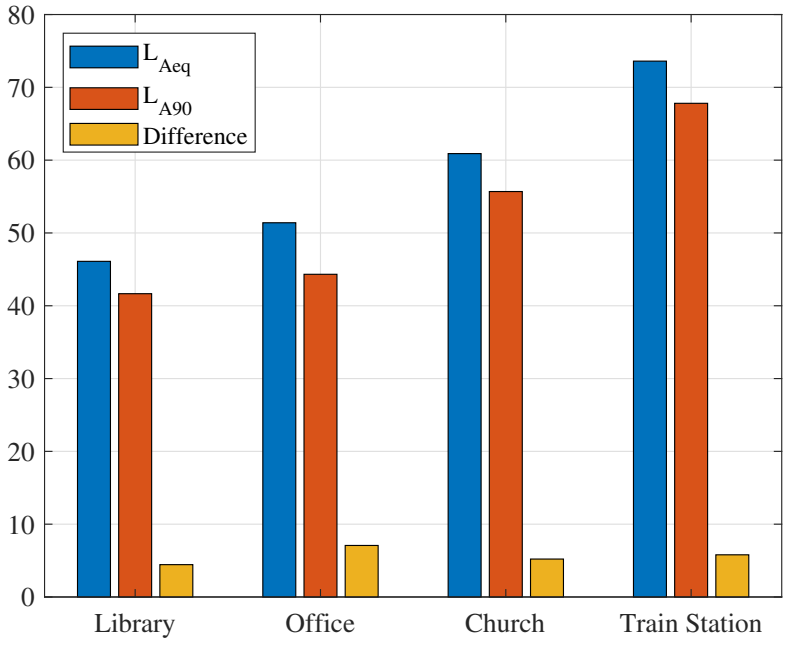


FIGURE 9.12: Equivalent A-weighted SPL L_{Aeq} and background noise level L_{A90} for four ambient noise samples from the ARTE Database [246].

.

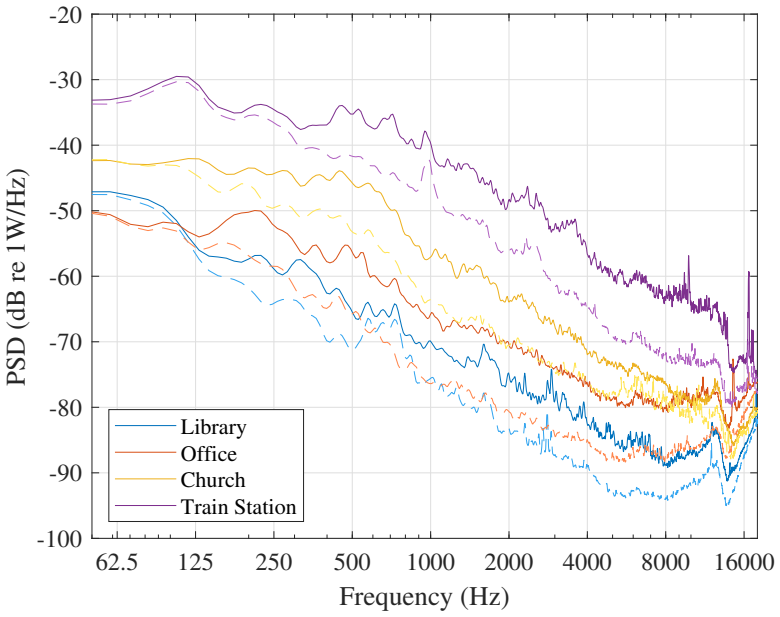
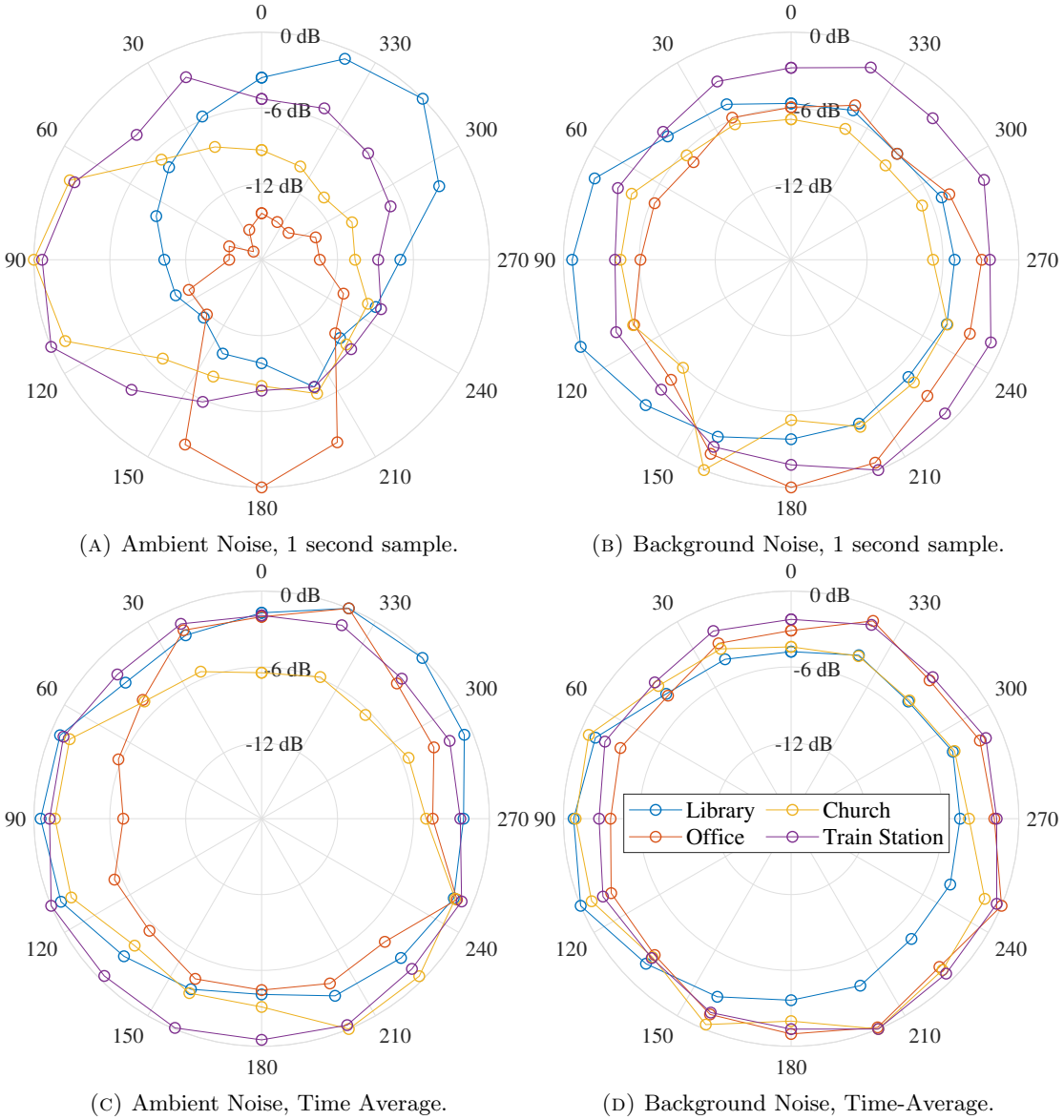


FIGURE 9.13: Solid Lines: Power Spectral Density of four ambient noise signals taken from the ARTE Database, as in Figure 9.8. The dashed lines show the corresponding PSDs of the associated background noise, defined as the quietest 10% of each signal.

The L_{A90} index is formed by analysing the signals at ⅛-second intervals, corresponding to "fast" time weighting [253], and calculating the A-weighted SPL over each interval. The value at the 10^{th} percentile of this dataset is defined as the L_{A90} value. To assess the spectrum of this background noise, all the ⅛-second segments of the original input signal with a level lower than the L_{A90} were concatenated, providing a continuous sample of the quietest 10% of each signal. Welch PSD estimates of both the ambient noise and the background noise are presented in Figure 9.13. From these spectra, it can be seen that the quietest proportion of the signals is dominated by low frequency energy - the largest differences between the solid and dashed lines occur in the speech frequency range. This, alongside manual analysis of contiguous quiet sections of each noise sample, confirms that the process of isolating the background noise from each recording has reduced the effect of discrete nearby talkers in each situation.

In order to assess the degree of diffuseness of these typical background noise samples, the original ambisonic recordings from the ARTE database [246] can be processed to yield the directional characteristics of the ambient sound field and that of the isolated background noise, using the method described by Weisser et al. [246]. In this method, the ambisonic recordings are decoded to a horizontal ring of 16 loudspeakers, and the relative A-weighted SPLs of these sources quantify the directionality, or diffusivity of the sound field. Figure 9.14 represents this data as polar directivity plots for each of the considered environments described in Table 9.1. Plots A and B each show a 1-second snapshot of the computed directivity patterns from the ambient noise and background noise respectively, and demonstrate that the ambient noise field contains, at times, highly directional and localisable sources of masking sound, whereas the background noise is substantially less directional overall, and the angular variation is also smoother, indicating either that multiple sources of background noise are operational, or that the microphone position is far enough from the source that the reverberant field is dominant over the direct sound. Both of these alternatives result in limited inter-aural correlation, and hence a reduction in the SRM. A similar trend is visible in Figures 9.14c and d, which show the time-averaged directivity of the

ambient and background noise over the duration of each recording. The lobes in the directivity of the ambient noise in the library, office and church scenes reveal the positions of small groups of talkers, but these are not as prominent in the corresponding background noise plots.



(A) Ambient Noise, 1 second sample.

(B) Background Noise, 1 second sample.

(C) Ambient Noise, Time Average.

(D) Background Noise, Time-Average.

FIGURE 9.14: Directional characteristics of the ambient and background sound fields from the church scene described in Table 9.1.

The diffuseness of the ambient and background noise can also be assessed using the Directivity Index, DI [254], which can be calculated using a similar process to the polar plots presented above, except the ambisonic signals are decoded to a full spherical array of sources surrounding the measurement position. For these results, a 30-channel spherical array was simulated, as the unaliased reproduction of fourth-order ambisonic signals requires a minimum of 25 sources. The DI is defined as the ratio of the maximum energy of a single source in this array to the average energy of all the sources. The DI values for the ambient and background noise fields in each space are presented in Table 9.2, and for three of the four environments, the directivity of the background noise is less than that of the corresponding ambient noise. In each case, the discrepancy from the ideal diffuse field of $DI = 0$ dB is due to the sources lying predominantly

Location	Ambient DI (dB)	Background DI (dB)
Library	4.7	4.9
Office	4.9	3.7
Church	5.2	4.2
Train Station	3.5	3.2

TABLE 9.2: Directivity Index (DI)values for the ambient and background noise in four environments from the ARTE Corpus [246].

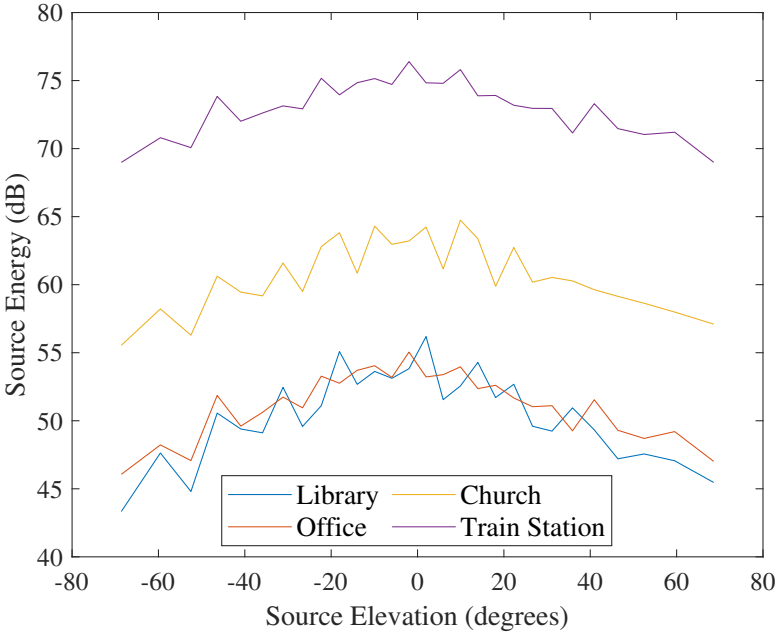
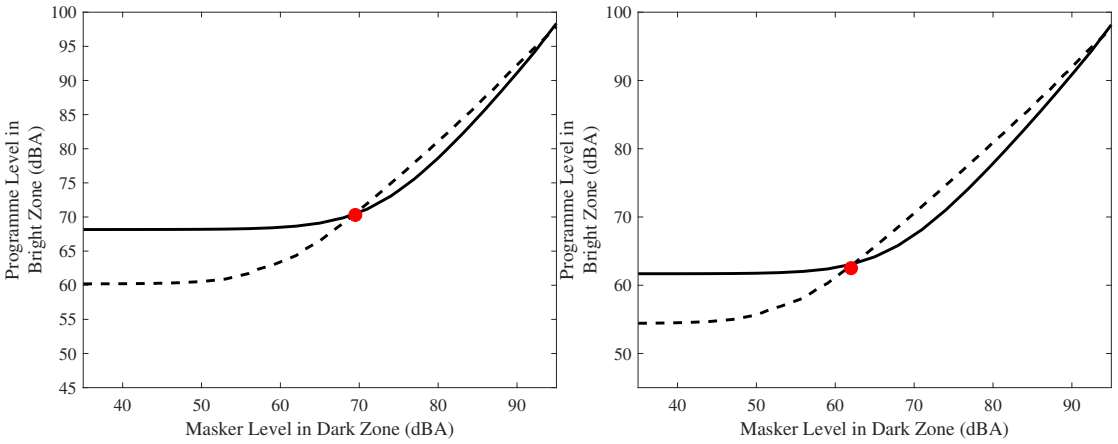


FIGURE 9.15: Equivalent source energy as a function of elevation for the background noise from four environments in the ARTE corpus [246].

on the horizontal plane, and the ceiling height in each space being significantly less than the dimensions of the floor area, as showed by the plots of source energy against elevation in Figure 9.15. This verifies that the analysis of the horizontal plane directivity presented in Figure 9.14 captures the potential dominant sources of SRM in each space. The stratification of the sound field also explains the relatively small differences between the computed DIs of the ambient and background noise samples, as the DI captures the full 3D directivity, thereby diluting the effect of the sources distributed around the horizontal plane.

Using the church scene from Table 9.1 as an example, it is possible to predict the optimal programme and masking signals based on both the ambient noise and the background noise, as indicated by the L_{A90} index. Figure 9.16 shows contours of SII in the bright and dark zones with variation in the programme and masker levels, for (A) the overall ambient noise in the environment, and (B) the background noise. As the SII metric does not account for the effect of SRM, the optimum programme and masker levels predicted using the ambient noise level are higher by 7.5 and 7.8 dB respectively, compared to those predicted using the background noise alone. The results from Figure 9.11 show that with one dominant competing noise source, spatially separated from the target source of speech, SRM values of 8 dB are observed. In this particular case, therefore, the ambient noise is expected to produce negligible additional masking over that provided by the diffuse background noise, despite being objectively louder by 5.2 dB.

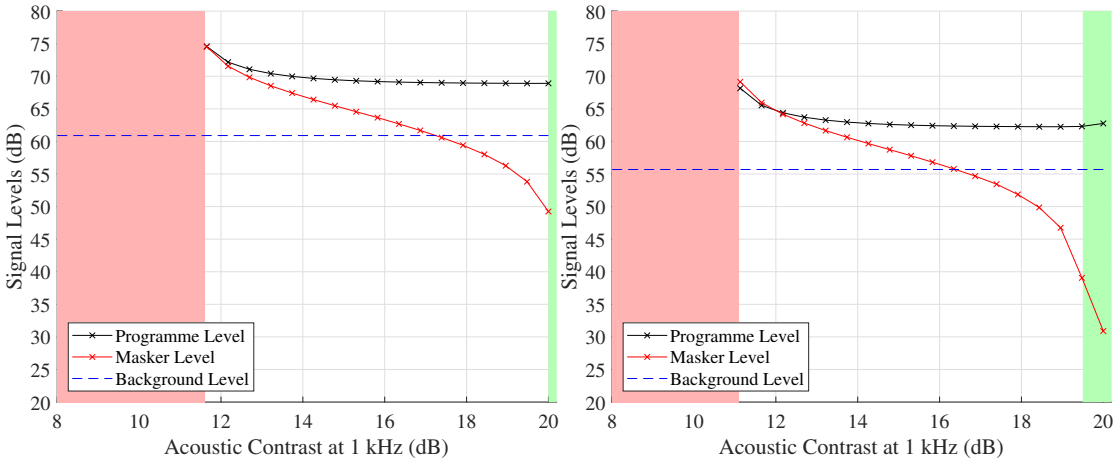
While there is a significant difference in signal levels when predictions are made based on the ambient noise and the background noise, the acoustic contrast requirements for private operation with and without additional masking are very similar. Comparison between Figures 9.17 A & B show that systems with mid-frequency acoustic contrast levels of 11-11.5 dB can provide sufficient privacy and bright zone intelligibility provided that an additional masking signal is used, and that acoustic contrast levels of 19.5-20 dB are required in order to omit additional artificial masking altogether. This similarity can be attributed to the similar spectral shape of the ambient and background noise in the church environment as displayed in Figure 9.13. In spaces with a greater discrepancy between the two spectra, in particular where the background noise spectra deviates significantly from a long-term average speech spectrum, higher acoustic contrast levels would be required to achieve a given level of privacy performance, as well as higher programme and masker signal levels.



(A) Church Ambient Noise, 60.9 dBA. (B) Church Background Noise, 55.7 dBA.

FIGURE 9.16: Contours of SII constraints in bright and dark zones with variation in the programme and masker signals, at $\text{SII}_d = 0.05$ and $\text{SII}_b = 0.60$. Left plot shows predictions based on the ambient noise in the church scene described in Table 9.1. Right plot shows the prediction based on the background noise in the same environment, i.e. the level exceeded 90% of the time. Red circles indicate the optimal programme and masker signals that satisfy both constraints.

The results presented in this section show that if SRM is not accounted for, the intelligibility reduction caused by ambient noise is likely to be overestimated by standard monaural intelligibility metrics. Through a combination of better-ear listening and binaural signal processing, the human auditory system is able to reject interference from discrete sources of masking that are not co-located with the source of speech. The background noise in a reverberant space is less likely to be dominated by these types of sources, and so is more reliable for use as a masker in personal audio contexts for two reasons. Firstly, and by definition, the level and spectrum of the background noise is relatively steady compared to the ambient noise, so the chance to glimpse the target speech is reduced. This increases the reliability of simple predictions of the masking based on the spectrum alone, such as those provided by the SII metric. Secondly, the background noise is in general more diffuse than the overall ambient noise field, resulting in less SRM than is the case with the overall ambient noise. This diffuse nature also provides a practical benefit - under this assumption, the effect of the background noise will be equal in both the bright and dark listening zones, and the effects on speech intelligibility can be evaluated with the same methods as the artificial masking. This simplifies the modelling of intelligibility reduction,



(A) Church Ambient Noise, 60.9 dBA. (B) Church Background Noise, 55.7 dBA.

FIGURE 9.17: Predictions of optimal programme and masker signal levels, based on intelligibility constraints of $SII_d = 0.05$ and $SII_b = 0.60$, for different levels of mid-frequency acoustic contrast. Left plot shows predictions based on the ambient noise in the church scene described in Table 9.1. Right plot shows the prediction based on the background noise in the same environment, i.e. the ambient noise level exceeded 90% of the time.

and reduces the hardware required for background noise measurement to a single microphone, located at a convenient position near the system. This type of background noise measurement is still likely to be required in some applications, due to long-term changes in the composition of the background noise, and this is discussed in the following section.

9.4 Temporal Variation in Ambient Noise

In the previous section, the spatial effects of ambient noise on personal audio system performance were explored, with a focus on how the background noise in a reproduction environment could be separated spatially from the ambient noise. Ignoring the effect of fluctuating, discrete sources of noise in the room is advantageous when specifying the required level of additional masking, as the effects of SRM would have to be calculated for these sources when assessing their effect on the intelligibility in the bright and dark zones.

In addition to the spatial separation of foreground and background sound sources, by definition, the background noise in a reproduction environment can also be isolated temporally from the overall ambient sound. Variations in the ambient noise level across fractions of a second can allow glimpses of the target speech to be unmasked [130]. These brief, but highly intelligible words or syllables significantly increase speech intelligibility, particularly of familiar or predictable sentences [255]. Even without the benefit of contextual information, the redundancy of spoken language means that even under a range of time-varying masking conditions, connected speech can often still be readily understood [256]. As discussed in the previous sections, increased intelligibility is beneficial in the bright zone, but detrimental in the dark zone where low intelligibility is desired. However, this does not necessarily mean that the speech intelligibility contrast between the listening zones is unaffected by modulation of the background noise. For example, if the speech in the bright zone is fully intelligible despite the presence of the constant background

noise, but the dark zone sound field is susceptible to glimpses of speech through the temporal variation in the overall ambient sound, this could compromise the privacy of the target listener.

This further suggests that when a combination of artificial masking and ambient noise is intended to be used in a speech privacy control system, the background noise level should be used to predict the required levels of artificial masking, or the acoustic contrast requirements of a loudspeaker array intended for use in a particular location. An additional benefit of this approach is that the level of the masker need not be adapted rapidly, i.e. over the time-scale of spoken words, to changes in the ambient sound level. As discussed in Chapter 4, it is advantageous for the masking signal to remain at a constant level, as level fluctuations can be regarded as a source of noise annoyance. Whilst this resolves the issues inherent with rapidly time-varying masking signal levels, the approach described thus far in this chapter has not yet addressed the effect of longer term changes to the background noise level, which may require adaptation throughout the course of a day. In public spaces, privacy is clearly only a concern when other people are sharing the space, and with any group of people comes the potential for fluctuation in the background noise due to changes in the level of occupancy.

To illustrate how speech privacy control systems are affected by changes in the background noise level, Figure 9.18 shows how the optimal programme and masker levels vary when the background noise level changes, for a system with the acoustic contrast profile shown in Figure 9.2, and intelligibility constraints of $\mathrm{SII}_d = 0.05$ and $\mathrm{SII}_b = 0.60$. The optimum levels of the programme and masker are strongly correlated with the background noise level, so for clarity, the ordinate of Figure 9.18 represents the programme and masker levels relative to the background noise level. The data points marked with black and red circles correspond to the previously discussed situation with a speech-shaped background noise at 60 dBA. These are displayed in Figure 9.6 using the same notation for ease of comparison. In general, the optimal programme and masker levels increase at a greater rate than the background noise. For example, at a background noise level of 40 dBA, the programme and masker levels must exceed the background by 5.5 and 3.5 dB respectively, whereas at a background noise level of 60 dBA, the programme and masker signals must be reproduced at 8.5 and 10 dB above the background. As the ACC process is a linear filtering operation, this effect can only be caused by non-linearities in the SII metric, as discussed earlier in Section 9.2. The red region in Figure 9.18 denotes the background noise level above which an optimal pair of programme and masker signals cannot be found using the described array configuration, as was encountered previously in the train station example shown in Figure 9.9d. Using this approach, data from noise surveys could be input into system simulations prior to installation. This would allow system integrators to determine the power requirements of the array and the required maximum level of acoustic contrast, based on the expected range of background noise levels.

As with previous plots of this type throughout this chapter, each pair of data points is the output of one pattern search optimisation run. With no optimising assumptions about the level or spectrum of the background noise or the programme material, each run took an average time of 11.1 seconds to complete, corresponding to an average of 93 SII evaluations per run on a desktop workstation. The run-time is subject to some random variability that depends on the size and shape of the feasible region subtended by the constraint contours with respect to the starting position of the pattern, but despite this, no run took longer than 25 seconds to complete. The design of the pattern search algorithm and the expected shape of the cost surface guarantees

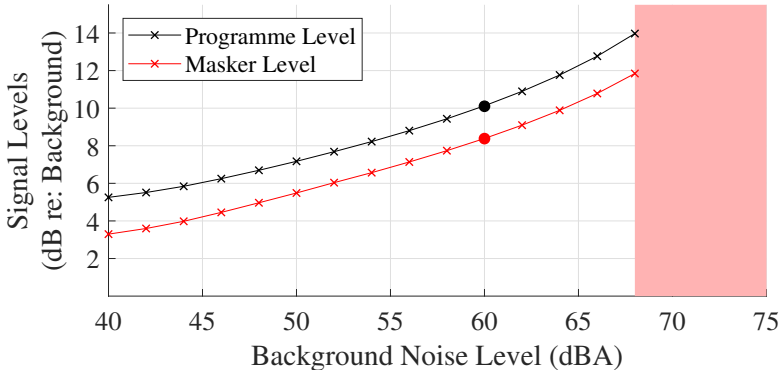


FIGURE 9.18: Optimal programme and masker signal levels with variation in speech-shaped background noise level, from an array with the acoustic contrast profile shown in Figure 9.2, and intelligibility constraints of $SII_d = 0.05$ and $SII_b = 0.60$.

that each iteration yields an improvement towards the optimal signal values, so the algorithm could be terminated after a pre-determined time interval, should this be necessary.

As the intention of the system is to adapt to long-term changes in the background noise level, rather than sudden changes in conditions caused by nearby, localisable sound sources, this approach is likely to be feasible, even with less powerful computational resources, and could be run at pre-determined time intervals, based on data from a longer measurement period. BS 4142, in its commentary on background noise measurements, suggests that a measurement period of at least 15 minutes is desirable for an accurate assessment of the background noise level [243]. This slow fluctuation in the background noise level has been observed in studies of open plan office sound masking systems [91, 257], and in response, systems to schedule the level of masking, or slowly adapt it based on the level observed by monitoring microphones, are available commercially [258]. Similar techniques could be employed in speech privacy control systems to reduce the level of additional masking input into the space, and to control the programme level to ensure good intelligibility for the target listener. In locations where significant fluctuation in the background noise level is common, system integrators should opt for designs where the majority of the masking signal level is directly controlled by the system, as opposed to relying on the masking provided by the background noise, as this provides maximum reliability for the central claim of speech privacy. A similar recommendation was made by Pirn [259] in a 1971 article on the variables affecting open plan office privacy:

> "To mask unwanted speech, we generally cannot rely on man-made activity noise because of its intermittent and often unpredictable character. If at times it dominates the ambient acoustical environment, such noise can be a boon to privacy seekers. On the other hand, it can also interfere with communication... Background noise in an open plan is invaluable. It must be continuous and unobtrusive, yet of an adequately high level to supply the necessary masking. The relative unpredictability of other sources suggests that electronically generated noise, evenly distributed via concealed loudspeakers, is the most reliable of several possible solutions." [259]

This suggests that in locations where the background noise is controlled globally, for example to reduce distraction in open plan environments, separate speech privacy control units for specific

purposes could be linked to the background noise system, thereby preserving the intended indoor soundscape.

Another application area where speech privacy control systems must operate in the presence of significant ambient noise is where sound zoning is integrated into in-vehicle telecommunications systems [10]. In road vehicles, noise from the powertrain, road-tyre interaction and ancillary systems such as air-conditioning are relatively steady compared to the modulation rate of speech, but vary much more rapidly than the background noise level in public buildings. Additionally, the difference between interior noise levels at idle and cruising speed in an average passenger car is around 25 dB [260], and modern vehicles with stop-start technology to prevent prolonged idling will result in an even wider range of background noise conditions. This forces the inclusion of continuously adjustable masker and programme levels, to ensure sufficient intelligibility and privacy for the target listener. These considerations potentially necessitate an alternative, faster strategy to predicting the masking effect of in-vehicle interior noise, such as is used in speed-dependent equalisation and volume control systems [261]. Masking predictions could either be based on microphone inputs, as has been described above, or alternatively by aggregating non-acoustical measurements from vehicle sensors. With the latter approach it may be possible to eliminate the computational demand of online masking predictions through the use of a lookup table. An additional consideration is that, inside a vehicle, powertrain and road/tyre noise can appear to originate from a combination of locations surrounding the listener [262], so estimations of the degradation to intelligibility would also have to take this spatial distribution of masking into account, as was discussed in Section 9.3.

9.5 Summary

In many envisioned circumstances where personal audio technology may be used to provide private sound zones, the surrounding environment will be noisy. Neglecting this ambient noise in the design of a private speech transmission system may lead to excessive masker levels and poor intelligibility in the bright zone, whereas systems that are designed to use a combination of artificial and natural background noise can afford to reduce the level of the additional masking signal, resulting in a reduced potential for noise annoyance. When a combination of artificial masking and background noise is used in a speech privacy control system, up to 8 dB less acoustic contrast is required for the same level of privacy, compared to the case where a single sound zoning process is used and all the masking is provided by the ambient noise. In either case, over-reliance on the ambient noise to provide masking has pitfalls, as the ability of this noise to mask the target speech must be carefully calculated, taking into account its spatial distribution and temporal profile.

When the ambient noise in a reproduction environment exhibits strong temporal modulation, and is formed by multiple, spatially distributed sources, the masking effect is reduced compared to the steady masking produced by the speech privacy control system, whose source appears co-located with the speech. Monaural intelligibility predictions based on the long term spectrum of this type of noise will over-predict the masking effect of the ambient noise, potentially leading to the privacy of target listeners being violated. To mitigate this spatial and temporal variance,

systems should be designed to rely only on the constant, diffuse masking provided by the quasi-stationary background noise in the reproduction environment. This background noise level can be captured from the statistics of a standard noise survey [243], and its effects on intelligibility can be predicted using simple metrics such as SII, as the spatial release from the masking provided by diffuse, continuous noise has been shown to be negligible [247].

Nevertheless, in buildings, daily occupancy patterns, mechanical ventilation settings and nearby traffic levels can cause the background noise level to vary, so systems must be designed to adapt to different background noise conditions. Likewise, systems installed in vehicles will be required to operate in a wide range of background noise levels, commensurate with changes in vehicle speed, engine load and the sound generated by ancillary vehicle systems. While the background noise level is undoubtedly related to its ability to mask speech, its spectral and temporal profile must equally be taken into account when predicting the intelligibility of speech in each listening zone. For practical reasons, it may not be possible to sample the sound fields within the bright and dark zones directly, but the diffuse field assumption means that real-time background noise measurements may be taken from a more convenient location nearby. The in-situ transfer response measurements required for the sound zoning process can be used to provide predictions of the reproduced zonal sound fields in the absence of background noise, which may then be updated with the background noise readings to form intelligibility estimates for each zone. These may then be used to adjust the programme and masker gains at a rate appropriate to track the temporal variation without causing undesirable fluctuations in the masking signal level.

Throughout this chapter, the main focus has been to investigate the effects of ambient noise on the objective, measurable goals of private personal audio system design. By isolating the steady, diffuse, background component of the ambient noise, an additional reliable source of masking has been established. Under several presented scenarios, the background noise can be combined with artificial masking to yield sufficient intelligibility in the bright zone and privacy in the dark zone. However, as discussed in Chapter 6, successful systems must not only provide an appropriate degree of speech intelligibility contrast, but the character of the additional masking should also be acceptable to passive listeners in the dark zone. The incorporation of background noise into the speech intelligibility calculations invites a parallel investigation into how this combination of masking signals is perceived. Of particular interest is the perceptual effect when the artificial masking and ambient noise are spectrally and temporally distinct; it is expected that the preference expressed for an artificial masking signal of a given character will depend on the situational context and the underlying background noise in which it is reproduced. A set of experimental procedures designed to ascertain the details of this potentially complex relationship are presented in the following chapter as suggestions for future work, alongside the main conclusions of this thesis.

Chapter 10

Summary, Conclusions and Future Work

This thesis has presented a range of simulated and experimental results concerning the design and evaluation of personal audio systems for private speech reproduction. In the systems described in this thesis, two listening zones are defined, and each receives a corresponding signal; the target speech material is focussed into the bright zone for the target listener, and a masking signal is focussed into the dark zone, which is occupied by other listeners from whom the target message should be kept private. Privacy is achieved by adjusting each signal based on the intelligibility of the target message in each zone and as an additional objective, the masking signal may also be adjusted to minimise the likelihood of annoying nearby listeners. In order to satisfy these objectives and constraints, an automated process for providing these adjustments has been described, based on data from speech intelligibility and subjective preference tests, and correlation with objective and subjective metrics. To assess the practical challenges of implementing speech privacy control systems in realistic acoustical conditions, the effects of reverberation and ambient noise on system performance have also been examined.

10.1 Conclusions

The relationships and dependencies between several variables in the speech privacy control problem have been identified. Through masking signal selection, gain adjustment, equalisation and regularisation, system designers can control the SNR in each listening zone, with the aim of providing the required level of intelligibility and privacy for the target listener. However, the relationship between these two indices depends on the temporal and spectral profile of the masker, the context, and the expectation of privacy in the particular scenario of interest. The result of this complex web of dependencies is that it is impossible to specify a general SNR target in each listening zone, as this calculated value will only be valid for the specific masker under consideration at that time. Speech intelligibility metrics measure the audibility of speech cues under a range of different forms of degradation, thereby enabling the dependence on the type of masker to be separated from the contextual variables. Consequently, rather than setting a

target SNR, a target value of a speech intelligibility metric can be set, and the corresponding SNR can then be found iteratively. The SII metric was found to be an appropriate metric to use for this purpose for several reasons:

- The primary means of intelligibility reduction is due to the additive noise from the masker, or the background noise in the reproduction environment, and the SII can be calculated rapidly from estimates of the spectrum of these constant sources of masking.
- The SII has also been widely used in clinical and research settings, giving rise to a large number of published transfer functions between SII values and the outcome of speech intelligibility tests. This allows for the true intelligibility of syllables, words and sentences to be estimated based on the predicted SII values in each listening zone.
- Links have been established between the SII and various descriptions of privacy. These relationships highlight the contextual differences in privacy expectations in closed and open plan spaces, and provide benchmark values for the dark zone SII target.

An important consideration in the provisioning of speech privacy control systems, besides the accurate prediction of zonal intelligibility, is the perceptual effect of introducing additional masking noise to an environment. To assess this, a combination of subjective metrics have been used to evaluate the psychoacoustic annoyance of the dark zone sound field. At low SNRs that correspond to the low levels of intelligibility expected in this zone, the sensation of loudness forms the dominant component of the psychoacoustic annoyance metric. However, the two objectives in the dark zone, of reducing the intelligibility and reducing the psychoacoustic annoyance, each require adjustments to the loudness of the masker in opposite directions. This restriction means that the less significant contributions of the sharpness, fluctuation strength and roughness metrics must also be included in the perceptual evaluation of a system.

To preface this perceptual evaluation, the acoustic contrast of various configurations of a loudspeaker array prototype were tested. Uniform linear arrays with a fixed number of elements exhibit a trade-off between low- and high-frequency acoustic contrast as the inter-element spacing is changed, due to corresponding changes in the array aperture and the spatial aliasing frequency. Spatial aliasing is particularly problematic in speech privacy control systems as prominent sidelobes in the directivity at high frequencies can lead to leakage of intelligible consonant sounds from the bright zone into the dark zone, significantly reducing privacy. Likewise, leakage of the masker into the bright zone can mask the ability of the target listener to understand these consonant sounds, further limiting system performance.

The perceptual relevance of this leakage between listening zones was investigated using objective speech intelligibility tests and paired preference subjective testing. Two loudspeaker array configurations and three masking conditions were tested, corresponding to three different approaches for setting the cut-off frequency of a low-pass filter applied to a speech-shaped noise masker. In the *wide* loudspeaker array condition, which caused spatial aliasing within the speech frequency range, measured SRTs were 6-20 dB lower than in the corresponding masking conditions using the *narrow* loudspeaker array, whose spatial aliasing frequency fell outside of the speech frequency range. Additionally, the SRT increased by 5-15 dB for both array configurations as the low-pass filter cut-off was increased in frequency. These results quantify the improvement to speech intelligibility caused by non-optimal coverage of the speech spectrum by the masker.

Listener preferences to the tested conditions were obtained by presenting pairs of simulated dark zone signals, which all corresponded to similar levels of privacy according to the measured SRTs. These paired preference results indicated a strong preference for speech-shaped maskers that had not undergone significant low-pass filtering, as these masking signals could be reproduced at a lower level in order to achieve the same amount of intelligibility reduction. Comments from listeners after the paired preference evaluations revealed a preference for maskers that were perceived to be "natural", as opposed to "artificial" in origin, suggesting that situational context may play a role in the acceptability of masking sound. A procedure for investigating this effect is presented as a suggestion for future work in the following section.

The results from the speech intelligibility and paired preference tests were compared against objective and subjective metric evaluations, in order to set design constraints and an optimisation goal for the design of the masking signal. By comparing the measured intelligibility of the presented stimuli against the computed SII across all tested conditions, the dark zone SII constraint was set at SII = 0.05. This corresponded to an average score of 50% words correct in the matrix sentence test. At this level of intelligibility, participants reported that speech could be considered private due to the additional information provided by the candidate word list that would not be available in practice. Table 3.1 shows that an SII constraint of 0.05 is stricter than the required level for "confidential privacy" as evaluated in open plan spaces (SII = 0.10), and approaches conditions that are comparable to the "minimal speech privacy" provided between closed rooms (SII = 0.03). Comparisons between the metric based evaluations of the presented signals with the results of the preference test showed that the loudness of the masker corresponded well with the predicted order of preference, and due to its strong and predictable links with signal intensity, was determined to be the most important perceptual index to minimise when designing for an acceptable masking signal. Additionally, the preference test results were found to correlate well with the instrumentally measured roughness, though this index is already low for constant, broadband noise maskers and is less straightforward to reduce without affecting other desired properties of the masking signal.

When speech privacy control systems are situated in reverberant spaces, the reverberation can impede the formation of listening zones and degrade the intelligibility of the target message. By placing the sound zones within the critical distance of the source, such that the direct field from the loudspeaker array dominates over the reverberant field in both the bright and dark zones, these effects can be minimised. Results from four reverberant spaces, simulated using an image source model, demonstrate that a higher level of both acoustic contrast and speech intelligibility contrast can be achieved by incorporating measurements of the room reverberation into the transfer responses used to generate the listening zones, compared to the case where a free-field radiation model is assumed. Experimental validation of one of the simulated cases yields similar conclusions, but the performance differential between using measured and free-field transfer responses was smaller. This was due to the physical loudspeaker array sources providing a lower level of excitation to the reverberant field than the monopole sources that were assumed in the image source modelling. The experimental results also quantified the effects of regularisation in the ACC process, when using free-field and measured transfer responses. When free-field responses were used, both under- and over-regularising the system led to lower levels of bright zone speech intelligibility compared to when the regularisation parameter was set optimally. However, for the tested condition, a bright zone SII in excess of 0.75, corresponding to a good

standard of communication, was achieved over a range of the regularisation parameter spanning five orders of magnitude, indicating relative insensitivity to the setting of this parameter. This result also indicates that a high level of system performance can be achieved using free-field responses, thereby avoiding the expense and relative inflexibility of conducting in-situ transfer response measurements. The use of free-field responses must be balanced against the required increase in the masking signal levels, a change that may be more or less tolerable depending on the level of ambient noise in the reproduction environment.

Ambient noise will reduce the intelligibility of speech in both the bright and dark listening zones of a speech privacy control system, and has the potential to be incorporated into predictions of the speech intelligibility, thereby reducing the level of artificial masking that must be output by the array. The advantage of incorporating this additional masking has been quantified by calculating the level of acoustic contrast that a system must reproduce in order to achieve a certain pair of speech intelligibility constraints. For the tested array configuration, 12 dB of mid-frequency acoustic contrast is required if additional masking is introduced, while over 20 dB contrast is required when a single sound zoning process is used and the masking effect is solely provided by the ambient noise. As well as requiring a greater degree of contrast, the latter configuration is also potentially impractical as the masking ability of ambient noise is difficult to predict reliably due to its temporal fluctuation and spatial distribution. However, the background noise in the reproduction environment, defined as the contributions to the ambient noise that are exceeded in level for 90% of the time, is temporally stable by definition, and has been shown using four typical examples to be more spatially diffuse than the corresponding samples of ambient noise. This results in a reduction in the SRM, which would have prevented the ambient noise from being useful as an additional source of masking in the speech privacy control problem. The spatial diffusivity of the background noise in the considered environments means that a single remote measurement can be assumed to correspond to the background noise measured in each listening zone. In addition, the relative temporal stability of the background noise provides time for an adaptive algorithm to update the masking predictions to account for long-term changes in the background noise, for example caused by variations in occupancy patterns, traffic noise ingress or ventilation settings.

10.2 Suggestions for Future Work

In Chapter 6, a series of listening tests were described. These consisted of a speech intelligibility test and a paired preference evaluation. Directly after the completion of the paired preference tests, each participant was asked to freely describe the reasons for their preference judgments, or any features of the presented signals that were useful to them in making their decisions. Many of the listeners attributed the stimuli in the test to naturally occurring or artificial noise sources, despite all of the presented stimuli being formed of filtered random noise. No situational context was provided to the listeners in the test, other than the surroundings of the soundproofed laboratory in which the test was conducted, but listeners unanimously preferred sounds that appeared to them to originate from natural sources, as opposed to artificial sources. This attribution of meaning to random noise signals suggests that even tenuous links to familiar sounds can significantly impact how masking signals are perceived.

When a speech privacy control system is evaluated in-situ, therefore, it is expected that the acceptability of the system will be strongly affected by the degree to which the masking signal matches the acoustical and visual context. For example, a speech privacy control system could be installed in a vehicle to facilitate private telecommunications [10]. It is conceivable that in order to reduce the conspicuity of the masking signal in this situation, the masker could be designed such that it has a similar character to the existing noise within the cabin, for example from the engine and tyres, or simulate ecologically valid noise sources such as air rushing through air conditioning vents. However, the spectral and temporal profile of such a signal may not lend itself to the effective masking of speech, and as a result may have to be increased in level or adjusted in spectral content in order to carry out its intended function. While speech-shaped noise has been shown in this thesis to be effective at masking speech, and preferred over low-pass filtered alternatives, it may be less acceptable in this situation, particularly in luxury vehicles where maintaining a "brand sound" is the subject of extensive investment [263].

A structured investigation into the context-dependence of masking signal preference is therefore proposed as the main suggestion for future work. As well as providing practical guidance for the design of masking signals in particular applications of speech privacy control, it is likely that such an investigation will also provide specific details on the required locations and sizes of the listening zones in different scenarios. In addition to this, the proposed investigation may reveal situation-dependent conclusions on how systems must address reverberation, and further clarity on the most acceptable methods to account for variation in the ambient noise level, beyond the general discussion of these factors that has been provided in this thesis.

Appendix A

Derivation of the Pressure Matching Method

As the name implies, the pressure matching method differs from ACC in its requirement of a specific sound pressure field $\mathbf{p}$ to be reproduced, or "matched". This $(N \times 1)$ complex vector, where N is the total number of bright and dark zone microphones, is defined at each frequency. Often in the literature, due to its ubiquity and compact representation, the elements of the target pressure field $\mathbf{p}$, corresponding to each bright zone microphone located at the 3D position vector $\mathbf{y}$, are defined by the plane wave

$$p_b(\mathbf{y}) = Ce^{-j\mathbf{k}\cdot\mathbf{y}}, \tag{A.1}$$

where $\mathbf{k}$ is the wavenumber vector, and this is used to form the vector $\mathbf{p}_b$. The remaining elements of $\mathbf{p}$, which define the dark zone, are given by a vector of zeros, such that

$$\mathbf{p} = \begin{bmatrix} \mathbf{p}_b \\ \mathbf{0} \end{bmatrix}. \tag{A.2}$$

The pressure generated by the source array when each element is weighted by the $M \times 1$ complex vector $\mathbf{q}$ is given by

$$\hat{\mathbf{p}} = \mathbf{Z}\mathbf{q}, \tag{A.3}$$

where, as described above for ACC, the $N \times M$ matrix of transfer responses $\mathbf{Z}$ could be estimated based on the system geometry, or specified by inputting measured transfer response data at each frequency.

The matching of the target pressure $\mathbf{p}$ to the pressure from the array $\hat{\mathbf{p}}$ is formalised by defining the complex error vector $\mathbf{e} = \mathbf{p} - \hat{\mathbf{p}}$. The overall error J is then formed by summing the squared contributions from each microphone to give

$$J = \mathbf{e}^H \mathbf{e} = [\mathbf{p} - \mathbf{Z}\mathbf{q}]^H [\mathbf{p} - \mathbf{Z}\mathbf{q}]. \tag{A.4}$$

Provided that the number of sensors exceeds or is equal to the number of sources, the solution to the least squares problem is given by

$$\mathbf{q}_0 = [\mathbf{Z}^H \mathbf{Z}]^{-1} \mathbf{Z}^H \mathbf{p}, \tag{A.5}$$

noting that in the square case where $N = M$, the pseudoinverse $[\mathbf{Z}^H \mathbf{Z}]^{-1} \mathbf{Z}^H$ is identical to the standard matrix inverse $\mathbf{Z}^{-1}$.

As was noted in the derivation for ACC, the matrix $\mathbf{Z}^H \mathbf{Z}$, which must be inverted, can be ill-conditioned for certain system geometries and at certain frequencies. This indicates the potential for small errors in the positioning of loudspeakers, or even numerical calculations to manifest themselves as large errors in the calculated optimal value of $\mathbf{q}$. To compensate for this eventuality, the cost function J in Equation A.4 may be modified to include a penalty term proportional to the array effort [63]:

$$J_{LS} = \mathbf{e}^H \mathbf{e} + \beta [\mathbf{q}^H \mathbf{q}], \tag{A.6}$$

where β is a regularisation parameter. The resulting least squares solution is given by

$$\mathbf{q}_{LS} = [\mathbf{Z}^H \mathbf{Z} + \beta \mathbf{I}]^{-1} \mathbf{Z}^H \mathbf{p}. \tag{A.7}$$

Appendix B

Comparison of Speech Reception Threshold between Native and Non-Native English Speakers

In the speech intelligibility and paired preference tests described in Chapter 6, 21 participants were recruited. Participants were eligible for testing if they were over the age of 18, had self-reported normal hearing, and were fluent in the English language. Of these, 12 were native English speakers, and 9 had English as an additional language. There is consensus in the literature that the ability to recognise speech in noise is dependent on native language status, and that this difference is not observed as significantly when considering the task of speech perception in quiet [264]. Despite this, it was decided not to require participants to have English as a first language, as a participant group containing only native speakers cannot adequately represent the entire listening population, for whom the designed systems must work well. The inclusion of non-native speakers gives information about the minimum SNR requirements in the bright zone, as this listener group are likely to be affected more significantly by the leakage of the masker into the bright zone.

To reduce the influence of language proficiency on the results of the intelligibility test, various factors in the test design were chosen to limit the difference in the SRT between native and non-native speakers of English. These design choices are outlined in this Appendix, based on a review of investigations into the intelligibility test conditions that yield the most significant differences in SRT between native and non-native speakers.

Informational masking, such as that caused by competing speakers or multi-talker babble affects native and non-native speakers differently [265, 266]. Engen demonstrated that English native speakers with no Mandarin lanugage proficiency received a larger release from masking in Mandarin babble than do native speakers of Mandarin [266]. This shows that familiarity with the language used as a babble-type or competing speech masker causes additional informational masking. Additionally, this result establishes a link between the effectiveness of a masker and the informational similarity between the intended speech and the masker. Therefore, choosing

a babble-type masker that uses samples from a particular language would affect listeners differently based on their proficiency in that language, in addition to the difficulties that non-native listeners already have when faced with energetic masking alone [265]. The consequence for the present speech intelligibility tests is that purely energetic maskers are the best choice for achieving similar SRTs between native and non-native listeners. The effect of the choice and number of talkers was studied by Bradlow and Pisoni, and it was found that using a single talker across a test resulted in better word recognition by non-native listeners, compared to a test containing multiple talkers, thus closing the SRT gap between native and non-native listeners [267].

Studies have also been carried out to determine the effect of test format, as well as content. In tests of the German matrix test [207] against the (German language) Triple Digit Test and an open-set sentence test, it was found that the reduced vocabulary of the matrix test decreased the effect of language skill on SRT, compared to the open-set sentence test [206]. The difference in mean SRT between non-native speakers across four levels of language proficiency and native speakers was 3 dB, but differences in SRT between the highest and lowest proficiency non-native groups was 2.5 dB [268]. This shows that unless specifically controlled for, different levels of language proficiency can significantly affect the standard deviation of the SRT among non-native speakers, but the listeners in the highest proficiency group could achieve similar performance to participants listening in their first language. In the study by Warzybok et al. [268], all study participants had only been living in Germany for a maximum of two months, so had limited exposure to natural language. Conversely, the non-native English-speaking participants in the tests described in this thesis had all been living and studying in England for significantly longer, so are more likely to have higher language proficiency than the average listeners in the study by Warzybok et al. Furthermore, when testing with meaningful sentences, it was found that non-native listeners were less able to utilise contextual cues to correctly guess misheard words, compared to native listeners [266, 268]. These results demonstrate that closer performance is achieved between native and non-native speakers for a closed-set, meaningless sentence test such as the matrix test, compared to alternative open-set tests with meaningful sentences.

The present discussion shows that the matrix test format, using steady, speech-shaped noise as a masker, and a single talker is an appropriate format for use with both native and non-native listeners, particular compared to open-set, familiar sentence-in-noise tests. Figure B.1 shows the distribution of SRTs, grouped by English language status; EFL indicates English is the participant's First Language, EAL means English is an Additional Language. The difference in SRTs between EAL and EFL groups Δ_{50} ranges from 1.7 to 6.6 dB, and consistently shows that English native speakers have a lower SRT, finding it easier to understand speech in noise than those with English as an additional language. The mean difference in SRT is 2.8 dB, consistent with the difference in SRT found by Warzybok et al. [268] using the German matrix test. The largest deviation was found for the W_A condition, in which consonant sounds were poorly masked - in this condition, native listeners were disproportionately able to correctly select words based on this information, compared to non-native listeners.

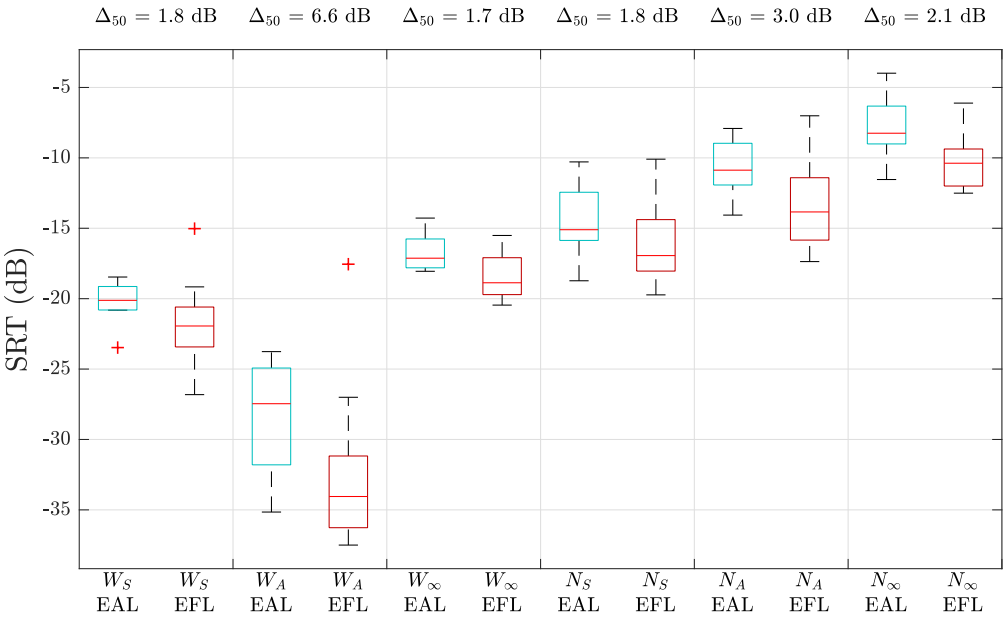


FIGURE B.1: Speech Reception Threshold distributions for each array configuration, grouped by language status. EAL = English as an Additional Language, EFL = English as First Language.

Appendix C

A Low-Cost Loudspeaker Array for Personal Audio with Enhanced Vertical Directivity

Parts of this Appendix were published as "A Low-Cost Loudspeaker Array for Personal Audio with Enhanced Vertical Directivity" in the proceedings of the 49^{th} International Congress and Exposition on Noise Control Engineering.

A key practical constraint placed on the personal audio system geometries discussed in this thesis has been to consider only linear loudspeaker array designs. This layout allows for compact loudspeaker arrays that can control sound fields in a single, usually horizontal plane. However, this arrangement provides no control of the vertical directivity, meaning that the reverberant field in the room will be unnecessarily excited when the loudspeaker array is in operation. This additional excitation can reduce the acoustic contrast that is achievable between the zones. In the analyses of personal audio system performance in reverberant spaces conducted by Simón-Gálvez et al. [11, 229], vertical directivity control was discussed as a way of reducing the energisation of the reverberant field, rather than solely focussing on controlling the azimuthal direction of beams. To avoid additional signal processing overhead, the vertical directivity was manually constrained by building the loudspeaker drivers into phase-shift enclosures [41, 229], and by effectively increasing the vertical aperture of the array. This was achieved by stacking four array elements vertically, and driving them in parallel. The resultant decrease in acoustic contrast between free-field and reverberant conditions was reduced by around 5 dB in the 1 - 8 kHz frequency region [11].

Following this concept, and with the intention of lowering the barrier to entry for practical personal audio system research, a design for a low-cost, 8-channel loudspeaker array design was made freely available online [240]. Figure C.1 shows the prototype 8-channel array, which includes on-board digital to analogue conversion using a consumer-grade USB sound card and amplification using 5-Watt Class D amplifier circuits. Each loudspeaker has an individual sealed cabinet, as shown in the exploded view in Figure C.2 and the entire cabinet design can be constructed from a single 1220 x 610mm sheet of 6mm plywood, again selected with cost reduction

FIGURE C.1: 8-Channel loudspeaker array prototype. Left: Front view; Right: Rear view with back panel removed.

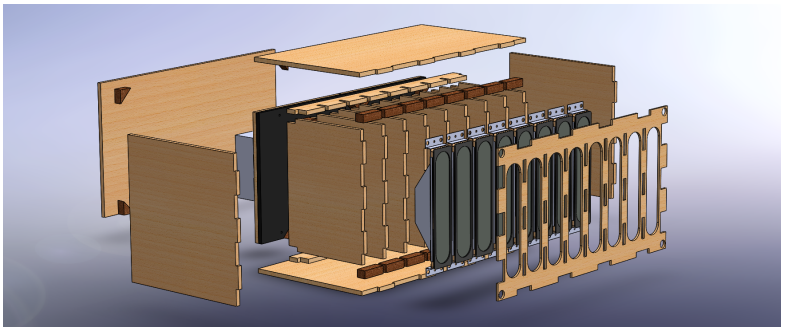


FIGURE C.2: Exploded view of the prototype 8-channel loudspeaker array design.

in mind. To enhance the vertical directivity of the array, without the additional expense of using multiple loudspeaker drivers in parallel, high aspect ratio loudspeaker drivers designed for fitment in televisions were used in the design. This choice maintains a small spacing between drivers to avoid spatial aliasing artefacts at high frequencies, as recommended in Section 5.4.

The remainder of this Appendix documents the anechoic performance of the system, using a similar methodology to that used in Section 8.4. The performance of the system is discussed when both measured and modelled transfer responses are used in the ACC process. This comparison is further motivated by a desire to reduce cost, as additional hardware and time is required to measure the transfer responses, and it is important to gauge the benefits of this additional expense, given that modelled transfer responses can be generated instantly with appropriate software. In the following discussion, two types of analytical transfer responses are considered, one where the array elements are modelled as point monopole sources, and the other where a baffled rectangular piston model is used to represent the high aspect ratio loudspeakers [269].

Figures C.3 and C.4 show the horizontal and vertical directivity of the array when it is configured to produce a single forward bright zone of width 30 degrees. The maximum horizontal directivity is achieved when measured transfer responses are used in the ACC process, and there are only marginal differences between the monopole and baffled rectangular piston models for the presented zonal geometry, across the frequency range. For all types of response, the rear radiation from the array is attenuated by 5 dB at 500 Hz and 10 dB at 2 kHz. The vertical directivity shown in Figure C.4 is unaffected by the choice of transfer response estimate as the generation of zonal filters uses no information from the vertically oriented measurement microphones. Extending the bright zone to a three-dimensional, 30° cone and including the off-axis vertical measurement positions in the dark zone would provide additional constraints to the

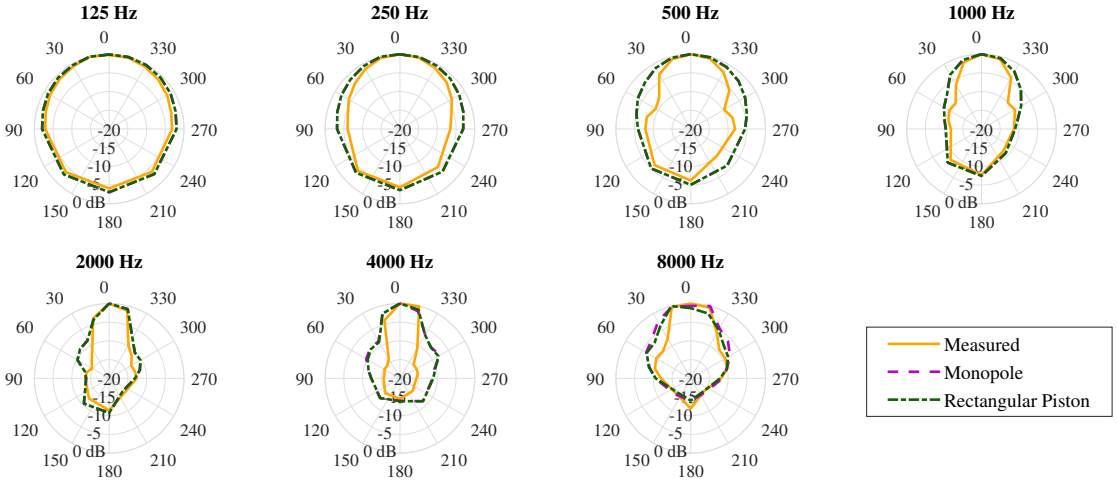


FIGURE C.3: Horizontal directivity with a single forward zone. Each plot shows directivity results averaged over an octave band.

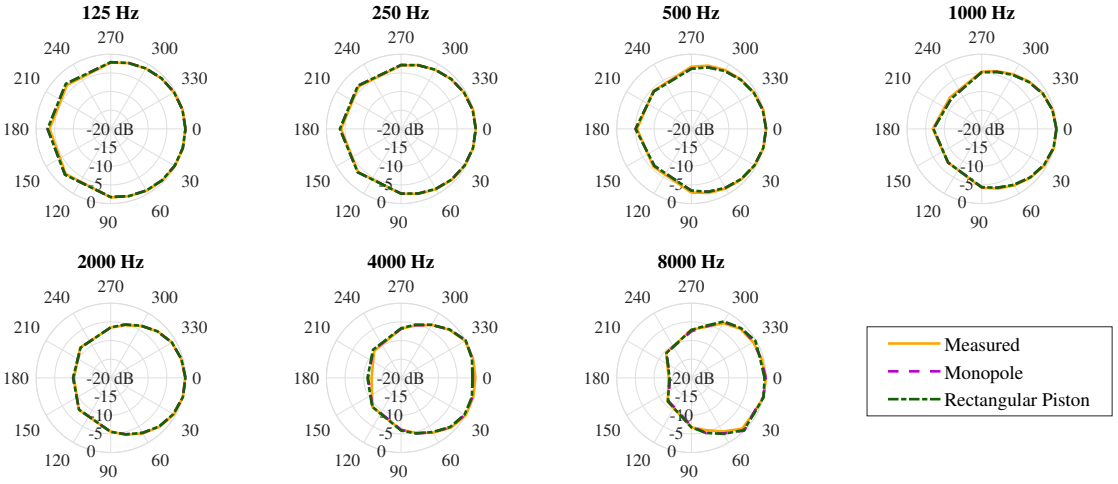


FIGURE C.4: Vertical directivity with a single forward zone. Each plot shows directivity results averaged over an octave band.

beamforming process in a direction in which the array has no beamforming capability. These additional constraints would only serve to increase the condition number of the squared transfer response matrix $[\mathbf{Z}_d^H \mathbf{Z}_d]$ from Equation 5.11 and thus may reduce robustness to environmental or sensitivity changes over time.

The dark zone in the ACC process is defined over the remaining horizontal angles, and the acoustic contrast performance and on-axis frequency response of the array with this configuration is presented in Figure C.5. At low frequencies, acoustic contrast is limited by the overall aperture of the array; for wavelengths significantly longer than this dimension, the entire array appears as a single monopole source. The most pronounced differences between the measured and modelled results occur at frequencies greater than 3 kHz. Below this frequency, in terms of acoustic contrast, the system based on measured responses outperforms those using modelled responses by around 5 dB. This is caused by mismatches in the sensitivity and the frequency response between the array elements. The impact of this mismatch was minimised by individually adjusting the gains of the in-built amplifier circuits such that broadband noise emitted from each loudspeaker,

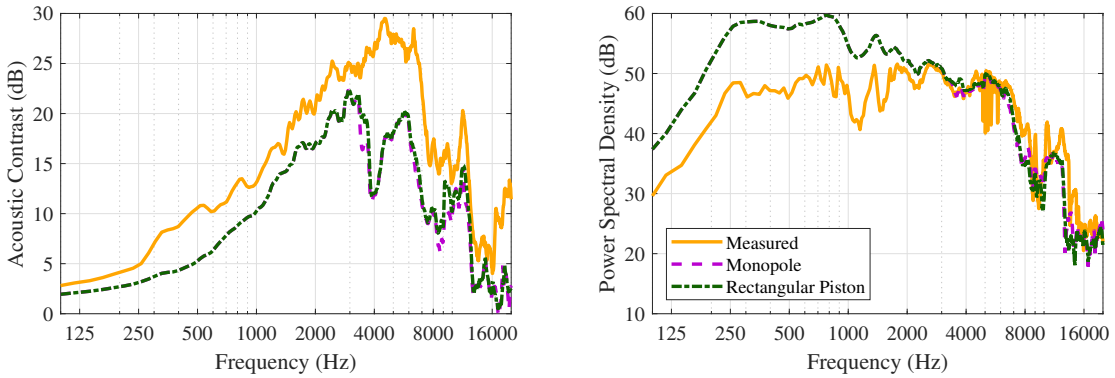


FIGURE C.5: Left: Experimental acoustic contrast using measured and modelled responses to optimise array filters. Right: On-axis (bright zone) SPL produced by array using measured and modelled responses.

recorded at an on-axis point in the far field, matched within 1 dB SPL prior to the transfer response measurements being taken.

In addition to the sensitivity issues described above, at frequencies above 3 kHz, discrepancies between the modelled and true transfer responses cause the acoustic contrast to decrease, while the measured contrast continues to increase up to 4.4 kHz, whereupon spatially aliased side-lobes begin to impinge on the dark zone. At 3 kHz, the wavelength of sound in air is comparable to the dimensions of the loudspeaker cabinet, meaning that both the monopole model and the model of the rectangular piston in an infinite baffle provide poor approximations to the true transfer responses. Given the marginal performance increase when moving from the simple monopole source model to the more complex rectangular piston model exhibited in Figures C.3-C.5, it is hypothesized that the mismatch between array elements is the dominant source of error between the modelled transfer responses, which assume perfectly matched drivers, and the true transfer responses. Accordingly, without measurements and subsequent digital correction for this mismatch, predicting the radiation pattern using a more detailed technique such as finite element modelling is unlikely to yield significantly better performance. In practical, reverberant environments, lower levels of acoustic contrast are expected overall due to increased leakage from the bright zone into the dark zone, as discussed in Chapter 8.

References

[1] Frank Fahy. Sound in Enclosures. In *Foundations of Engineering Acoustics*, Chapter 9, pages 236–269. Elsevier, 1st Edition, 2001.

[2] Stephen J. Elliott, Jordan Cheer, Harry Murfet, and Keith R. Holland. Minimally radiating sources for personal audio. *The Journal of the Acoustical Society of America*, 128(4):1721–8, 2010. doi: 10.1121/1.3479758.

[3] Jon Francombe, Russell Mason, Martin Dewhirst, and Søren Bech. Elicitation of attributes for the evaluation of audio-on-audio interference. *The Journal of the Acoustical Society of America*, 136(5):2630–2641, 2014. doi: 10.1121/1.4898053.

[4] W. F. Druyvesteyn and J. Garas. Personal Sound. *Journal of the Audio Engineering Society*, 45(9):685–701, 1997. URL `http://www.aes.org/e-lib/browse.cfm?elib=7843`. Accessed 31/07/2020.

[5] Ji-Ho Chang, Chan-Hui Lee, Jin-Young Park, and Yang-Hann Kim. A realization of sound focused personal audio system using acoustic contrast control. *Journal of the Acoustical Society of America*, 125(4):2091–2097, 2009. doi: 10.1121/1.3082114.

[6] Marcos F. Simón Gálvez, Stephen J. Elliott, and Jordan Cheer. Personal Audio Loudspeaker Array as a Complementary TV Sound System for the Hard of Hearing. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E97.A(9):1824–1831, 2014. doi: 10.1587/transfun.E97.A.1824.

[7] Jordan Cheer, Stephen J. Elliott, and Marcos F. Simón Gálvez. Design and implementation of a car cabin personal audio system. *Journal of the Audio Engineering Society*, 61(6): 412–424, 2013. URL `http://www.aes.org/e-lib/browse.cfm?elib=16832`. Accessed 31/07/2020.

[8] Jacob Donley, Christian Ritz, and W. Bastiaan Kleijn. Improving Speech Privacy in Personal Sound Zones. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016. doi: 10.1109/ICASSP.2016.7471687.

[9] Jacob Donley, Christian H. Ritz, and W. B. Kleijn. Multizone Soundfield Reproduction With Privacy and Quality Based Speech Masking Filters. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26(6):1037–1051, 2018. doi: 10.1109/TASLP.2018.2798804.

[10] Kudzayi Chwioko, Delphine Nourzad, and Xavier Vinamata. Apparatus and Method for Privacy Enhancement, March 15th 2018. WIPO Patent WO2018046185A1.

[11] Marcos F. Simón-Gálvez, Stephen J. Elliott, and Jordan Cheer. The effect of reverberation on personal audio devices. *The Journal of the Acoustical Society of America*, 135(5):2654–2663, 2014. doi: 10.1121/1.4869681.

[12] Marek Olik, Philip J. Jackson, and Philip Coleman. Influence of low-order room reflections on sound zone system performance. In *Proc. Meetings on Acoustics*, volume 19, 2013. doi: 10.1121/1.4800873.

[13] Khan Richard Baykaner, Christopher Hummersone, Russell Mason, and Søren Bech. The acceptability of speech with interfering radio program material. In *Proc. 136th Audio Engineering Society Convention*, Berlin, 2014. URL `http://www.aes.org/e-lib/browse.cfm?elib=17167`. Accessed 31/07/2020.

[14] Jung-Woo Choi and Yang-Hann Kim. Generation of an acoustically bright zone with an illuminated region using multiple sources. *Journal of the Acoustical Society of America*, 111(4):1695–1700, 2002. doi: 10.1121/1.1456926.

[15] Ferdinando Olivieri, Filippo Maria Fazi, Philip A Nelson, and Simone Fontana. Comparison of Strategies for Accurate Reproduction of a Target Signal with Compact Arrays of Loudspeakers for the Generation of Zones of Private Sound and Silence. *Journal of the Audio Engineering Society*, 64(11):905–917, 2016. doi: 10.17743/jaes.2016.0045.

[16] Philip Coleman, Philip J. B. Jackson, Marek Olik, Martin Møller, Martin Olsen, and Jan Abildgaard Pedersen. Acoustic contrast, planarity and robustness of sound zone methods using a circular loudspeaker array. *The Journal of the Acoustical Society of America*, 135 (4):1929–1940, 2014. doi: 10.1121/1.4866442.

[17] Taewoong Lee, Jesper Kjaer Nielsen, and Mads Graesbøll Christensen. Signal-Adaptive and Perceptually Optimized Sound Zones with Variable Span Trade-Off Filters, 2019. URL `https://arxiv.org/abs/1911.10016v2`. Accessed 31/07/2020.

[18] Jon Francombe, Philip Coleman, Marek Olik, Khan Baykaner, Philip J B Jackson, Russell Mason, Søren Bech, Jan Abildgaard Pedersen, and Martin Dewhirst. Perceptually Optimized Loudspeaker Selection for the Creation of Personal Sound Zones. In *Proc. 52nd Audio Engineering Society Conference*, Guildford, 2013. URL `http://www.aes.org/e-lib/browse.cfm?elib=16907`. Accessed 31/07/2020.

[19] Nasim Radmanesh and Ian S. Burnett. Reproduction of independent narrowband soundfields in a multizone surround system and its extension to speech signal sources. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, 2011. doi: 10.1109/ICASSP.2011.5946440.

[20] Terence Betlehem and Paul D. Teal. A constrained optimization approach for multi-zone surround sound. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, 2011. doi: 10.1109/ICASSP.2011.5946434.

[21] Martin Cooke, Catherine Mayo, and Cassia Valentini-Botinhao. Hurricane natural speech corpus - higher quality version, [sound], 2019. doi: 10.7488/ds/2482.

[22] Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, [sound], 2017. doi: 10.7488/ds/1994.

[23] Stuart John Hall. *The Development of a New English Sentence in Noise Test and an English Number Recognition Test.* MSc Thesis, University of Southampton, 2006.

[24] Daniel Wallace and Jordan Cheer. Optimisation of Personal Audio Systems for Intelligibility Contrast. In *Proc. 144th Audio Engineering Society Convention*, Milan, Italy, 2018. URL `http://www.aes.org/e-lib/browse.cfm?elib=19413`. Accessed 31/07/2020.

[25] Daniel Wallace and Jordan Cheer. Design and evaluation of personal audio systems based on speech privacy constraints. *Journal of the Acoustical Society of America*, 147(4):2271–2282, 2020. doi: 10.1121/10.0001065.

[26] Daniel Wallace and Jordan Cheer. The Design of Personal Audio Systems for Speech Transmission using Analytical and Measured Responses. In *Proc. 44th International Conference on Acoustics, Speech and Signal Processing*, Brighton, 2019. doi: 10.1109/ICASSP.2019.8683269.

[27] Daniel Wallace and Jordan Cheer. Combining Artificial and Natural Background Noise in Personal Audio Systems. In *Proc. 10th IEEE Sensor Array and Multichannel Signal Processing Workshop*, Sheffield, UK, 2018. doi: 10.1109/SAM.2018.8448847.

[28] Francis Rumsey. Sound Field Control. *Journal of the Audio Engineering Society*, 61(12):1046–1050, 2013. URL `http://www.aes.org/e-lib/browse.cfm?elib=17080`. Accessed 31/07/2020.

[29] P.A. Nelson and S.J. Elliott. *Active Control of Sound.* Academic Press, 1992.

[30] H. Zhu, R. Rajamani, and K. A. Stelson. Active control of acoustic reflection, absorption, and transmission using thin panel speakers. *The Journal of the Acoustical Society of America*, 113(2):852–870, 2003. doi: 10.1121/1.1534834.

[31] Harry F. Olson and Everett G. May. Electronic Sound Absorber. *Journal of the Acoustical Society of America*, 25(6):1130–1136, 1953. doi: 10.1121/1.1907249.

[32] A.J. Berkhout. Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, 93(5):2764–2778, 1993. doi: 10.1121/1.405852.

[33] Duane H. Cooper and Jerald L. Bauck. Prospects for Transaural Recording. *Journal of the Audio Engineering Society*, 37(1/2):3–19, 1989. URL `http://www.aes.org/e-lib/browse.cfm?elib=6108`. Accessed 31/07/2020.

[34] Francis Rumsey. Sound Field Control: Personal Sound Zones and Moving Listeners. *Journal of the Audio Engineering Society*, 64(10):808–813, 2016. URL `http://www.aes.org/e-lib/browse.cfm?elib=18521`. Accessed 31/07/2020.

[35] Philip J. Jackson, Finn Jacobsen, Philip Coleman, and Jan Abildgaard Pedersen. Sound field planarity characterized by superdirective beamforming. In *Proc. Meetings on Acoustics*, volume 19, Montreal, 2013. doi: 10.1121/1.4800877.

[36] International Organisation for Standardisation. ISO 3382 Acoustics - Measurement of room acoustic parameters – Part 3: Open plan offices, 2013.

[37] Leo Beranek and Tim Mellow. Acoustic Components. In *Acoustics - Sound Fields, Transducers and Vibration*, Chapter 4, pages 119–198. Academic Press, 2019.

[38] Frank Fahy. The vibrating circular piston and the cone loudspeaker. In *Foundations of Engineering Acoustics*, Chapter 6, pages 126–129. Elsevier, 1st Edition, 2001.

[39] Maya Mochizuki, Ayumu Osumi, and Youichi Ito. Privacy protection method for speech using small speakers placed around a head. In *Proc. 43rd International Congress on Noise Control Engineering*, Melbourne, 2014.

[40] Harry F. Olson. Gradient Loudspeakers. *Journal of the Audio Engineering Society*, 21 (2):86–93, 1973. URL `http://www.aes.org/e-lib/browse.cfm?elib=2006`. Accessed 31/07/2020.

[41] Thomas J Holmes. The 'Acoustic Resistance Box' - A Fresh Look at an Old Principle. In *Proc. 77th Audio Engineering Society Convention*, Hamburg, 1985. URL `http://www.aes.org/e-lib/browse.cfm?elib=11549`. Accessed 31/07/2020.

[42] F. Joseph Pompei. *Sound From Ultrasound: The Parametric Array as an Audible Sound Source*. Ph.D. Thesis, Dept. of Architecture, MIT, 2002. URL `http://hdl.handle.net/1721.1/7987`. Accessed 31/07/2020.

[43] John D. Meyer and Paul J. Kohut. Broadband acoustical transmitting system, October 13th 1997. US Patent US5821470.

[44] Meyer Sound Laboratories Inc. SB-1 Resources, 1997. URL `https://meyersound.com/download/sb-1-resources/`. Accessed 31/07/2020.

[45] Marcos F. Simón Gálvez, Dylan Menzies, and Filippo Maria Fazi. Dynamic audio reproduction with linear loudspeaker arrays. *Journal of the Audio Engineering Society*, 67(4): 190–200, 2019. doi: 10.17743/jaes.2019.0007.

[46] Ferdinando Olivieri, Mincheol Shin, Filippo M. Fazi, Philip A. Nelson, and Peter Otto. Loudspeaker array processing for multi-zone audio reproduction based on analytical and measured electroacoustical transfer functions. In *Proc. 52nd Audio Engineering Society Conference*, Guildford, 2013. URL `http://www.aes.org/e-lib/browse.cfm?elib=16897`. Accessed 31/07/2020.

[47] Peter J. Westervelt. Parametric Acoustic Array. *The Journal of the Acoustical Society of America*, 35(4):535–537, 1963. doi: 10.1121/1.1918525.

[48] Thomas D. Kite, John T. Post, and Mark F. Hamilton. Parametric array in air: Distortion reduction by preprocessing. *The Journal of the Acoustical Society of America*, 103(5):2871, 1998. doi: 10.1121/1.421645.

[49] F. Joseph Pompei. The use of airborne ultrasonics for generating audible sound beams. *Journal of the Audio Engineering Society*, 47(9):726–731, 1999. URL `http://www.aes.org/e-lib/browse.cfm?elib=12092`. Accessed 31/07/2020.

[50] Nobuo Tanaka and Motoki Tanaka. Active noise control using a steerable parametric array loudspeaker. *The Journal of the Acoustical Society of America*, 127(6):3526–3537, 2010. doi: 10.1121/1.3409483.

[51] Directional Audio. Ultrasonic speaker hire, 2020. URL `https://www.directionalaudio.co.uk/directional-speaker-hire/`. Accessed 30/07/2020.

[52] Woon Seng Gan, Jun Yang, and Tomoo Kamakura. A review of parametric acoustic array in air. *Applied Acoustics*, 73(12):1211–1219, 2012. doi: 10.1016/j.apacoust.2012.04.001.

[53] Mincheol Shin, Sung Q. Lee, Filippo M. Fazi, Philip A. Nelson, Daesung Kim, Semyung Wang, Kang Ho Park, and Jeongil Seo. Maximization of acoustic energy difference between two spaces. *The Journal of the Acoustical Society of America*, 128(1):121–131, 2010. doi: 10.1121/1.3438479.

[54] Ole Kirkeby and Philip A. Nelson. Reproduction of plane wave sound fields. *The Journal of the Acoustical Society of America*, 94(5):2992–3000, 1993. doi: 10.1121/1.407330.

[55] Barry D. Van Veen and Kevin M. Buckley. Beamforming: A Versatile Approach to Spatial Filtering. *IEEE ASSP Magazine*, 5(2):4–24, 1988. doi: 10.1109/53.665.

[56] C.L. Dolph. A Current Distribution for Broadside Arrays Which Optimizes the Relationship between Beam Width and Side-Lobe Level. *Proc. IRE*, 34(6):335–348, 1946. doi: 10.1109/JRPROC.1946.225956.

[57] P.A. Nelson, A.R.D. Curtis, S.J. Elliott, and A.J. Bullmore. The minimum power output of free field point sources and the active control of sound. *Journal of Sound and Vibration*, 116(3):397–414, 1987. doi: 10.1016/S0022-460X(87)81373-1.

[58] Stephen J. Elliott, Jordan Cheer, Jung-Woo Choi, and Youngtae Kim. Robustness and Regularization of Personal Audio Systems. *IEEE Transactions on Audio, Speech and Language Processing*, 20(7):2123–2133, 2012. doi: 10.1109/TASL.2012.2197613.

[59] Mark Poletti. An investigation of 2D multizone surround sound systems. In *Proc. 125th Audio Engineering Society Convention*, San Francisco, 2008. URL `http://www.aes.org/e-lib/browse.cfm?elib=14703`. Accessed 31/07/2020.

[60] Marek Olik, Jon Francombe, Philip Coleman, Philip J B Jackson, Martin Olsen, Martin Møller, Russell Mason, and Søren Bech. A Comparative Performance Study of Sound Zoning Methods in a Reflective Environment. In *Proc. 52nd Audio Engineering Society Conference*, Guildford, 2013. URL `http://www.aes.org/e-lib/browse.cfm?elib=16914`. Accessed 31/07/2020.

[61] Finn Jacobsen, Martin Olsen, Martin Møller, and Finn T. Agerkvist. A Comparison of Two Strategies for Generating Sound Zones in a Room. In *Proc. 18th International Congress on Sound and Vibration*, Rio De Janiero, 2011. URL `https://orbit.dtu.dk/files/5677256/ICSV18FJ[1].pdf`. Accessed 31/07/2020.

[62] Terence Betlehem, Wen Zhang, Mark A. Poletti, and Thushara D. Abhayapala. Personal Sound Zones. *IEEE Signal Processing Magazine*, 32(2):81–91, 2015. doi: 10.1109/MSP.2014.2360707.

[63] Mincheol Shin, Filippo M. Fazi, Philip A. Nelson, and Fabio C. Hirono. Controlled sound field with a dual layer loudspeaker array. *Journal of Sound and Vibration*, 333(16):3794–3817, 2014. doi: 10.1016/j.jsv.2014.03.025.

[64] Ferdinando Olivieri, Filippo Maria Fazi, Simone Fontana, Dylan Menzies, and Philip Arthur Nelson. Generation of Private Sound with a Circular Loudspeaker Array

and the Weighted Pressure Matching Method. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 25(8):1579–1591, 2017. doi: 10.1109/TASLP.2017.2700945.

[65] Ji-Ho Chang and Finn Jacobsen. Sound field control with a circular double-layer array of loudspeakers. *Journal of the Acoustical Society of America*, 131(6):4518–4525, 2012. doi: 10.1121/1.4714349.

[66] Taewoong Lee, Jesper Kjaer Nielsen, Jesper Rindom Jensen, and Mads Graesbøll Christensen. A Unified Approach to Generating Sound Zones Using Variable Span Linear Filters. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, 2018. doi: 10.1109/ICASSP.2018.8462477.

[67] Stephen J. Elliott and Matthew Jones. An active headrest for personal audio. *The Journal of the Acoustical Society of America*, 119(5):2702, 2006. doi: 10.1121/1.2188814.

[68] Jordan Cheer, Stephen J. Elliott, Youngtae Kim, and Jung-Woo Choi. Practical Implementation of Personal Audio in a Mobile Device. *Journal of the Audio Engineering Society*, 61 (5):290–300, 2013. URL `http://www.aes.org/e-lib/browse.cfm?elib=16825`. Accessed 31/07/2020.

[69] X. Liao, S. Elliott, J. Cheer, and S. Zheng. Design of a Loudspeaker Array for Personal Audio in a Car Cabin. *Journal of the Audio Engineering Society*, 65(3):226–238, 2017. doi: 10.17743/jaes.2016.0065.

[70] Marcos F. Simón Gálvez and Stephen J. Elliott. The design of a personal audio superdirective array in a room. In *Proc. 52nd Audio Engineering Society Conference*, Guildford, 2013. URL `http://www.aes.org/e-lib/browse.cfm?elib=16918`. Accessed 31/07/2020.

[71] University of Surrey. Perceptually Optimised Sound Zones, 2018. URL `http://iosr.surrey.ac.uk/projects/POSZ/index.php`. Accessed 31/07/2020.

[72] Marek Olik, Philip J. B. Jackson, Philip Coleman, and Jan Abildgaard Pedersen. Optimal source placement for sound zone reproduction with first order reflections. *The Journal of the Acoustical Society of America*, 136(6):3085–3096, 2014. doi: 10.1121/1.4898423.

[73] Philip Coleman, Philip J. Jackson, Marek Olik, Martin Olsen, Martin Møller, and Jan Abildgaard Pedersen. The influence of regularization on anechoic performance and robustness of sound zone methods. In *Proc. Meetings on Acoustics*, volume 19, Montreal, 2013. doi: 10.1121/1.4799031.

[74] Philip Coleman, Philip Jackson, Marek Olik, and Jan Abildgaard Pedersen. Optimizing the planarity of sound zones. In *Proc. 52nd Audio Engineering Society Conference*, Guildford, 2013. URL `http://www.aes.org/e-lib/browse.cfm?elib=16916`. Accessed 31/07/2020.

[75] Jon Francombe, Russell Mason, Martin Dewhirst, and Søren Bech. A Model of Distraction in an Audio-on-Audio Interference Situation with Music Program Material. *Journal of the Audio Engineering Society*, 63(1/2):63–77, 2015. doi: 10.17743/jaes.2015.0006.

[76] Jon Francombe, Russell Mason, Martin Dewhirst, and Søren Bech. Modelling listener distraction resulting from audio-on-audio interference. In *Proc. Meetings on Acoustics*, volume 19, 2013. doi: 10.1121/1.4799636.

[77] J. Rämö, S. Bech, and S. H. Jensen. Real-time perceptual model for distraction in interfering audio-on-audio scenarios. *IEEE Signal Processing Letters*, 24(10):1448–1452, Oct 2017. doi: 10.1109/LSP.2017.2733084.

[78] Khan Richard Baykaner, Christopher Hummersone, Russell Mason, and Søren Bech. The prediction of the acceptability of auditory interference based on audibility. In *Proc. 52nd Audio Engineering Society Conference*, Guildford, 2013. URL `http://www.aes.org/e-lib/browse.cfm?elib=16908`. Accessed 31/07/2020.

[79] Hugo Fastl and Eberhard Zwicker. *Psychoacoustics: Facts and models*. Springer, 2007. doi: 10.1007/978-3-540-68888-4.

[80] T. Houtgast and H. J. M. Steeneken. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *Journal of the Acoustical Society of America*, 77(3):1069–1077, 1985. doi: 10.1121/1.392224.

[81] Martin Bo Møller and Martin Olsen. Sound Zones: On Envelope Shaping of FIR Filters. In *Proc. 24th International Congress on Sound and Vibration*, London, 2017.

[82] Jacob Donley and Christian Ritz. Multizone Reproduction of Speech Soundfields: A Perceptually Weighted Approach. In *Proc. APSIPA Annual Summit and Conference*, 2015. doi: 10.13140/RG.2.1.3319.2162.

[83] Jungsoo Kim and Richard de Dear. Workspace satisfaction: The privacy-communication trade-off in open-plan offices. *Journal of Environmental Psychology*, 36:18–26, 2013. doi: 10.1016/j.jenvp.2013.06.007.

[84] M. Navai and J.A. Veitch. Acoustic Satisfaction in Open-Plan Offices : Review and Recommendations IRC-RR-151. Technical report, Institute for Research in Construction, 2003. doi: 10.4224/20386513.

[85] A. Haapakangas, E. Kankkunen, V. Hongisto, P. Virjonen, D. Oliva, and E. Keskinen. Effects of five speech masking sounds on performance and acoustic satisfaction. implications for open-plan offices. *Acta Acustica united with Acustica*, 97(4):641–655, 2011. doi: 10.3813/AAA.918444.

[86] W. J. Cavanaugh, W. R. Farrell, P. W. Hirtle, and B. G. Watters. Speech Privacy in Buildings. *The Journal of the Acoustical Society of America*, 34(4):475–492, 1962. doi: 10.1121/1.1918154.

[87] Jukka Keränen, Petra Virjonen, David Oliva Elorza, and O. Valtteri Hongisto. Design of room acoustics for open offices. *Scandinavian Journal of Work, Environment and Health, Supplement*, 4:46–49, 2008.

[88] A. C. C. Warnock. Acoustical privacy in the landscaped office. *Journal Of The Acoustical Society Of America*, 53(6):1535–1543, 1973. doi: 10.1121/1.1913498.

[89] T.R. Horrall. Personal sound masking system, May 3rd 2005. US Patent 6,888,945.

[90] Valtteri Hongisto, David Oliva, and Laura Rekola. Subjective and objective rating of spectrally different pseudorandom noises—Implications for speech masking design. *The Journal of the Acoustical Society of America*, 137(3):1344–1355, 2015. doi: 10.1121/1.4913273.

[91] Lucas Lenne, Patrick Chevret, and Julien Marchand. Long-term effects of the use of a sound masking system in open-plan offices: A field study. *Applied Acoustics*, 158:107049, 2020. doi: 10.1016/j.apacoust.2019.107049.

[92] Jung-Min Lee, Tae-Woong Lee, Jin-Young Park, and Yang-Hann Kim. Generation of a private listening zone; acoustic parasol. In *Proc. 20th International Congress on Acoustics*, Sydney, 2010.

[93] Jin Young Park, Ji Ho Chang, and Yang Hann Kim. Generation of independent bright zones for a two-channel private audio system. *Journal of the Audio Engineering Society*, 58 (5):382–393, 2010. URL `http://www.aes.org/e-lib/browse.cfm?elib=15452`. Accessed 31/07/2020.

[94] Yefeng Cai, Ming Wu, and Jun Yang. Design of a time-domain acoustic contrast control for broadband input signals in personal audio systems. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, 2013. doi: 10.1109/ICASSP.2013.6637665.

[95] International Telecommunications Union. Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, 2001.

[96] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7):2125–2136, 2011. doi: 10.1109/TASL.2011.2114881.

[97] Anadi Chaman, Yu-Jeh Liu, Jonah Casebeer, and Ivan Dokmanić. Multipath-enabled private audio with noise, 2018. URL `https://arxiv.org/abs/1811.07065`. Accessed 31/07/2020.

[98] Yu-Jeh Liu, Jonah Casebeer, and Ivan Dokmanić. Cocktails, but no party: Multipath enabled private audio. In *Proc. International Workshop on Acoustic Signal Enhancement*, Tokyo, 2018. URL `https://arxiv.org/abs/1809.05862`. Accessed 31/07/2020.

[99] John S. Bradley and Bradford N. Gover. Describing Levels of Speech Privacy in Open-Plan Offices. Technical report, National Research Council Canada, 2003. doi: 10.4224/20378524.

[100] Institute of Electrical and Electronics Engineers. IEEE 297:1969 Recommended Practice for Speech Quality Measurements. Technical report, IEEE, 1969.

[101] ASTM International. E1374-18 Standard Guide for Office Acoustics and Applicable ASTM Standards, 2018.

[102] B. Hagerman. Sentences for Testing Speech Intelligibility in Noise. *Scandinavian Audiology*, 11(2):79–87, 1982. doi: 10.3109/01050398209076203.

[103] Robert L. Sherbecoe and Gerald A. Studebaker. Audibility-index functions for the connected speech test. *Ear and Hearing*, 23(5):385–398, 2002. doi: 10.1097/00003446-200210000-00001.

[104] Holosonics. Audio Spotlight Case - In-Store Campaign, 2016. URL `https://www.youtube.com/watch?v=8HGln6ooGyg`. Accessed 21/09/2020.

[105] Geo. A. Campbell. Telephonic intelligibility. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 19(109):152–159, 1910. doi: 10.1080/14786440108636784.

[106] H. Fletcher and J. C. Steinberg. Articulation Testing Methods. *Bell System Technical Journal*, October:806–854, 1929. doi: 10.1002/j.1538-7305.1929.tb01246.x.

[107] N. R. French and J. C. Steinberg. Factors Governing the Intelligibility of Speech Sounds. *Journal of the Acoustical Society of America*, 19(1):90–119, 1947. doi: 10.1121/1.1916407.

[108] American National Standards Institute. ANSI S3.5-1969 American National Standard Methods for the Calculation of the Articulation Index, 1969.

[109] ASTM International. E1130-16 Objective Measurement of Speech Privacy in Open Plan Spaces Using Articulation Index, 2016.

[110] ANSI. ANSI/ASA S12.70 American National Standard Criteria for Evaluating Speech Privacy in Healthcare Facilities, 2016.

[111] International Organisation for Standardisation. ISO 9921:2003(en) Ergonomics - Assessment of Speech Communication, 2003.

[112] Arthur S. House, Carl E. Williams, Michael H. L. Hecker, and K. D. Kryter. Articulation-Testing Methods: Consonantal Differentiation with a Closed-Response Set. *The Journal of the Acoustical Society of America*, 37(1):158–166, 1965. doi: 10.1121/1.1909295.

[113] A. Boothroyd. Developments in Speech Audiometry. *International Journal of Audiology*, 7(3):368, 1968. doi: 10.3109/05384916809074343.

[114] Richard H. Wilson and Wendy B. Cates. A comparison of two word-recognition tasks in multitalker babble: Speech recognition in noise test (SPRINT) and words-in-noise test (WIN). *Journal of the American Academy of Audiology*, 19(7):548–556, 2008. doi: 10.3766/jaaa.19.7.4.

[115] Birger Kollmeier and Matthias Wesselkamp. Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *The Journal of the Acoustical Society of America*, 102(4):2412–2421, 1997. doi: 10.1121/1.419624.

[116] John Bench, Åse Kowal, and John Bamford. The BKB (Bamford-Kowal-Bench) Sentence Lists for Partially-Hearing Children. *British Journal of Audiology*, 13(3):108–112, 1979. doi: 10.3109/03005367909078884.

[117] Robyn M. Cox, Genevieve C. Alexander, and Christine Gilmore. Development of the Connected Speech Test. *Ear and Hearing*, 8(5):191–126, 1987. doi: 10.1097/00003446-198710001-00010.

[118] Benjamin W. Y. Hornsby. The Speech Intelligibility Index : What is it and what's it good for? *The Hearing Journal*, 57(10):10–17, 2004.

[119] Susan Scollie. SII Predictions of Aided Speech Recognition. *The Hearing Journal*, 57(9), 2004. doi: 10.1097/01.HJ.0000292838.52117.b1.

[120] Alexandra Macpherson and Michael A. Akeroyd. Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey. *Trends in Hearing*, 18:10–16, 2014. doi: 10.1177/2331216514537722.

[121] Jesper Jensen and Cees H. Taal. An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 24(11):2009–2022, 2016. doi: 10.1109/TASLP.2016.2585878.

[122] International Electrotechnical Commission. IEC 60268-16:2011 Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index, 2011.

[123] H.J.M. Steeneken and T. Houtgast. A physical method for measuring speech transmission quality. *The Journal of the Acoustical Society of America*, 67(1):318–326, 1980. doi: 10.1121/1.384464.

[124] American National Standards Institute. ANSI/ASA S3.5-1997 (R2017) Methods for Calculation of the Speech Intelligibility Index, 1997.

[125] Chaslav V. Pavlovic. The speech intelligibility index standard and its relationship to the articulation index, and the speech transmission index. *Journal of the Acoustical Society of America*, 119(5):3326, 2006. doi: 10.1121/1.4786372.

[126] V. Hongisto. A model predicting the effect of speech of varying intelligibility on work performance. *Indoor Air*, 15(6):458–468, 2005. doi: 10.1111/j.1600-0668.2005.00391.x.

[127] P. Virjonen, J. Keränen, R. Helenius, J. Hakala, and O. V. Hongisto. Speech privacy between neighboring workstations in an open office - A laboratory study. *Acta Acustica united with Acustica*, 93(5):771–782, 2007.

[128] International Organisation for Standardisation. ISO 7240-24:2016 Fire detection and fire alarm systems - Part 24: Fire alarm loudspeakers, 2016.

[129] Robert W. Peters, Brian C. J. Moore, and Thomas Baer. Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *The Journal of the Acoustical Society of America*, 103(1):577–587, 1998. doi: 10.1121/1.421128.

[130] Martin Cooke. A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119(3):1562–1573, 2006. doi: 10.1121/1.2166600.

[131] Yan Tang. *Speech intelligibility enhancement and glimpse-based intelligibility models for known noise conditions*. Ph.D. Thesis, Universidad del País Vasco, 2014.

[132] Yan Tang, Martin Cooke, Bruno M. Fazenda, and Trevor J. Cox. A metric for predicting binaural speech intelligibility in stationary noise and competing speech maskers. *The Journal of the Acoustical Society of America*, 140(3):1858–1870, 2016. doi: 10.1121/1.4962484.

[133] Yan Tang, Richard J. Hughes, Bruno M. Fazenda, and Trevor J. Cox. Evaluating a distortion-weighted glimpsing metric for predicting binaural speech intelligibility in rooms. *Speech Communication*, 82:26–37, 2016. doi: 10.1016/j.specom.2016.04.003.

[134] Jesper Jensen. estoi.m, 2016. URL `http://kom.aau.dk/~jje/code/estoi.m`. Accessed 31/07/2020.

[135] Jesper Jensen and Cees H. Taal. Speech Intelligibility Prediction Based on Mutual Information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(2): 430–440, 2014. doi: 10.1109/TASLP.2013.2295914.

[136] C. S. Watson, W. J. Kelly, and H. W. Wroton. Factors in the discrimination of tonal patterns. II. Selective attention and learning under various levels of stimulus uncertainty. *The Journal of the Acoustical Society of America*, 60(5):1176–1186, 1976. doi: 10.1121/1.381220.

[137] Xihong Wu and Jing Chen. A computational model for assessment of speech intelligibility in informational masking. *Frontiers of Electrical and Electronic Engineering in China*, 7 (1):107–115, 2012. doi: 10.1007/s11460-012-0189-8.

[138] Lennart Magnusson. Speech intelligibility index transfer functions and speech spectra for two Swedish speech recognition tests. *Scandinavian Audiology*, 25(1):59–67, 1996. doi: 10.3109/01050399609047557.

[139] Ryan W. McCreery. *Audibility as a Predictor of Speech Recognition and Listening Effort.* Ph.D. Thesis, University of Nebraska-Lincoln, 2011. URL `https://digitalcommons.unl.edu/cehsdiss/104`. Accessed 31/07/2020.

[140] Karl D. Kryter. Validation of the Articulation Index. *The Journal of the Acoustical Society of America*, 34(11):1698–1702, 1962. doi: 10.1121/1.1909096.

[141] Robert W. Young. Re-Vision of the Speech-Privacy Calculation. *Journal of the Acoustical Society of America*, 38(4):524–530, 1965. doi: 10.1121/1.1909735.

[142] J. S. Bradley. The acoustical design of conventional open plan offices. *Canadian Acoustics - Acoustique Canadienne*, 31(2):23–31, 2003. URL `https://jcaa.caa-aca.ca/index.php/jcaa/article/view/1523`. Accessed 31/07/2020.

[143] Bradford N. Gover and John S. Bradley. Measures for assessing architectural speech security (privacy) of closed offices and meeting rooms. *The Journal of the Acoustical Society of America*, 116(6):3480–3490, 2004. doi: 10.1121/1.1810300.

[144] ASTM International. E2638-10 (2017) Objective Measurement of the Speech Privacy Provided by a Closed Room, 2017.

[145] Markus Müller-Trapet and Bradford N Gover. Relationship between the privacy index and the speech privacy class. *Journal Of The Acoustical Society Of America*, 145(5):EL435–441, 2019. doi: 10.1121/1.5109049.

[146] Takahiro Tamesue, Shizuma Yamaguchi, and Tetsuro Saeki. Study on achieving speech privacy using masking noise. *Journal of Sound and Vibration*, 297(3-5):1088–1096, 2006. doi: 10.1016/j.jsv.2006.05.012.

[147] H.L.F. Helmholtz. *On the sensations of tone.* Longmans, Green and Co, 1895. doi: 10.1016/0016-0032(54)90054-X.

[148] Ulrich Widmann. *Ein Modell der Psychoakustischen Lästigkeit von Schallen und seine Anwendung in der Praxis der Lärmbeurteilung*. Ph.D. Thesis, TU München, 1992.

[149] H. Fletcher and W.A. Munson. Loudness, its Definition, Measurement and Calculation. *Journal of the Audio Engineering Society*, 5(1924):82–108, 1933. doi: 10.1121/1.1915637.

[150] International Organization for Standardization. ISO 226:2003 Acoustics – Normal equal-loudness-level contours, 2003.

[151] Brian C. J. Moore. Development and Current Status of the "Cambridge" Loudness Models. *Trends in Hearing*, 18:1–29, 2014. doi: 10.1177/2331216514550620.

[152] Brian R. Glasberg and Brian C. J. Moore. A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society*, 50(5):331–342, 2002. URL `http://www.aes.org/e-lib/browse.cfm?elib=11081`. Accessed 31/07/2020.

[153] Dominic Ward, Cham Athwal, and Munevver Kokuer. An efficient time-varying loudness model. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 13–16, 2013. doi: 10.1109/WASPAA.2013.6701884.

[154] Esben Skovenborg and Søren H. Nielsen. Evaluation of Different Loudness Models with Music and Speech Material. In *Proc. 117th Audio Engineering Society Convention*, San Francisco, 2004.

[155] British Standards Institute. BS ISO 532-1 : 2017 Acoustics — Methods for calculating loudness. Part 1: Zwicker Method, 2017.

[156] Ulrich Widmann. A Psychoacoustic Annoyance Concept for Application in Sound Quality. In *Proc. Noise-Con 1997*, 1997.

[157] Josef Schlittenlacher, Takeo Hashimoto, Sonoko Kuwano, and Seiichiro Namba. Overall loudness of short time-varying sounds. In *Proc. 43rd International Congress on Noise Control Engineering*, Melbourne, 2014.

[158] André Fiebig and Roland Sottek. Contribution of peak events to overall loudness. *Acta Acustica united with Acustica*, 101(6):1116–1129, 2015. doi: 10.3813/AAA.918905.

[159] Dominic Ward. *Applications of Loudness Models in Audio Engineering*. Ph.D. Thesis, Birmingham City University, 2017. URL `http://www.open-access.bcu.ac.uk/7228/1/PhDThesis.pdf`. Accessed 31/07/2020.

[160] Genesis Acoustics. Loudness Toolbox, 2020. URL `https://web.archive.org/web/20190509124114/http://genesis-acoustics.com/en/pages/post/sonie.php`. Accessed 31/07/2020.

[161] Hugo Fastl. Temporal Masking Effects: II. Critical Band Noise Masker. *Acta Acustica united with Acustica*, 36(5):317–331, 1976.

[162] Harvey Fletcher. Auditory Patterns. *Reviews of Modern Physics*, 12:47–66, 1940.

[163] E. Zwicker. Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2):248–248, 1961. doi: 10.1121/1.1908630.

[164] Daniel Wallace. Code for Psychoacoustic Metrics used in "Practical Audio System Design for Private Speech Reproduction", 2020. doi: 10.5258/SOTON/D1486.

[165] Sonoko Kuwano and Seiichiro Namba. Sharpness evaluation of temporally varying sounds. In *Proc. 45th International Congress and Exposition on Noise Control Engineering*, Hamburg, 2016.

[166] Institute of Electrical and Electronics Engineers. IEEE 269, 2020. URL `standards.ieee.org/downloads`. Accessed 28/07/2020.

[167] W. Aures. Ein Berechnungsverfahren der Rauhigkeit. *Acta Acustica united with Acustica*, 58(5):268–281, 1985.

[168] P Daniel and R. Weber. Psychoacoustical Roughness: Implementation of an Optimized Model. *Acta Acustica united with Acustica*, 83(1):113–123, 1997.

[169] Ronnie P. N. Duisters. *The modelling of auditory roughness for signals with temporally asymmetric envelopes*. Ph.D. Thesis, Technische Universiteit Eindhoven, 2005.

[170] Densil Cabrera, Sam Ferguson, and Emery Schubert. PsySound3: software for acoustical and psychoacoustical analysis of sound recordings. In *Proc. 13th International Conference on Auditory Display*, Montreal, 2007. URL `www.psysound.org`. Accessed 31/07/2020.

[171] Wouter Dreschler, Hans Verschuure, C. Ludvigsen, and S. Westermann. ICRA noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. *Audiology*, 40(3):148–157, 2001. doi: 10.3109/00206090109073110.

[172] Sabine Hochmuth, Birger Kollmeier, Thomas Brand, and Tim Jürgens. Influence of noise type on speech reception thresholds across four languages measured with matrix sentence tests. *International Journal of Audiology*, 54:62–70, 2015. doi: 10.3109/14992027.2015.1046502.

[173] Sarah McGuire and Patricia Davies. An Overview of Methods To Quantify Annoyance Due To Noise With Application To Tire-Road Noise. Technical report, Ray W. Herrick Laboratories, 2008. URL `http://www.igga.net/File/PurdueOverviewofMethodstoQuantifyAnnoyance.pdf`. Accessed 28/07/2020.

[174] European Parliament and Council of the European Union. Assessment and management of environmental noise (EU Directive). *Official Journal of the European Communities*, L189: 12–25, 2002. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32002L0049`. Accessed 28/07/2020.

[175] R.H. Lyon. *Designing for Product Sound Quality*. Marcel Dekker Inc., New York, 1st Edition, 2000.

[176] International Telecommunications Union. Recommendation P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications, 2004.

[177] International Telecommunications Union. Recommendation P.863: Perceptual Objective Listening Quality Assessment, 2011.

[178] International Telecommunications Union. Recommendation P.861: Objective quality measurement of telephone-band (300-3400 Hz) speech codecs, 1998.

[179] Matthew Jones and Stephen J. Elliott. Personal audio with multiple dark zones. *The Journal of the Acoustical Society of America*, 124(6):3497–3506, 2008. doi: 10.1121/1.2996325.

[180] C. House, S. Dennison, D. G. Morgan, N. Rushton, G. V. White, J. Cheer, and S. Elliott. Personal Spatial Audio in Cars Development of a loudspeaker array for multi-listener transaural reproduction in a vehicle. In *Proc. Institute of Acoustics*, volume 39. pt. 2, 2017.

[181] GRAS Sound and Vibration. Head and Torso Simulators, 2020. URL `www.gras.dk/products/head-torso-simulators-kemar`. Accessed 28/07/2020.

[182] Department for Education. Acoustic design of schools: performance standards. Technical Report February, Department for Education, 2015. doi: 10.1016/j.jns.2003.09.014.

[183] Heinrich Kuttruff. *Room Acoustics*. Elsevier, 3rd Edition, 1991.

[184] A. W. Bronkhorst and R. Plomp. Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *The Journal of the Acoustical Society of America*, 92(6):3132–3139, 1992. doi: 10.1121/1.404209.

[185] Angela Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Proc. 108th Audio Engineering Society Convention*, Paris, 2000. doi: 10.1109/ASPAA.1999.810884.

[186] Ole Kirkeby, Philip A. Nelson, Hareo Hamada, and Felipe Orduna-Bustamante. Fast deconvolution of multichannel systems using regularization. *IEEE Transactions on Speech and Audio Processing*, 6(2):189–194, 1998. doi: 10.1109/89.661479.

[187] Fiete Winter, Frank Schultz, Gergely Firtha, and Sascha Spors. A Geometric Model for Prediction of Spatial Aliasing in 2.5D Sound Field Synthesis. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 27(6):1031–1046, 2019. doi: 10.1109/TASLP.2019.2892895.

[188] Michael A. Akeroyd, John Chambers, David Bullock, Alan R. Palmer, A. Quentin Summerfield, Philip A. Nelson, and Stuart Gatehouse. The binaural performance of a cross-talk cancellation system with matched or mismatched setup and playback acoustics. *The Journal of the Acoustical Society of America*, 121(2):1056–1069, 2007. doi: 10.1121/1.2404625.

[189] Ferdinando Olivieri, Filippo Maria Fazi, Philip A. Nelson, Mincheol Shin, Simone Fontana, and Lang Yue. Theoretical and experimental comparative analysis of beamforming methods for loudspeaker arrays under given performance constraints. *Journal of Sound and Vibration*, 373:302–324, 2016. doi: 10.1016/j.jsv.2016.03.005.

[190] Khan Baykaner, Philip Coleman, Russell Mason, Philip J.B. Jackson, Jon Francombe, Marek Olik, and Søren Bech. The relationship between target quality and interference in sound zones. *Journal of the Audio Engineering Society*, 63(1-2):78–89, 2015. doi: 10.17743/jaes.2015.0007.

[191] Reinier Plomp. Auditory handicap of hearing impairment and the limited benefit of hearing aids. *Journal of the Acoustical Society of America*, 63(2):533–549, 1978. doi: 10.1121/1.381753.

[192] British Society of Audiology. Practice Guidance: Assessment of speech understanding in noise in adults with hearing difficulties, 2019. URL `https://www.thebsa.org.uk/wp-content/uploads/2019/04/OD104-80-BSA-Practice-Guidance-Speech-in-Noise-FINAL.Feb-2019.pdf`. Accessed 28/07/2020.

[193] Robert S. Bolia, W. Todd Nelson, Mark A. Ericson, and Brian D. Simpson. A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, 107(2):1065–1066, 2000. doi: 10.1121/1.428288.

[194] John R. Foster and Mark P. Haggard. The Four Alternative Auditory Feature test (FAAF)-linguistic and psychometric properties of the material with normative data in noise. *British Journal of Audiology*, 21(3):165–174, 1987. doi: 10.3109/03005368709076402.

[195] J. Ousey, S. Sheppard, T. Twomey, and A. R. Palmer. The ihr-mccormick automated toy discrimination test—description and initial evaluation. *British Journal of Audiology*, 23 (3):245–249, 1989. doi: 10.3109/03005368909076506.

[196] Mark E. Lutman, Stuart J. Hall, and Sheetal Athalye. Development of a telephone hearing test. In *Proc. Institute of Acoustics*, volume 28, Southampton, 2006. URL `http://resource.isvr.soton.ac.uk/staff/pubs/PubPDFs/Pub8744.pdf`. Accessed 31/07/2020.

[197] Michael Nilsson, Sigfrid D. Soli, and Jean A. Sullivan. Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*, 95(2):1085–1099, 1994. doi: 10.1121/1.408469.

[198] Brian Taylor. Speech-in-noise tests: How and why to include them in your basic test battery. *The Hearing Journal*, 56(1):40–43, 2003.

[199] Bjorn Hagerman. Efficiency of speech audiometry and other tests. *British Journal of Audiology*, 27(6):423–425, 1993. doi: 10.3109/03005369309076719.

[200] Thomas Brand and Birger Kollmeier. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *The Journal of the Acoustical Society of America*, 111(6):2801–2810, 2002. doi: 10.1121/1.1479152.

[201] D. N. Kalikow, K. N. Stevens, and L. L. Elliott. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, 61(5):1337–1351, 1977. doi: 10.1121/1.381436.

[202] Anne Anastasi. Validity: Basic Concepts. In *Principles of Psychological Testing*, page 139. Macmillan, 4th Edition, 1976.

[203] Nancy Aarts, Kathy R Duncan, and Nancy L Aarts. A comparison of the HINT and Quick SIN tests. *Journal of Speech-Language Pathology and Audiology*, June, 2006.

[204] Richard L. Freyman, Uma Balakrishnan, and Karen S. Helfer. Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America*, 115(5):2246–2256, 2004. doi: 10.1121/1.1689343.

[205] Hannah Domenica Semeraro. *Developing a measure of auditory fitness for duty for military personnel*. Ph.D. Thesis, University of Southampton, 2015. URL `https://eprints.soton.ac.uk/388043/1/HDSemeraro_PhD_ISVR_FEE_080216.pdf`. Accessed 31/07/2020.

[206] Birger Kollmeier, Anna Warzybok, Sabine Hochmuth, Melanie A. Zokoll, Verena Uslar, Thomas Brand, and Kirsten C. Wagener. The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International Journal of Audiology*, 54:3–16, 2015. doi: 10.3109/14992027.2015.1020971.

[207] K. Wagener, V. Kühnel, and B. Kollmeier. Development and Evaluation of a German Sentence Test. *Zeitschrift für Audiologie*, 31(1):1–32, 1999.

[208] Kirsten Wagener, Jane Lignel Josvassen, and Regitze Ardenkjær. Design, optimization and evaluation of a Danish sentence test in noise. *International Journal of Audiology*, 42 (1):10–17, 2003. doi: 10.3109/14992020309056080.

[209] Dale Robinson Hewitt. *Evaluation of an English Speech-in-Noise Audiometry Test*. MSc Thesis, University of Southampton, 2008.

[210] HearCom. Matrix sentence test, 2020. URL `http://hearcom.eu/prof/DiagnosingHearingLoss/AuditoryProfile/SpatialHearing.html`. Accessed 28/07/2020.

[211] Søren Bech and Nick Zacharov. *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley & Sons, 2006.

[212] David Oliva, Valtteri Hongisto, and Annu Haapakangas. Annoyance of low-level tonal sounds – Factors affecting the penalty. *Building and Environment*, 123:404–414, 2017. doi: 10.1016/j.buildenv.2017.07.017.

[213] Valtteri Hongisto, Johanna Varjo, David Oliva, Annu Haapakangas, and Evan Benway. Perception of water-based masking sounds-long-term experiment in an open-plan office. *Frontiers in Psychology*, 8(JUL):1–14, 2017. doi: 10.3389/fpsyg.2017.01177.

[214] Jennifer A. Veitch, John S. Bradley, Louise M. Legault, Scott Norcross, and Jana M. Svec. Masking Speech in Open-Plan Offices with Simulation Ventilation Noise: Noise-Level and Spectral Composition Effects on Acoustic Satisfaction. Technical report, Institute for Research in Construction, 2002. doi: 10.4224/20386334.

[215] Warren E. Blazier Jr. Revised noise criteria for application in the acoustical design and rating of hvac systems. *Noise Control Engineering*, 16(2):64–73, 1981. doi: doi:10.3397/1.2832172.

[216] Robert C. Chanaud. Progress in Sound Masking. *Acoustics Today*, 3(4):21–26, 2007. doi: 10.1121/1.2961158.

[217] Nathan Van Ness. 5 Reasons "Adaptive" Sound Masking Negatively Affects Sound Masking Performance and Occupant Comfort, 2019. URL `https://cambridgesound.com/5-reasons-adaptive-sound-masking-negatively-affects-sound-masking-performance-and-occupant-comfort/`. Accessed 28/07/2020.

[218] E. Colin Cherry. Some Experiments on the Recognition of Speech, with One and with Two Ears. *Journal Of The Acoustical Society Of America*, 25(5):975–979, 1953.

[219] International Telecommunications Union. ITU-T P.800 Methods for objective and subjective assessment of quality, 1996.

[220] International Telecommunication Union. ITU-R Recommendation BS.1284-1 General methods for the subjective assessment of sound quality, 2019.

[221] International Telecommunication Union. ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems, 2001.

[222] Maria Perez-Ortiz and Rafal K. Mantiuk. A practical guide and software for analysing pairwise comparison experiments, 2017. URL `http://arxiv.org/abs/1712.03686`. Accessed 31/07/2020.

[223] Natanya Ford, Francis Rumsey, and Bart De Bruyn. Graphical Elicitation Techniques for Subjective Assessment of the Spatial Attributes of Loudspeaker Reproduction – A Pilot Investigation. In *Proc. 110th Audio Engineering Society Convention*, Amsterdam, 2001. URL `http://www.aes.org/e-lib/browse.cfm?elib=9984`. Accessed 31/07/2020.

[224] Thibaud Leclère, David Théry, Mathieu Lavandier, and John F. Culling. Speech intelligibility for target and masker with different spectra. In *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing*, Cham, 2016. doi: 10.1007/978-3-319-25474-6_27.

[225] Terence Betlehem and Thushara D. Abhayapala. Theory and design of sound field reproduction in reverberant rooms. *The Journal of the Acoustical Society of America*, 117(4): 2100–2111, 2005. doi: 10.1121/1.1863032.

[226] Taffeta M. Elliott and Frédéric E. Theunissen. The modulation transfer function for speech intelligibility. *PLOS Computational Biology*, 5(3), 2009. doi: 10.1371/journal.pcbi.1000302.

[227] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979. doi: 10.1121/1.382599.

[228] Ji-Ho Chang, Jin-Young Park, and Yang-Hann Kim. Scattering effect on the sound focused personal audio system. *The Journal of the Acoustical Society of America*, 125(5):3060, 2009. doi: 10.1121/1.3101453.

[229] Marcos F. Simón Gálvez, Stephen J. Elliott, and Jordan Cheer. A superdirective array of phase shift sources. *The Journal of the Acoustical Society of America*, 132(2):746–756, 2012. doi: 10.1121/1.4733556.

[230] Eric A. Lehmann and Anders M. Johansson. Diffuse reverberation model for efficient image-source simulation of room impulse responses. *IEEE Transactions on Audio, Speech and Language Processing*, 18(6):1429–1439, 2010.

[231] Eric A. Lehmann. Fast image-source method: Matlab code, 2010. URL `http://www.eric-lehmann.com/fast_ISM_code/index.html`. Accessed 28/07/2020.

[232] Leo Beranek and Tim Mellow. Radiation and scattering of sound by the boundary value method. In *Acoustics - Sound Fields, Transducers and Vibration*, Chapter 12, pages 553–603. Academic Press, 3rd Edition, 2019. doi: 10.1016/C2017-0-01630-0.

[233] Wallace C. Sabine. Architectural acoustics. In *Proc. American Academy of Arts and Sciences*, volume 42. 1906. doi: 10.2307/20022177.

[234] Murray Hodgson. When is diffuse-field theory applicable? *Applied Acoustics*, 49(3):197–207, 1996. doi: 10.1016/S0003-682X(96)00010-2.

[235] M. Poletti, F. M. Fazi, and P. A. Nelson. Sound-field reproduction systems using fixed-directivity loudspeakers. *The Journal of the Acoustical Society of America*, 127(6):3590–3601, 2010. doi: 10.1121/1.3409486.

[236] Matti Karjalainen, Tuomas Paatero, John N. Mourjopoulos, and Panagiotis D. Hatziantoniou. About room response equalization and dereverberation. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 5:183–186, 2005. doi: 10.1109/ASPAA.2005.1540200.

[237] Panagiotis D. Hatziantoniou and John N. Mourjopoulos. Generalized Fractional-Octave Smoothing of Acoustic Responses. *Journal of the Audio Engineering Society*, 48(4):259–280, 2000. URL `http://www.aes.org/e-lib/browse.cfm?elib=12070`. Accessed 31/07/2020.

[238] Matti Karjalainen, Esa Piirilä, Antti Järvinen, and Jyri Huopaniemi. Loudspeaker response equalisation using warped digital filters. In *Proc. of NorSig-96*, 1996. URL `http://users.spa.aalto.fi/mak/PUB/Norsig96_EQ.pdf`. Accessed 28/07/2020.

[239] Bengt-Inge Dalenbäck. CATT, 2018. URL `www.catt.se`. Accessed 28/07/2020.

[240] Daniel Wallace. Dataset for: A Low-Cost Loudspeaker Array for Personal Audio with Enhanced Vertical Directivity, 2020. doi: 10.5258/SOTON/D1218.

[241] Adelbert W. Bronkhorst. The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple Talker Conditions. *Acta Acustica united with Acustica*, 86(1):117–128, 2000.

[242] J. Swaminathan, C. R. Mason, T. M. Streeter, V. Best, E. Roverud, and G. Kidd. Role of Binaural Temporal Fine Structure and Envelope Cues in Cocktail-Party Listening. *Journal of Neuroscience*, 36(31):8250–8257, 2016. doi: 10.1523/JNEUROSCI.4421-15.2016.

[243] BSI. BS 4142:2014+A1:2019 Methods for rating and assessing industrial and commercial sound, 2019.

[244] Michael Caley and Peter Georgiou. Application of Signal Detection Theory to Alarm Audibility in a Locomotive Cabin Environment. In *Proc. Acoustics*, 2004.

[245] Mark A. Abramson. *Pattern Search Filter Algorithms for Mixed Variable General Constrained Optimization Problems*. Ph.D. Thesis, Rice University, 2002.

[246] Adam Weisser, Jörg M. Buchholz, Chris Oreinos, Javier Badajoz-Davila, James Galloway, Timothy Beechey, and Gitte Keidser. The Ambisonic Recordings of Typical Environments (ARTE) database. *Acta Acustica united with Acustica*, 105(4):695–713, 2019. doi: 10.3813/AAA.919349.

[247] William A. Yost. Spatial release from masking based on binaural processing for up to six maskers. *The Journal of the Acoustical Society of America*, 141(3):2093–2106, 2017. doi: 10.1121/1.4978614.

[248] Gary L. Jones and Ruth Y. Litovsky. A cocktail party model of spatial release from masking by both noise and speech interferers. *The Journal of the Acoustical Society of America*, 130(3):1463–1474, 2011. doi: 10.1121/1.3613928.

[249] N. I. Durlach. Equalization and Cancellation Theory of Binaural Masking-Level Differences. *The Journal of the Acoustical Society of America*, 35(8):1206–1218, 1963. doi: 10.1121/1.1918675.

[250] John F. Culling, Monica L. Hawley, and Ruth Y. Litovsky. The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. *The Journal of the Acoustical Society of America*, 116(2):1057–1065, 2004. doi: 10.1121/1.1772396.

[251] Monica L. Hawley, Ruth Y. Litovsky, and John F. Culling. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, 115(2):833–843, 2004. doi: 10.1121/1.1639908.

[252] Jürgen Peissig and Birger Kollmeier. Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. *The Journal of the Acoustical Society of America*, 101(3):1660–1670, 1997. doi: 10.1121/1.418150.

[253] British Standards Institution. BS EN 61671-1:2013 Electroacoustics — Sound level meters, 2013.

[254] Bradford N. Gover, James G. Ryan, and Michael R. Stinson. Microphone array measurement system for analysis of directional and spatial variations of sound fields. *The Journal of the Acoustical Society of America*, 112(5):1980–1991, 2002. doi: 10.1121/1.1508782.

[255] Joost M. Festen and Reinier Plomp. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, 88(4):1725–1736, 1990. doi: 10.1121/1.400247.

[256] George A. Miller and J. C. R. Licklider. The Intelligibility of Interrupted Speech. *Journal of the Acoustical Society of America*, 22(2):167–173, 1950.

[257] Valtteri Hongisto and Jukka Ker. Simple model for the acoustical design of open-plan offices. *Acta Acustica united with Acustica*, 90:481–485, 2004.

[258] Soft dB Inc. Sound volume automatic adjustment method and system, February 2 2012. US Patent 8116461B2.

[259] Rein Pirn. Acoustical Variables in Open Planning. *The Journal of the Acoustical Society of America*, 49(5A):1339–1345, 1971. doi: 10.1121/1.1912506.

[260] D. J. Thompson and J. Dixon. Vehicle Noise. In Frank Fahy and John Walker, editors, *Advanced Applications in Acoustics, Noise and Vibration*, Chapter 6, page 258. CRC Press, 2018.

[261] Markus Christoph. Speed Dependent Equalising Control System, August 25th 2015. US Patent US9118290B2.

[262] Woomin Jung, Stephen J. Elliott, and Jordan Cheer. Identifying Interior Noise Sources in a Vehicle Cabin Using the Inverse Method. In *Proc. 23rd International Congress on Sound & Vibration*, Athens, 2016.

[263] Richard H. Lyon. Product sound quality: From perception to design. *Sound and Vibration*, 37(3):18–23, 2003. doi: 10.1121/1.4743110.

[264] Lynn Hansberry Mayo, Mary Florentine, and Søren Buus. Age of second-language acquisition and perception of speech in noise. *Journal of Speech, Language, and Hearing Research*, 40(3):686–693, 1997. doi: 10.1044/jslhr.4003.686.

[265] Lisa Kilman, Adriana Zekveld, Mathias Hällgren, and Jerker Rönnberg. The influence of non-native language proficiency on speech perception performance. *Frontiers in Psychology*, 5(651), 2014. doi: 10.3389/fpsyg.2014.00651.

[266] Kristin J. Van Engen. Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble. *Speech Communication*, 30(52):943–953, 2010. doi: 10.1038/mp.2011.182.doi.

[267] Ann R. Bradlow and David B. Pisoni. Recognition of spoken words by native and non-native listeners: Talker-, listener, and item-related factors. *Journal of the Acoustical Society of America*, 106(4):2074–2085, 1999. doi: 10.1038/jid.2014.371.

[268] Anna Warzybok, Thomas Brand, Kirsten C. Wagener, and Birger Kollmeier. How much does language proficiency by non-native listeners influence speech audiometric tests in noise? *International Journal of Audiology*, 54:88–99, 2015. doi: 10.3109/14992027.2015.1063715.

[269] Philip P. Morse and K. Uno Ingard. *Theoretical Acoustics*. McGraw-Hill, New York, 1st Edition, 1968.