Audio Zooming in a Drone Surveillance System for Police Evidence Gathering

Stephen Stroud

A thesis submitted in partial fulfilment of the requirements of
Liverpool John Moores University for the degree of
Doctor of Philosophy

December 2025

Declaration of Originality

No part of the work referred to in this thesis has been submitted to support an application for another degree or qualification for any university or institute of learning. The author alone has produced the research contained in this thesis.

Acknowledgements

I would like to thank my supervision team, particularly Dr Karl Jones, for his support and encouragement throughout this research project. I would also like to thank the other supervisors on the project, Dr Gerard Edwards, Colin Robinson, and Dr David Ellis.

I am also extremely grateful to my family, especially Sarah and Harvey, for all their patience and understanding during the previous years of research.

The Audio and Music Production Programme staff at the Byrom Street Campus of Liverpool John Moores University were a massive help during discussions regarding research and audio engineering in general. The Applied Forensic Technology Research Group members also need a special mention.

Abstract

This thesis documents the work on an innovative audio-zooming system intended to be attached to and then deployed on a drone, which is then able to be used for Police surveillance and evidence gathering. The central hypothesis is that a lightweight array of omni-directional microphones, when integrated with an advanced beamforming algorithm, noise reduction, and filtering techniques, can effectively detect and target noise, then separate and isolate the desired sound sources. This capability could significantly enhance the collection of auditory evidence from complex and noisy environments. This, in turn, could support the Police in law enforcement efforts to secure reliable convictions via forensic evidence-gathering methods.

The thesis begins with an introduction to the historical challenges associated with the Cocktail Party Problem, a term describing the human ability to detect and focus on a single sound source in a noisy environment with competing, interfering sources, and the longstanding attempts of the scientific community to replicate this biological phenomenon through engineering solutions. A literature review follows, exploring the multi-disciplinary aspects of sound localisation and sound separation problems, covering advancements in signal processing, machine learning, and acoustic engineering. Prior to the experimental approaches being covered, the acoustic-physics and mathematical foundations underpinning beamforming and noise-suppression theory are explored. Building on that groundwork, there is a focus on the algorithms used for directional sound capture, such as the Minimum Variance Distortionless Response (MVDR) beamformer, together with advanced audio techniques such as spectral filtering to improve legibility. The discussion proceeds to the equipment that make the experiments possible. MATLAB® [1]serves as a flexible environment for simulation, while a set of custom-built microphone arrays gathers the raw acoustic data. An examination of the experimental approach rounds things out by explaining, step by step, how the listening trials were designed and executed, including array configurations, test-room conditions, and the data-analysis pipeline that converts recordings into interpretable results. The results from the experiments are presented and analysed to evaluate the performance of the proposed system, including metrics on signal-to-noise ratio improvements, directional accuracy, and computational efficiency.

The conclusion summarises the findings, affirming the potential and limitations of the proposed system while suggesting pathways for future research, such as integrating real-time processing capabilities and testing in varied environmental conditions. This thesis contributes to the growing field of audio and video engineering by offering a novel framework for drone-mounted auditory surveillance, opening avenues for enhanced law enforcement and forensic practices.

[1] MATLAB is a registered trademark of The MathWorks, Inc.

Table of Contents

List of Figures

List of Tables

Glossary

Ambisonic	*A multi-channel audio format. Captures vertical and horizontal signals to create the impression of surround sound.*
Azimuth	*Refers to the angle of a received signal compared to the horizontal plane of a receiver.*
Beamforming	*Commonly known as spatial filtering. It is a signal-processing technique used in sensor arrays to measure signals coming from specific directions.*
Direction of arrival	*Describes the location of a sound source relative to the receiver position.*
Omni-Directional	*A 360-degree polar pattern for a microphone that can capture signals uniformly from all directions.*
Pink Noise	*Pink noise has a relatively even distribution of energy across frequencies. Its intensity decreases in a stair-step fashion, so each octave has the same energy.*
Super-Cardioid	*A more directional microphone polar pattern.*

Nomenclature

r	The Radial distance (Meters)
φ	The Azimuth angle (Radians)
θ	The Polar angle (Radians)
c	The Speed of sound (ms^{-1})
f	Frequency (Hertz)
λ	Wavelength (Metres)
t	Time (seconds)

List of Acronyms

ASR	*Automatic speech recognition*
AMNOR	*Adaptive Microphone-array for Noise Reduction*
AZ	*Audio Zooming*
BSS	*Blind Source Separation*
CASA	*Computational Auditory Scene Analysis*
CCTV	*Closed-Circuit Television*
CPP	*Cocktail Party Problem*
DAW	*Digital Audio Workstation*
DirAC	*Directional Audio Coding*
DOA	*Direction of Arrival*
DRR	*Direct to Reverberant Ratio*
ESPRIT	*Estimation of Signal Parameters via Rotational Invariance Technique*
GCC-PHAT	*Generalised Cross-Correlation Phase Transform*
HSS	*Harmonic Structure Stability*
ICA	*Independent Component Analysis*
ILD	*Interaural Level Difference*
ITD	*Interaural Time Difference*
KEMAR	*Knowles Electronic Manikin for Acoustical Research*
MATLAB™	*Matrix Laboratory*
MIMO	*Multiple Input Multiple Output*
MINT	*Multiple Input/Output Inverse Theorem*
MixIT	*Mixture Invariant Training*
MOM	*Mixtures of Mixtures*
MUSIC	*Multiple Signal Classification*

MVDR	*Minimum Variance Distortionless Response*
RMS	*Root Mean Square*
SIMO	*Single Input Multiple Output*
SI-SDR	*Scale-Invariant Signal-to-Distortion Ratio*
SNR	*Signal-to-Noise Ratio*
SOSM	*Second-Order Statistical Measures*
SPL	*Sound Pressure Level*
SRP-PHAT	*Steered-Response Power Phase Transform*
SSL	*Sound Source Localisation*
STFT	*Short-Time Fourier Transform*
STOI	*Short-Time Objective Intelligibility*
SUDO RM-RF	*Successive Downsampling and Resampling of Multi-Resolution Features*
SVD-PHAT	*Singular Value Decomposition - Phase Transform*
SVM	*Support Vector Machine*
TDOA	*Time Difference on Arrival*
THD	*Total harmonic distortion*

Universal Constants

Speed of Sound in Air	*Approximately 343 meters per second (ms^{-1}) at room temperature (20°C).*
Threshold of Human Hearing	*Approximately 20 microPascals (µPa) are typically equivalent to 0dB SPL (Sound Pressure Level).*
Human Hearing Range	*20Hz to 20kHz*

Chapter 1: Introduction

1.1 Motivation and Outline of the Research

The inspiration for this research was the tragic murder of a young girl called Ava White in Liverpool city centre on the 25th of November 2021. CCTV recorded the visual and the audio scenes leading up to the crime, however, because of the interfering environmental noise, the audio captured was unusable in court. If a technology was available to the Police which allowed a desired portion of the audio to be zoomed into, similar to the way a camera lens can zoom into a visual scene, then prosecutions such as this could be achieved in a more timely manner. More details on the murder of Ava White are available in Chapter 2.

Taken together, this case and similar incidents provide the primary motivation for this thesis: to investigate whether an audio-zooming system could improve the evidential value of audio in noisy urban surveillance footage. The scope of the work is restricted to microphone-array-based methods suitable for mounting on Police-style drones, with an emphasis on post-event forensic analysis rather than real-time deployment.

The research began with an investigation of Audio Zooming. Audio zooming refers to a concept where a remote user can select a listening position within an audio scene, examining the audio related to that specific location. The listening position would be represented as a point with known coordinates relative to a microphone array. These coordinates would be mapped to a direction of arrival (azimuth and elevation values) and could then be used to construct a steering vector. When appropriate weights are applied to the multichannel microphone signals, the output is thereby spatially linked to the selected location and approximates the signal that would be captured by a virtual microphone directed at that point. The concept of Audio Zooming is heavily influenced by the challenge of solving 'the cocktail party problem' (CPP).

The CPP is a long-established proposed challenge in acoustic signal processing that describes the human ability to focus on a single conversation or sound source within a noisy environment filled with overlapping and interfering sounds. This phenomenon was first formally noted by Colin Cherry in 1953, who highlighted the impressive ability of the human auditory system to locate, isolate and focus on a specific voice in a noisy room whilst simultaneously filtering out irrelevant or unwanted background sounds Cherry (1953).

The CPP symbolises a classic scientific challenge, which would demand complex solutions, including localisation and sound separation, which the human brain performs almost automatically. However, it still presents significant challenges for engineers attempting to replicate this set of processes using artificial systems and technologies. The task is difficult owing to the mixtures of various sound waves that simultaneously reach the ears (or microphones, in the case of replications) from different directions, blending into a composite signal. Identifying and extracting an individual source from this mixture requires advanced and highly complex computational approaches that mimic our biological systems.

Finding a technical solution to the CPP would greatly benefit speech recognition and generation, hearing aids, and the focus of this study, audio surveillance. The complexity of the problem is partly due to the issues of distinguishing between overlapping sound sources and adapting to real-world environmental challenges such as reverb and reflections, which quickly become unpredictable in a noisy scene. Interfering signals can come from multiple directions with various sound pressure levels and frequency makeup.

Over the last seventy years, multiple strategies have been proposed to address this complex problem, including manipulating microphone pickup patterns, beamforming techniques, adaptive noise cancellation, sound source separation, and independent component analysis (ICA). In recent years, progress in developing machine learning and neural networks has been promising to improve the results of sound source separation in noisy environments. Even so, creating a technological system that matches human performance remains a significant challenge for researchers due to the complex problems faced when dealing with the physics and unpredictable nature of real-world soundscapes.

The core research problem addressed in this thesis is how to recover intelligible target speech from distant, low-level sources embedded in complex, noisy and reverberant urban soundscapes, using a drone-mounted microphone array that must remain lightweight, low-power and robust enough for operational Police use.

The nature of CPP suggests that creating audio zooming techniques for effective sound source isolation in noisy settings would be sensible. Understanding and addressing the challenge is crucial for applications such as drone-mounted auditory surveillance systems, where precise and targeted sound capture can aid Police in law enforcement and evidence gathering.

While the idea of a machine solving the "cocktail party problem" was first introduced in the 1950s (Cherry, 1953), technology still struggles to replicate what the brain can do easily. In any typical audio environment, the targeted audio signals will be blended with interfering environmental sounds (Hawley, Litovsky and Culling, 2004), resulting in difficulty extracting the desired sounds from the surrounding noise. Therefore, it is necessary to develop a system that removes spurious audio, leaving only sounds of interest to the user (Huang, Benesty and Chen, 2006).

This work proposes the development of a process for receiving video footage (with its associated audio track). As the image is aimed in a particular direction or optically 'zoomed in', the focus of the audio will be correspondingly narrowed. The research focused on drones as the platform for a microphone array. The drone model is based upon those that the United Kingdom Police use for surveillance, such as the DJI M300 RTK[2], a rugged and heavy-duty model used by Merseyside Police and North Wales Police.

Figure 1.1. The DJI Matrice 300 RTK Drone. (DJI, 2024)

In the Ava White case, the CCTV surveillance audio was hampered by environmental sounds, which made focusing on the suspect's speech extremely difficult.

This work introduces a model that permits different array configurations, and the kind of scene simulations that allow the user to adjust to various sizes and dimensions. The model applies Minimum Variance Distortionless Response (MVDR) beamforming techniques, accounting for

[2] https://store.dji.com/uk/product/matrice-300-rtk-and-dji-care-plus?vid=111261

the reflections one would expect when an audio signal is projected in an urban environment. MVDR was chosen over simpler fixed beamformers such as delay-and-sum because it can better suppress interfering noise and control sidelobes while preserving the target direction, yet remains computationally feasible for a constrained drone platform; alternative beamformers and these trade-offs are discussed in Chapter 3. The model's algorithm also considers drone noise, reverberation and interference.

This research additionally required the design of an adaptive microphone array system for audio zooming and noise reduction (Kataoka and Ichinose, 1990), capable of deployment in conjunction with a video camera on a drone and the development of algorithms to perform audio zooming synchronised to the zooming of the video stream, whilst simultaneously removing the sound of the drone's motors and any wind noise captured.

In summary, the thesis proposes and evaluates a drone-mounted audio-zooming system based on lightweight microphone arrays, MVDR beamforming and subsequent noise-reduction stages, with the goal of improving the intelligibility and forensic usefulness of recorded surveillance audio. A detailed overview of the remaining chapters and their individual contributions is provided in Section 1.5.

1.2 Novel Contributions

This thesis presents a drone-suitable audio-zooming system. It covers hardware design and algorithms and validates them in simulation, a controlled booth, and outdoor trials. There are several areas of novelty documented.

A low-weight sixteen-microphone circular array was designed for carriage by a UK Police drone. The design, fabrication approach and wiring layout are documented, and acoustic and electrical characterisation was performed on the finished array.

A practical MVDR beamformer-based audio-zoom algorithm was implemented in MATLAB for the microphone arrays geometry, with a lightweight post-production spectral filter tuned to the interference seen in testing. The whole processing chain used in the results is specified and repeatable.

A multi-environmental evaluation was conducted using standard metrics. The same SNR and STOI procedures were applied in simulation, in a controlled booth, and in outdoor tests. Measurable improvements were observed in post-processing in each setting.

Lastly, robustness to array error and microphone loss was investigated in a MATLAB simulation and published separately (Stroud *et al.*, 2023). The results were that even with the loss of multiple sensors, the beamforming could still take place with minimal loss of focus.

These contributions move audio zooming from concept to a documented prototype with quantified behaviour in both controlled and outdoor scenes. Real-time operation and a complete evidential workflow are explicitly left to future work.

1.3 Aims and Objectives

1.3.1 Aim

The aim of this research was to develop a process where recorded audio from a microphone array can be sonically 'zoomed' into a specific geographic area. The work was carried out with a specific emphasis on Police drones as the platform for a microphone array.

1.3.2 Objectives

1. Development of a microphone array using low-weight microphones.

   Drones are increasingly being used for capturing audio and video due to their flexibility and range of coverage. However, the power consumption and weight of drone audio systems present a significant barrier to truly efficient drone-based capture. Their design mandates that they be lightweight, compact and low-power to maintain the drone's flight stability and extend its operational time. These are all necessary factors for truly superior aerial audio capture.

   In this thesis, this objective is addressed by designing and constructing prototype microphone arrays and validating their performance through controlled sound-booth experiments, simulations and outdoor field tests.

2. Development of a noise rejection system**.**

   A signal conditioning system makes the audio signals more useful by filtering out unwanted noise. Unwanted noise comes from the drone's motors, airflow, and other sources, which are common to all drones. Since the Police require clear audio to facilitate prosecutions, a system is needed to maintain the audio signal's integrity and filter out

anything that could likely compromise legibility. In this thesis, “noise rejection” refers specifically to spatial suppression of interferers using beamforming (placing nulls and controlling sidelobes in the directions of unwanted sources), whereas “noise reduction” refers to the subsequent spectral processing that attenuates residual broadband noise such as rotor and wind noise. Together, these stages form the proposed noise-reduction chain for drone-mounted audio capture. The proposed approach combines spatial filtering using MVDR beamforming with a subsequent spectral noise-suppression stage tailored to rotor, wind and environmental noise, implemented and evaluated in a MATLAB simulation framework and in real recordings.

3. Development of an audio zooming algorithm**:**

   Audio zooming sharpens the focus on a particular sound source, just as optical zoom lets a video camera get closer to a subject and maintain clarity. The audio zoom algorithm should aim the polarity of the microphone array toward the desired target while filtering out unwanted sounds. Recent research in UAV acoustics has shown that lightweight microphone arrays mounted on multi-rotor drones can be used with beamforming to localise and enhance ground-level sound sources in outdoor environments (Hoshiba *et al.*, 2017**;** Salvati *et al.*, 2020)**.** These systems typically employ delay-and-sum, MUSIC or MVDR-style beamformers to steer a main lobe towards the region of interest while placing nulls towards dominant interferers. In parallel, commercial drone-mounted acoustic imaging cameras now overlay beamformed sound-intensity maps onto high-resolution video, effectively providing an acoustic “zoom” for leak detection and industrial inspection (e.g. CRY2626G UAV acoustic imager) (Crysound, 2025).

   These developments demonstrate that drone-mounted microphone arrays are a practical platform for directional audio capture and motivate the audio-zooming approach taken in this thesis. In this work, audio zooming is realised by applying MVDR-based beamforming and post-filtering so that the output approximates a virtual directional microphone pointed at the region of interest. The algorithm is tested on simulated crime-scene simulations and field test recordings to quantify gains in SNR and intelligibility.

1.4 Reflection on Project Aim and Objectives

1.4.1 Aim

This research aimed to develop a process where recorded audio from a microphone array can be sonically 'zoomed' into a specific geographic area. The work was carried out with a particular emphasis on UK Police drones as the platform for a microphone array. That aim has been met: the prototype significantly raises the intelligibility of recorded speech in ideal conditions and continues to improve intelligibility levels even when drone noise, wind, and traffic are present.

1.4.2 Objectives

1.4.2.a Objective 1. Development of a microphone array using low-weight microphones

The first objective was to design a low-mass array suitable for integration with a UK Police drone. The objective has been achieved. The array met the mechanical and electrical targets and remained within the payload capacity of a DJI M300-class airframe. In controlled tests and simulations, the circular sixteen-element design maintained main-lobe focus and showed less than a 1dB deficit relative to an idealised octagon while simplifying fabrication. Fault-introduction runs demonstrated minimal degradation under sensor loss, which supports operational robustness.

The main trade-offs are weight, power and environmental coupling. Added mass shortens endurance and can move the centre of gravity away from the optimal point for stable hover. Wind loading and rotor wash increase low-frequency vibration and flow noise at the capsules. Cable routing and mechanical isolation need careful attention to limit microphonics. At the electronics level, capsule matching, bias stability and preamp equivalent input noise set the floor for beamformer performance in quiet scenes. In short, the array geometry is sound and field-tolerant, but long-term endurance and excessive low-frequency noise remain the governing constraints for a flight-worthy system.

1.4.2.b Objective 2. Development of a noise-rejection system

The second objective sought an adaptive noise-rejection system able to suppress rotor and ambient interference. The objective has been achieved with some constraints. Combining MVDR spatial filtering with a lightweight spectral-subtraction stage reduced rotor tones and broadband ambient noise and delivered measurable SNR and STOI gains in booth, simulation and field

conditions. The chain suppressed dominant low-frequency energy from approximately 150 to 600Hz and improved intelligibility when interference and early reflections were present. The design is simple, explainable and compatible with evidential logging.

Limitations reflect scene dynamics. Wind gusts, rapid attitude changes and non-stationary interferers reduce cross-channel coherence, which weakens spatial filtering and forces more aggressive spectral masks. Parameter choices, therefore, remain scene dependent. A calibrated noise reference helps, but impulsive sources and dense music require alternative priors or Wiener post-filters to avoid distortion. The system is effective and reproducible, but automatic tuning for fast-changing outdoor scenes would be covered with future work.

1.4.2.c Objective 3. Development of an audio-zooming algorithm

The third objective concerned an audio-zoom algorithm that could perform an acoustic zoom similar to an optical zoom. This objective has been achieved as a demonstrable prototype. Scene-steered beamforming produced a practical acoustic zoom that can be linked to digital visual framing. Grid steering and lobe shaping worked in simulation and outdoor tests, and performance remained acceptable with simulated element failure. This brings the cocktail-party problem into a drone-sized array domain and shows that a classical beamformer plus a modest post-filter can recover speech from cluttered scenes without heavy machine-learning infrastructure.

Constraints are primarily computational. The MATLAB prototype is not real-time. A deployable system must maintain beam coherence during yaw, pitch and roll, and handle time-varying Doppler and airframe vibration without retraining or operator retuning. Reverberant courtyards and hard façades also reduce spatial selectivity at low frequencies so that some scenes will require hybrid steering plus post-filtering. The algorithm achieves useful audio zoom in practice, but embedded implementation and motion-aware coherence control are needed for live operations.

Collectively, all planned objectives have been achieved in demonstrable prototype form and satisfy the aim of the project.

1.5 Summary

This chapter set out the “cocktail-party” challenge and showed why Police drones would benefit from an audio-zooming capability. By examining how the human auditory system can focus on one person's speech amid competing voices, the chapter highlights the benefit of giving machines a comparable skill.

The project aims to build an adaptive microphone-array platform that keeps multiple channels in step while the drone is in flight. Achieving clear recordings is difficult because rotor noise and background sounds mask weak speech cues.

To meet that challenge, the study combines MVDR beamforming with a noise-suppression stage and could link the audio focus to the drone's optical zoom in post production. The result is a surveillance tool that improves monitoring and supports forensic playback. The design is intended to remain effective when some microphones fail or when wind and traffic noise rise because the circular array provides overlapping coverage from multiple directions and the beamforming/noise-reduction stages are designed to tolerate missing channels and changing noise conditions; the robustness of this approach is examined in detail in the simulation and field experiments reported in later chapters.

Additionally, this research emphasises forensic audio analysis post-event. Critical audio evidence from crime scenes often suffers degradation due to interfering sounds, significantly hampering investigative clarity. The proposed framework enables clearer audio extraction and separation retrospectively from complex soundscapes, increasing the reliability and effectiveness of audio evidence in legal contexts.

Regarding the thesis structure, the remainder of this document is organised into eight chapters.

Chapter 2 recalls the investigation that set this research project in motion, focusing on the murder of Ava White, the available evidence and the legal context, and shows how an audio-zoom method could help detectives pick out and understand low-level speech masked by louder foreground sounds.

Chapter 3 presents a literature review of audio zooming, the Cocktail Party Problem, sound source localisation, source separation and noise-reduction methods, with particular emphasis on microphone arrays, beamforming and related forensic audio techniques, and identifies gaps that motivate the present work.

Chapter 4 describes the initial experimental approach in a controlled sound-booth environment, detailing the microphone and loudspeaker configurations, custom hardware builds, test procedures and early proof-of-concept experiments used to explore the feasibility of audio zooming.

Chapter 5 introduces the simulation approach, revisiting key acoustic and spatial-filtering theory before instantiating these equations in a MATLAB-based model that encodes the drone payload constraints, microphone geometry and rotor-noise characteristics, and applies this model to both a reflection-controlled booth and a reconstruction of the Ava White crime scene.

Chapter 6 extends this simulation framework to Liverpool John Moores University's Exemplar Houses, describing the design of the circular microphone array and loudspeaker rig, the physics-based simulations of the site and the subsequent self-contained field tests that validate the algorithm in a realistic outdoor setting.

Chapter 7 presents the experimental results from the sound-booth trials, simulations and Exemplar House tests, including signal-to-noise ratio improvements, intelligibility metrics and polar response characteristics for the proposed audio-zoom system. There is also a discussion and interpretation of these results in the context of the existing literature, assessing the system's strengths and limitations and its suitability for forensic deployment.

Chapter 8 draws together the main conclusions of the research, summarises the novel contributions, reflects on whether the original aim and objectives have been met and outlines potential directions for future work.

Chapter 2: The Murder of Ava White

2.1 Overview

In order to better understand the context of the project, it would be useful to be aware of the details of the criminal case that inspired the work. On 25th November 2021, the fatal stabbing of twelve-year-old Ava White (see Figure 2.1) during Liverpool's Christmas lights switch-on exposed a critical gap in current urban surveillance. While multiple CCTV cameras captured video, the accompanying audio was seemingly drowned out by crowd noise, traffic and a nearby busker. This chapter reconstructs the incident from the confrontation on School Lane, to the time of the teenage perpetrator's conviction in July 2022, and catalogues the evidence that ultimately shaped this research. By examining where microphone technology proved insufficient and how that shortfall influenced the investigation, this Chapter sets the tone for the experimental, simulation and field studies presented in the rest of the thesis.

Figure 2.1. Portrait of Ava White released after the Sentencing. (Merseyside Police, 2022).

2.2 Background and Timeline

On the evening of Thursday 25 November 2021, Liverpool city centre hosted a Christmas lights switch-on event on Church Street. At around 8:40 pm, twelve-year-old Ava White confronted a group of four local teenage boys on School Lane after seeing that they had filmed her on Snapchat (BBC, 2022). During the brief altercation, one boy stabbed Ava once in the neck with a 7.5 cm knife before he and the group ran away. Ava staggered to nearby Church Alley, where she collapsed; she was taken to Alder Hey Children's Hospital and died later that night at approximately 10:16 pm.

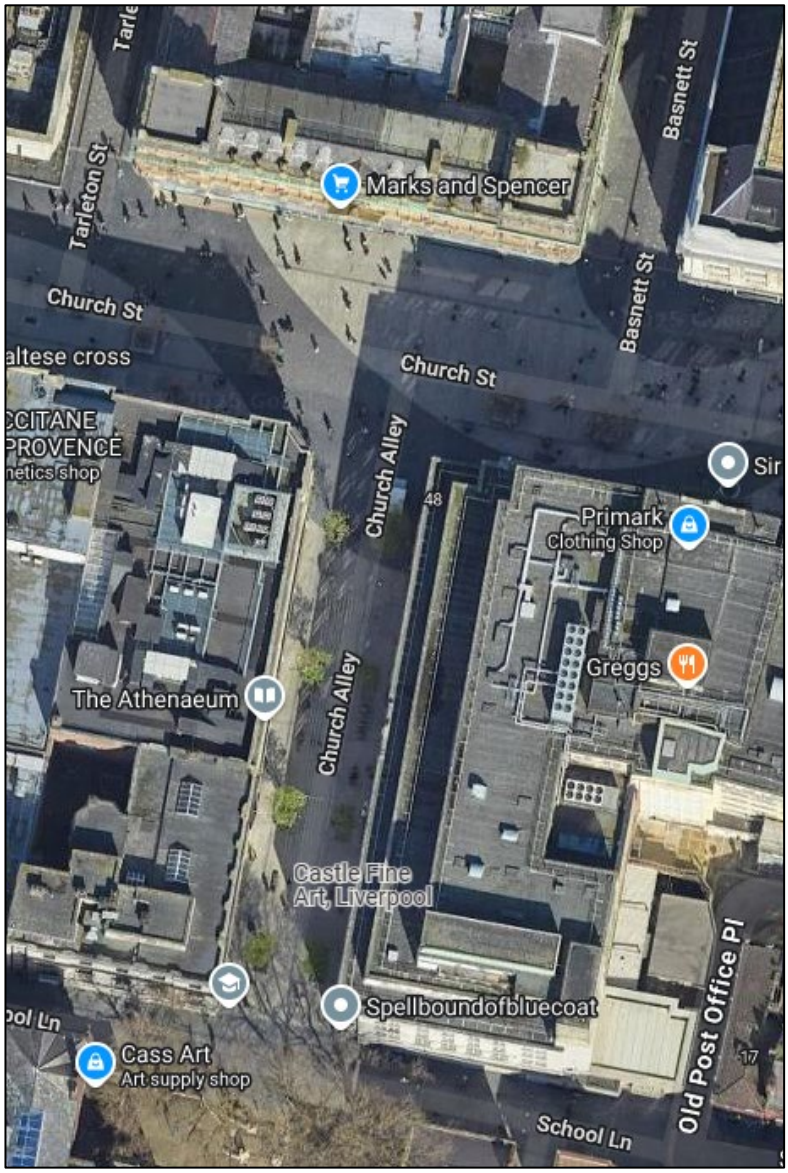

Figure 2.2. Google Maps Satellite view of the murder scene (53.40452, -2.98377) (Google Maps, 2025a).

The suspects, who cannot be named for legal reasons, were identified using witness accounts and CCTV footage from the surrounding streets. All four were arrested on 28 November, but three were later released without charge. One boy was charged with murder and possession of a knife and remanded in custody. A plea hearing in February 2022 saw him admit carrying the knife but deny murder. A jury trial at Liverpool Crown Court between 9 and 24 May 2022 considered CCTV, video evidence and eyewitness testimony. On 24 May the jury returned a unanimous guilty verdict, and on 11 July 2022 the judge imposed a life sentence with a minimum term of 13 years (The Guardian, 2022).

2.3 Available Evidence

The case relied almost entirely on CCTV, video and eyewitness accounts.

Since the incident happened during a Christmas Lights switch-on, there was a large amount of environmental noise from the crowd, traffic and a busker. The Liverpool city-centre cameras belonging to the council and local retailers did not have specialised audio capture arrays and instead relied on standard consumer microphones (Merseyside Police, 2022).

Witnesses claimed that it was a Snapchat video of Ava, filmed on the boy's phone, that triggered the initial dispute. While some video evidence was filmed at the scene, no useful or intelligible audio was recorded beyond crowd noise. The jury would later rely on visual analysis and eyewitness testimony (Judiciary of England and Wales, 2022).

2.4 The Importance of Audio

Typical urban Closed-Circuit Television (CCTV) microphone specifications include a single electret-condenser omnidirectional Ø 6 mm capsule. It is cheap, robust, and picks up 360 ° audio, so the camera does not have to be aimed at a target (Nodac, 2019).

These microphones typically have a frequency range of 40Hz –16kHz. The Verifact-A model can pick up 40Hz-15kHz (Louroe, 2010), while the Axis P-series built-in microphone captures 50Hz-16kHz. This covers the human speech band with enough headroom for the 16kHz audio that tends to be used by IP codecs.

These CCTV cameras typically use Broadband Automatic gain control and fixed 12dB/oct pre-emphasis. For example, the Nodac OCB-CM40 lists *"Built-in AGC circuit to adapt to environment"* (Nodac, 2019). Regarding sensitivity, the OCB-CM40 possesses 58dB re 1 V/Pa

(Nodac, 2019). The high sensitivity lets a single capsule reach 20-30 m outdoors without interfering barriers.

In terms of self-noise, which impacts the signal-to-noise ratio (SNR), the model 71423 add-on microphone has an SNR ≈ 60dB for a 40dB-SPL source, which results in a 34dB-A equivalent self-noise (CCTV Camera World, 2025). High-quality studio microphone manufacturer Neumann's microphone-noise guide classifies 20–23dB-A as a high self-noise figure for a studio microphone and suggests that levels of 24dB-A and above are unworthy of studio use (Neumann, 2024).

Therefore, the CCTV built-in microphone would be considered to have a high self-noise by studio standards. The noise floor is acceptable for picking up loud voices, but it would make the quiet ambience sound noisy.

Regarding recording audio, CCTV cameras typically encode incoming analogue audio at 16kHz / 16-bit PCM, a lower resolution compared to the modern, standard studio recording convention of 48kHz / 24-bit PCM. However, the CCTV audio sample rate is adequate for capturing environmental noise and works with the built-in microphone's 16kHz acoustic limit without wasting bandwidth. As per the Nyquist–Shannon sampling theorem, 16kHz sampling can only capture up to 8kHz F_{max} accurately, therefore, unless there is appropriate filtering, 16kHz could produce aliasing (Shannon, 1949).

In the Ava White case, the collected CCTV recordings, if the microphones were live, would contain heavy crowd and traffic noise, which would interfere with any desirable verbal evidence as it could mask the 1–4kHz speech band in the audio footage.

Had a specialised microphone array been mounted above School Lane, beamforming with side-lobe suppression, followed by a filter focused on the target area, it would have been possible to improve the signal-to-noise ratio enough to capture any incriminating speech. The resulting gain in intelligibility could have expedited the investigation and reduced any ambiguities.

2.5 Legal Considerations

The admissibility of enhanced audio in legal cases is something that should be addressed. Part 19 of the Criminal Procedure Rules states that expert evidence, such as digitally enhanced audio, provided that it is relevant, scientifically reliable and from an unbroken evidential chain, is admissible (Crown Prosecution Service, 2022).

The party tendering the material must supply the original file, the enhancement log and a report that explains the technique in terms that a lay Jury can follow.

Regarding the issue of proportionality versus privacy, Paragraph 3.2.2 of the Surveillance Camera Code of Practice puts a strong presumption against routine audio capture in public spaces. Only a "pressing need" can override that presumption (Home Office, 2021). Liverpool City Council's CCTV policy adopts the same stance and specifies that its city-centre cameras are configured for video-only operation (Liverpool City Council, 2022).

The procedures for dealing with what might legally be considered AI-processed evidence require careful consideration. When beamforming, filtering or enhancement is applied to audio, every transformation step must be recorded, time-stamped and signed off. Software versions, parameter sets and cryptographic hash values (fixed-length digital fingerprints such as SHA-256) for input and output files should be preserved so that an independent expert can reproduce the result (Crown Prosecution Service, 2022).

The legal impact of audio evidence on the Ava White case is difficult to judge. The attack meets the serious incident threshold, yet no public source shows that intelligible speech was recorded. Given the policy settings above, any audio recording that does exist is likely to contain only diffuse ambient noise, crowd, traffic and street music rather than point-source speech.

2.6 Summary

The Ava White case raises questions about how current urban surveillance systems can fall short when investigators need intelligible audio. Liverpool's city centre cameras often record video only or rely on single, low-sensitivity microphones that struggle against interfering crowd, traffic and music noise. Technical analysis of typical CCTV capsules and Liverpool City Council's privacy policy confirms that any captured audio would likely be limited to general ambience rather than focused speech. As a result, the court could not draw on audio evidence, instead relying on visual footage and eyewitness testimony.

In comparable future incidents, a police-operated drone equipped with a beamforming microphone array could be deployed above the scene to capture directional, noise-suppressed audio of specific individuals or interactions that are not intelligible on fixed CCTV microphones. The absence of any usable audio evidence in this case therefore highlights the need for a system that (i) provides spatial selectivity towards defined regions of interest, (ii) improves the signal-

to-noise ratio of weak speech masked by diffuse noise and dominant interferers, (iii) tolerates reverberant urban sound fields and changing noise conditions, and (iv) produces outputs that can be documented and reproduced to forensic standards. These requirements define the problem that the proposed audio-zooming system must solve and motivate the beamforming and noise-reduction approaches developed in Chapters 4 to 6.

The next chapter, the Literature Review, surveys existing work on microphone arrays, beamforming and audio zooming and positions the present research within that context.

Chapter 3: Literature Review

3.1 Overview

This chapter explores the academic research behind audio zooming, the Cocktail Party Problem (CPP), sound source localisation (SSL), and source separation. Additionally included is a review of the noise reduction problem, which reveals a genuine concern for relevant real-world scenarios, such as drone-based sound processing. The review identifies gaps in the literature, thereby establishing the context and motivation for novel research. What follows is a review of the evolution of audio technologies relevant to the research question.

Audio zooming, the cocktail party problem, sound source localisation, source separation, and noise reduction are all deep research areas with substantial prior work. However, much of this work has been developed either for laboratory conditions or for applications such as teleconferencing, hearing aids and general speech enhancement, rather than for police-operated drones in noisy urban scenes. As a result, there is limited guidance on how to design a lightweight, drone-mounted microphone array and processing chain that can deliver forensic-grade, reproducible audio zooming in the kinds of environments illustrated by the Ava White case. The remaining sections of this chapter therefore examine each component technology in more depth and identify where additional work is required to support the drone-based surveillance scenario that this thesis addresses.

3.2 Audio Zooming and The Cocktail Party Problem

3.2.1 Audio Zoom and Directional Capture

This subsection reviews early work on directional microphones and audio zoom concepts that pre-date modern array-based processing. The term ‘Audio Zooming' refers to a concept where a remote user can select a listening position within an audio scene, examining the audio related to that specific location.

In an attempt to tackle the Cocktail Party Problem (CPP), a method to achieve audio zooming must first be determined. Spatial filtering, commonly known as beamforming, was one of the earliest methods to be employed (Capon, 1969). These techniques address the filtering of unwanted noise by slightly changing the sounds collected from different microphones. The human auditory system has two main skills: it can tell where a sound is coming from, and then

focus on that sound. Beamformers aim to work in a similar way to ultimately make target sounds clearer and more focused. In the same way that a camera's zoom lets you focus on something specific far away, beamforming enables you to focus on sound in a similar way.

Delay-and-sum beamforming is an early technique used in the field of spatial filtering. It relies on basic principles described by G. C. Southworth (1946) in his seminal writing on acoustic wavefront reconstruction.

Delay-and-sum, sometimes referred to as time delay beamforming, works by aligning and then summing any received signals to create a final focused signal. Southworth was one of the first researchers to deeply understand how sound behaves in air and realised that sound waves could be thought of and treated like light as a wave of differing frequencies. Southworth's paper (1946), which focused on classic audio engineering techniques, put forward the theory that it is important to understand and appreciate how acoustic wavefronts can be aligned and combined, leading to the idea that a receiver can 'hear' a focused signal from a single source even when it is mixed with other competing sounds. Southworth made important discoveries in many areas, such as medicine, engineering, and physics, in addition to advances in the development of hearing aids.

Olson and Preston (1949) first introduced the idea of a machine that could "zoom" into an audio scene. They achieved this during the development of the Single Ribbon cardioid Microphone, which was an improvement over any previously developed microphone for eliminating sounds to the rear. Olson and Preston's microphone additionally had an adaptive cardioid pickup pattern. When recording higher frequencies, for example at 10 kHz, the microphone exhibited increasingly super-cardioid behaviour, with a narrower forward lobe and a small rear lobe, in contrast to the standard cardioid polar pattern observed at 1 kHz, where the response is strongest to the front, reduced at the sides and strongly attenuated at the rear. Olson and Preston's microphone introduced adaptive pickup patterns, a groundbreaking first step in directional audio capture. Since it was just a single sensor, it did not have the necessary possible adjustment for many of today's applications, including surveillance. Its design prioritises high-frequency sounds, for which it has excellent directionality. This would limit its effectiveness in complex acoustic scenes, and its likely size and power requirements would make it ill-suited for today's portable systems.

3.2.2 The Cocktail Party Problem

The Cocktail Party Problem literature characterises the perceptual abilities of human listeners and underpins many engineering approaches to audio zoom.

The concept of the "Cocktail Party Problem" came from Cherry (1953), whose name is now coupled with this important work. The CPP highlights one of the brain's exceptional capabilities, namely auditory attention to competing sound sources. The classic example of CPP is a situation in which a person is trying to have a conversation while at a noisy party (multiple voices, cutlery and crockery clashing and perhaps some background music). Somehow, this individual at the party manages to remain focused on their conversation. The theory is that, in this scenario, a person has an innate auditory switch that allows them to almost effortlessly filter out most of the surrounding sounds and focus on desired sound sources.

This selective auditory focus is analogous to visual pattern completion, where readers can recognise words even when some letters are missing or displaced. In noisy environments, the auditory system exploits brief moments when the background level drops to obtain clearer fragments of the target speech, and then uses linguistic and contextual knowledge to fill in the masked segments. While a microphone array and associated signal-processing algorithms cannot fully replicate the human cognitive process, the system is analogous, in that it enhances sound from a desired direction and suppresses competing sources, thereby making the target speech more accessible to a human listener.

Although Cherry's work was foundational, laying the groundwork for auditory scene analysis and future audio separation research, real-world engineering solutions to the cocktail party problem remain elusive. A reason for this seems to be that, as computing has advanced over the years, intelligent systems still struggle to deal with interfering frequencies and complex audio events within an auditory scene. To be more human-like, these systems would require a deeper understanding of sounds at an increasingly complex level and, even more, to understand their relationship to the whole scene (Bronkhorst, 2015). In short, machines must achieve an elusive kind of "robustness" at the level of perception and cognition that living things seem to do effortlessly.

3.2.3 Beamforming and Array Based Zoom

A complementary line of work applies classical beamforming and spatial filtering to realise audio zoom with microphone arrays rather than single adaptive microphones.

Capon (1969) introduced the first Minimum Variance Distortionless Response (MVDR) beamformer in 1969. It was a significant step forward in spatial filtering. Unlike early techniques such as delay-and-sum, the MVDR beamformer combines the array's inputs and filters more sophisticatedly. It does this by changing the weights of the array sensors. This is preferable because the method can treat interference as distortion, which will then be reduced. This would, in theory, result in a clearer output signal from the target. Interference and noise from other directions are reduced because the MVDR beamformer uses the spatial covariance matrix of the received signals to attenuate unwanted sounds.

The first reference to an audio zoom, an attempt to create a sonic equivalent of the then-recently introduced video zoom, was first introduced by JVC in 1980 (Ishigaki *et al.*, 1980). A second-order gradient unidirectional microphone was developed for JVC with a frequency response of 100Hz – 10kHz. This experiment relied on using two well-matched electret microphones, doubling as drivers for the audio zoom's necessary phase and amplitude manipulations.

The idea of "audio zoom," was a novel step towards linking auditory and visual technologies. However, this early implementation was limited in scope and practical application. This imposed constraints and potentially limited cost-effectiveness and scalability. This system's operational fidelity was further challenged by requiring very precise phase and amplitude manipulations that may be computationally intensive in a virtual world. While the frequency response range of 100Hz to 10kHz is well suited for basic speech and environmental sounds, it does not encompass the full human auditory spectrum.

The influential review by Veen and Buckley (1988) revisited fundamental beamforming techniques. These include the delay-and-sum method and MVDR. Veen and Buckley compared their strengths and weaknesses, which were discussed earlier in this chapter. To summarise, the delay-and-sum method is simple and computationally inexpensive. It performs adequately in undemanding scenarios but is limited in scenarios where the interference overpowers the desired signal. MVDR offers much better performance in challenging, reverberant situations. It provides good noise cancellation and distortionless response when summing the signals. This means it can separate and control signals that are close together and could be ideal for employment in real-world surveillance operations.

Matsumoto and Naono (1989) achieved a dual channel stereo effect where, with respect to the listener, one of the microphone pair had to operate as an omnidirectional microphone in front of

the sound source, while the other had to function as a super-cardioid with the kind of sensitivity and pickup pattern which, for all practical purposes, could cause it to operate as a "spaced pair" stereo zoom microphone. The dependence on static microphone arrangements could potentially restrict functionality in complex or variable soundscapes.

3.2.4 CASA and Machine Learning Approaches

Beyond classical array processing, Computational Auditory Scene Analysis and, more recently, machine-learning approaches have been proposed as routes towards solving the Cocktail Party Problem.

In answer to the question, "Is it possible to build a machine capable of solving it *[The Cocktail Party Problem]* in a satisfactory manner?" (Haykin and Chen, 2005), the authors turned to the work of Wang and Brown (1999), concluding that by advancing Machine learning, and Computational Auditory Scene Analysis (CASA), in particular, it may be possible to develop a simulation of a soundscape and record, isolate and download a desired sound signal. Equation (3.1), is from a review by Haykin and Chen (2005) of Wang and Brown's work, who developed a two-layer oscillator network that performed audio source separation:

$$C = \frac{\sum x(t)y(t)}{\sqrt{\sum x^2(t) \sum y^2(t)}} \tag{3.1}$$

Real-world audio complexity poses a challenge for CASA-based methods (Brown and Cooke, 1994). They have difficulty with overlapping audio, transient sounds, and quickly changing environments. The two-layer network of oscillators first proposed at MIT in 1992 was successful in controlled conditions, but is possibly too computationally demanding to be useful in real-time applications. While audio science has moved on since the 1990s, there is still a significant gap between the performance of computational models and human listeners.

While increases in computational power may make such models more tractable in real time, it is not yet clear whether this will close the gap to the flexibility and robustness of human auditory perception.

These CASA and machine learning-based approaches have enchanced audio zoom and Cocktail Party Problem research, particularly in terms of modelling complex, non-linear mixtures and exploiting linguistic context. Deep neural beamformers and end-to-end CPP systems can deliver strong separation and recognition performance on curated datasets that can sometimes exceed what simple delay-and-sum or MVDR can achieve on their own. However, this performance

comes at the cost of heavy training requirements, sensitivity to dataset bias and a loss of transparency: it becomes difficult to explain why a given model fails in a particular real-world scene. These trade-offs are especially problematic in forensic applications, where fast reproducibility and interpretability are as important as raw enhancement performance.

3.2.5 Perceptual Acoustic Zoom and DirAC

A further branch of work focuses on perceptual "acoustic zoom" based on manipulating distance cues, diffuseness and directional parameters.

Schultz-Amling *et al*. (2010) published an Acoustical Zooming paper investigating Directional Audio Coding (DirAC). Directions of arrival and sound diffuseness are the two facets of sound that DirAC treated as parameters. In their system, the sound field is analysed to estimate the dominant direction of arrival and separate more directional components from diffuse background energy, and the current dominant direction is treated as the sound of interest. While the work from these authors was intended to advance teleconferencing in terms of the user interface, the idea of automatically steering a video image to the active talker could have interesting potential parallels in the use of drones to steer audio and video to a particular location of interest on the ground. While the work of Schultz-Amling *et al.*'s work on DirAC was innovative (2010), leveraging such sound parameters as direction of arrival and sound diffuseness, it nonetheless would have limitations when applied in broader contexts, such as using drones to steer audio to specified locations. The authors' initial focus on teleconferencing seems to tilt the DirAC work towards something highly specialised: a system likely optimised for controlled, indoor environments. The work suggests that DirAC may not be the best candidate for unpredictable real-world outdoor environments.

Further research on acoustical zooming using a multi-microphone array was undertaken by Van Waterschoot *et al*. (2013). Their paper presented a general theory that allowed for independent control over the sound source levels. This, in turn, allowed for the creation of an acoustic zoom effect. An interesting part of the research was that it did not rely on a computationally expensive source separation algorithm, as previous work had done. Instead, it used two less computationally demanding algorithms: one that reduced spatial noise and the other that reduced spectral noise. It was economical in weight and cost and employed cheap, lightweight microphones attached to a simple array.

The Acoustic Zoom (AZ) effect is based on modifying one or more cues associated with the perception of sound source distance. Numerous factors combine to enable the auditory system to perceive auditory distance; sound intensity is the first and perhaps the most fundamental. Several other factors also contribute. The direct-to-reverberant ratio (DRR), or how much of a sound reaches the listener directly, as opposed to via reflections; spectral distortion, or how much a sound's frequency content has changed (and how much it ought to have changed, according to an auditory system's guesses as to how far it has travelled); and differences in the times at which a sound reaches each ear (the Interaural Time Difference, or ITD) and the levels at which it is received (the Interaural Level Difference, or ILD). Identifying how a sound moves is a cue to its distance. Relying on affecting changes in perceptual cues such as the intensity of sound, the DRR, and spectral distortion presents potential limitations. These cues can change in unpredictable ways in complex, real-world environments that contain multiple kinds of noise, strong reverberations, and an overall dynamic soundscape. The effectiveness of this method seems likely to decrease under such circumstances, especially when applied to situations such as drone surveillance or outdoor recording, where environmental factors and interference would likely occur (Schulte-Fortkamp and Jordan, 2023).

Thiergart, Kowalczyk and Habets (2014) agreed that the most effective means of achieving Acoustic Zooming was through the use of spatial filters. In a similar manner, Christensen *et al*. (2016) used a full-rank Wiener subspace filter with dynamic rank limiting to enhance speech in a simulation involving a circular microphone array. Both studies advanced understanding, but at the same time, reveal some inherent challenges to the approaches taken. The spatial filtering techniques used in the studies are highly array-dependent, array configuration, and source-localisation sensitive. That is, they only perform well when the sound sources being studied are located where they are supposed to be, in a controlled laboratory space, and when the array itself is perfectly intact. So, while it is a relatively computationally inexpensive approach, it may not be robust enough to deal with rapidly changing complex auditory scenes.

Wilson (2017) found that in CPP situations, it is natural for humans to increase the signal-to-noise ratio. This was accomplished binaurally by using interaural time differences, interaural level differences, and interaural decorrelation. The mechanisms by which the auditory system solves sound-space problems, essentially the processes by which auditory perception "engineers" the auditory experience, calls to mind the kinds of gain, pan, and other audio engineering techniques one might use to achieve the same result with a sound system. Yet, while the human auditory system adapts effortlessly to intricate sounds, artificially applying

these methods still poses a problem. The technologies that seek to duplicate this process often fall short of our auditory system's finely detailed spatial perception and adaptability, especially in noisy, dynamic situations where Police surveillance typically occurs.

Zhang *et al*. (2022) considered speech recognition in a CPP environment, using a neural beamformer-based frontend and an automatic speech recognition (ASR) backend, which is end-to-end optimised. The findings presented here reinforce the view that spatial filtering (beamforming), offers a practical route to both sound-source localisation and audio zooming. Even so, several drawbacks remain. Neural beamformers, while generally resilient to noise and interference, impose a heavy computational load. Implementing them in real time on size, weight and power-constrained platforms such as small drones therefore becomes a serious problem to consider. End-to-end learning frameworks add a second obstacle: they rarely converge without time-consuming data curation, and even extensive training sets can produce models that overlook features critical to other parts of the processing chain (Webb *et al.*, 2019).

The literature on audio zooming and the CPP demonstrates that directional microphones, classical beamforming and more recent multi-microphone and audio-visual methods can substantially enhance a target source in complex mixtures. Nevertheless, most of these systems assume static or handheld platforms, relatively controlled acoustic conditions and do not explicitly account for rotor noise, platform motion or evidential chain requirements. There is also comparatively little work that links audio zoom directly to a moving video frame in a way that is practical for post-event forensic analysis of CCTV-style material. This thesis responds to these gaps by developing a drone-oriented audio zooming pipeline based on MVDR beamforming and post-filtering, explicitly designed to synchronise with video zoom and to operate under the noise, weight and reproducibility constraints of police surveillance.

3.3 Drone Microphone Array Signal Processing

3.3.1 Beamforming and Localisation with Drone-Mounted Arrays

Over the past decade there has been a growing body of work on microphone-array signal processing specifically for drones. Studies such as Salvati *et al.* (2018) demonstrated that a compact microphone array mounted on a micro-air vehicle (MAV) can be used with beamforming to localise acoustic sources while the platform is hovering. Their experiments explicitly considered the impact of rotor noise and showed that, under steady-flight conditions, spatial filtering is still able to recover meaningful directional information.

Go and Choi (2021) extended this line of work by combining beamforming on a drone-embedded array with navigation and attitude data from the flight controller, improving direction-of-arrival estimates by fusing acoustic and inertial information.

3.3.2 Learning-Based Sound Localisation

Subsequent contributions have increasingly exploited learning-based localisation, for example Wang and Cavallaro (2022) and Wang, Clayton and Rossberg (2023) who proposed deep-learning-assisted sound source localisation from a flying drone, using neural networks to compensate for array perturbations and platform motion that degrade classical direction-of-arrival algorithms.

More recent work has moved towards multi-drone and tracking scenarios. Yen et al. (2024) investigated cooperative sound source tracking using multiple UAVs, incorporating rotor-noise-informed SNR estimation into their algorithms. By explicitly modelling the egonoise characteristics of each platform, they were able to improve tracking robustness in very low SNR conditions and during manoeuvres. These studies collectively show that both classical and learning-based array processing can be adapted to handle platform motion and strong, non-stationary propulsion noise, at least for localisation and tracking tasks.

3.3.3 Egonoise Reduction

A parallel thread of research focuses on rotor egonoise reduction and speech enhancement directly from on-board microphones. Wang and Cavallaro (2020) proposed a blind source separation framework for egonoise suppression on multi-rotor drones, treating the rotor noise as a dominant but unknown source and using BSS techniques to recover the residual target signals at extremely low input SNRs. Manamperi *et al*. (2024) built on this by presenting an audio signal-enhancement pipeline for drone-embedded microphones that combines array processing with post-filtering, again targeting recordings dominated by broadband rotor noise. These systems demonstrate that it is possible to substantially improve the quality of on-board recordings by explicitly modelling or separating the egonoise component, but they are typically evaluated on controlled datasets and generic speech or environmental sounds rather than on weak, far-field conversational speech in dense urban scenes.

3.3.4 Drone Audition Pipelines

Survey work on drone audition places these algorithms within broader end-to-end systems for environmental and bioacoustic monitoring. Wang, Clayton and Rossberg (2023) reviewed drone audition for bioacoustic monitoring, outlining typical end-to-end pipelines that integrate array design, egonoise mitigation, localisation and downstream analysis. Their survey emphasises the importance of jointly considering hardware configuration, signal processing and the target application when designing UAV-based listening systems. However, most of the applications discussed are ecological or environmental rather than forensic, and the evaluation criteria are framed around detection and classification performance rather than intelligibility or evidential suitability.

Taken together, this body of work confirms that drones are viable platforms for microphone-array processing and that both traditional beamforming and modern learning-based methods can address key challenges such as rotor egonoise, platform motion and low SNR. Nonetheless, the primary focus has been on localisation, tracking, generic enhancement or bioacoustic monitoring, with comparatively little attention to the recovery of intelligible, far-field conversational speech in complex urban environments or to the documentation and reproducibility requirements of forensic practice. The present thesis builds on these drone audition techniques but redirects them towards a crime-scene-motivated audio-zooming problem: designing a lightweight array and processing chain suitable for Police-style drones that can improve the intelligibility of target speech linked to video evidence, while remaining transparent enough to support an evidential workflow.

3.3.5 Onboard Microphone Array Design

Several groups have built microphone arrays that are physically integrated with drone platforms. Salvati *et al.* (2018) mounted a compact four-microphone uniform linear array, repurposed from a PlayStation Eye device, on top of a micro-air vehicle and used conventional delay-and-sum beamforming to localise ground-based acoustic events while hovering. The array aperture was kept small to respect weight and form-factor limits, and the study showed that meaningful direction-of-arrival estimates could still be obtained despite strong rotor noise, provided the drone remained in steady flight.

Go and Choi (2021) extended this idea to a denser, 32-channel time-synchronised MEMS phased array mounted on a quadrotor, using beamforming for 3D acoustic source localisation. Their design illustrates the trade-off between array aperture and practicality: the larger planar array

improved localisation accuracy at a ground range of ~150 m but increased payload mass and required careful mounting relative to the rotors.

Other work has explored spherical and irregular 3D arrays. Hoshiba *et al.* (2024) used a 16-channel spherical MEMS array (12 microphones in the lower hemisphere and 4 in the upper hemisphere of a 110-mm body) suspended 600 mm below a DJI Inspire 2 via a rigid pipe, and evaluated several MUSIC-type algorithms for robust localisation under dynamic ego-noise. In contrast, Manamperi et al. (2024) embedded 15 dual-microphone MEMS modules (30 channels) directly on the drone frame—under the motors, on the landing gear and along the arms—to capture ego-noise and target signals in situ. Their system was used both for localisation and for multichannel Wiener filtering plus GMM-based post-filtering, and demonstrates the other extreme of the design space: an irregular, body-conformal array that tightly follows the airframe but experiences severe scattering and low SNR near the propellers.

Many speech-enhancement and ego-noise-reduction studies use similar on-board arrays, often via shared datasets. These designs are typically optimised for algorithm evaluation providing rich spatial information and strong rotor noise rather than for low-weight, evidence-oriented field deployment. The emphasis is on demonstrating rotor noise suppression or localisation accuracy, not on manufacturability, ruggedisation or forensic traceability.

Table 3.X summarises representative onboard microphone arrays from the literature in terms of platform, array configuration and main processing goals.

Table 3.1. Onboard microphone arrays from the literature.

Study	Platform	Sensors	Geometry	Processing
Salvati et al. (2018)	Parrot Bebop 1 MAV, array on top of body	4 electret mics (PlayStation Eye)	Short ULA, small aperture	Delay-and-sum beamforming
Go & Choi (2021)	Quadrotor with underside phased array	32 time-synced MEMS mics	Planar phased array	Beamforming + spectral subtraction + nav fusion
Hoshiba et al. (2024)	DJI Inspire 2, array 0.6 m below via pipe	16 MEMS mics	110-mm spherical array (12 lower, 4 upper)	SEVD-MUSIC, GEVD-MUSIC, HIST-MUSIC-3D
Manamperi et al. (2024)	Phantom-class UAV, array embedded on frame	30 MEMS channels (15 dual-mic modules)	Irregular body-conformal array on arms, motors	Multichannel Wiener filtering + GMM post-filtering
Wang & Cavallaro (2020)	Multi-rotor UAVs (DREGON dataset)	Up to tens of onboard mics	Various: arm- and body-mounted arrays	Blind source separation for ego-noise reduction

3.4 Sound Source Localisation

3.4.1 Definitions and Classical TDOA/DOA Framework

This subsection introduces the basic formulation of sound source localisation and the main families of algorithms used to estimate direction of arrival. Addressing the CCP often demands that the system first locate the target signals before attempting any form of source separation. SSL refers to estimating the exact position of an emitter within an acoustic scene. Locations are normally expressed in spherical coordinates, comprising an azimuth angle (φ), an elevation or polar angle (θ), and a radial distance (*r*). Typically, only the Azimuth angle (φ) and the polar angle (θ) are needed to estimate the Direction of Arrival (DOA).

SSL is typically achieved using one of three methods: Calculating the Time Difference on Arrival, Beamforming or Subspace Methods (Argentieri, Danès and Souères, 2015). These methods are all potential frameworks for SSL in controlled settings, yet they each have limitations that could make them less than ideal for real-world applications. They can all struggle with computational efficiency, robustness, and adaptation to changing sonic environments. When

attempting to achieve SSL in a real-world outdoor environment with a surveillance system, perhaps a hybrid, adaptive approach would be more sensible (Jekateryńczuk and Piotrowski, 2024).

3.4.2 Psychoacoustic and Audio-Visual Localisation Studies

Some work has examined localisation from a perceptual or audio-visual perspective. Yost, Dye and Sheft (1996) introduced a Knowles Electronic Manikin head for Acoustical Research (KEMAR) into the research field. Sound sources were added, and the argument was put forth that spatial cues are neither necessary nor sufficient for sound source determination. However, their methods have potential issues, such as the test subjects being able to repeat the sounds on the simple setup as often as desired before issuing an answer.

By 2018, trained neural network solutions to the problem of sound localisation and separation were becoming popular. Owens and Efros (2018) tested a joint audio and visual model that used a neural network learning approach to achieve sound localisation. The detection of temporal misalignment achieved localisation. The obvious drawback to these methods was the computational power and time needed to train and run the network. As such, by 2019, alternative approaches were being developed in the form of algorithms that did not rely on neural networks for sound localisation.

3.4.3 Subspace and High Resolution Methods

High-resolution subspace methods such as ESPRIT and MUSIC, and more recent variants like SVD-PHAT, aim to improve localisation accuracy beyond basic TDOA schemes. Paulraj, Roy and Kailath (1985) developed their Estimation of Signal Parameters via the Rotational Invariance Technique (ESPRIT), which required significantly fewer computations than previous methods of SSL.

Alternatively, Schmidt (1986) employed a microphone array to develop their Multiple Signal Classification (MUSIC) approach. The method involves online computation of eigenvectors, however, while producing good localisation results, the computational power needed to employ the process made real-time implementation challenging, but MUSIC only required half the amount of microphones of ESPRIT to achieve similar results.

Singular Value Decomposition-Phase Transform (SVD-PHAT) was proposed to improve localisation accuracy and reduce algorithm complexity. SVD-PHAT localises multiple sound

sources more accurately than SRP-PHAT, with a reduction in the root mean square (RMS) error of up to 0.0395 radians (Grondin and Glass, 2019). While SVD-PHAT showed great progress, it may be argued that it was too dependent on accurate phase information to be practical in environments where the sound is very reverberant or where there is a lot of direct sound. These are essentially the same limitations for which standard beamforming has been criticised. Another issue with SVD-PHAT is that, even though there were performance gains over SRP-PHAT (and thus over standard beamforming), they were not large enough to make it a clear choice for sound localisation systems that need to be physically realised and that operate across a wide range of acoustic environments, such as a sound localisation system mounted on a drone.

3.4.4 Ad Hoc and Drone-Mounted Localisation

Liaquat *et al*. (2021) suggested Sound Localisation for Ad hoc Microphone Arrays. A microphone array with only four microphones was used to limit the algorithm's complexity and computational time. On average, this sound localisation system was 96.7% accurate, a significantly improved result over the deep learning and neural network methods which were 71.46% accurate. This process of sound localisation was based on time delay. It required the microphones in the ad-hoc microphone array to be time-synchronised and arranged in a particular geometry. Their sound localisation method involved two estimation steps: the DOA estimates and the distance estimates.

The method depends on the microphones' exact time synchronisation and specific geometric arrangements, which presents the practical challenges of achieving and maintaining such synchronisation, particularly in dynamic or outdoor environments with more unpredictable environments. Achieving and maintaining synchronisation in a dynamic or outdoor environment is challenging. Regarding the system's robustness, there is a dependence on accurate DOA and distance estimates, which may present difficulties in highly reverberant spaces.

Manamperi *et al*. (2022) researched sound source localisation using an irregularly shaped microphone array attached to a drone. The predictable problem was the noise of the drone's motors/propellers, which led to a low signal-to-noise ratio (SNR). Their work proposed cross-correlation based on the DOA estimation technique using the time difference of arrival (TDOA). The method could estimate the position in 3D space for simultaneously active multiple sound sources on the ground at low SNR Conditions (-30dB). The results obtained from this method increase the feasibility of localisation under extreme SNR levels.

While the method showed potential, it still had drawbacks. It depends on cross-correlation and time-difference-of-arrival (TDOA), making it susceptible to environmental problems such as wind, reflections, and the kind of dynamic interference that people and vehicles will bring. For the system to work effectively in real-world situations, the computational power needed could push the limits of what is possible with lightweight drone-based processors. And while the Manamperi *et al.* (2022) paper showed promise in the controlled conditions of the laboratory, the question of how well it would work in a more chaotic real-world environment remained unanswered.

Existing sound source localisation research offers a wide range of tools, from ITD/ILD and GCC-PHAT methods through to subspace algorithms such as MUSIC and ESPRIT, and more recent learning-based direction-of-arrival estimators. These approaches are effective for fixed arrays in relatively stable acoustic conditions, but far fewer studies consider irregular or compact arrays mounted on moving drones, where array geometry perturbations, wind, reflections and low SNR can all degrade localisation performance. Moreover, localisation is rarely integrated with a complete audio-zooming workflow that runs from array design through to intelligibility metrics on realistic crime-scene-like material.

In this thesis, the localisation principles reviewed here are embedded into a practical array geometry and beam steering strategy for drone carriage, and are evaluated through a combination of controlled booth tests, simulations and Exemplar House experiments in later chapters.

3.5 Source Separation

3.5.1 CASA and Early Computational Approaches

Early work on source separation focused on Computational Auditory Scene Analysis and model-based multichannel frameworks. There was a consensus within the literature that the CPP is a challenging problem to solve technologically, as both interfering sources and reverberation need to be controlled.

Computational Auditory Scene Analysis (CASA) (Brown and Cooke, 1994) was essentially a source segregation system. It could construct a recreation of an audio scene and separated the elements by frequency contours before synthesising a waveform. While promising, CASA still presented some difficulties. It depended on accurate frequency contour extractions, making it susceptible to noise and distortion, which are common in real-world environments. The computational effort it takes to reconstruct and synthesise the waveforms perhaps limits the

applicability to post-production situations where one has enough resources, such as a desktop computer. Finally, CASA displays issues in dealing with dynamic acoustic environments when several sources of sound overlap or happen quickly, one after the other.

A potential solution to these shortcomings was proposed by Huang, Benesty and Chen (2006) who used blind SIMO (Single Input Multiple Output) for interference identification, which was converted from MIMO (Multiple Input Multiple Output) before being processed with MINT (Multiple Input/Output Inverse Theorem) for dereverberation purposes. This method of blindly identifying and cancelling the interfering sources allowed the targeted signal to be extracted.

Accurate mathematical modelling of the target and interference sources is the first step in the signal processing of MIMO systems. However, in real-world scenarios, the direct characterisation of these signals is often imprecise owing to the inherent variability of the environment and noise. The signal processing chain will not work properly without accurate target and interference modelling. At best, it will yield an imprecise estimate of the array output, from which the target state can be inferred. The work of Makino *et al*. (2005) has shown steps toward addressing these major limitations, which are threefold: (1) mobility provides rich data for characterisation of the target and interference during the signal processing; (2) precise mathematical modelling (obtained from the characterisation) yields better performance in MIMO systems; and (3) a high-precision signal processing chain allows inversion to work correctly.

3.5.2 ICA and Blind Source Separation

Smita, Biswas and Solanki (2007) discussed Independent Component Analysis (ICA), which uses Blind Source Separation (BSS), Harmonic structure stability (HSS), and a Second Order Statistical Measures (SOSM) approach to achieve source separation. Smita, Biswas and Solanki concluded that signal separation remained a complex problem with no reliable methods for the general case. While classifiers such as Support Vector Machine (SVM) are generally more accurate than traditional classifiers, ICA is a reliable machine-learning approach to solving source separation in audio mixtures, which could then be applied to the CPP.

This conclusion was supported by Shlens (2014), who stated that:

"The goal of ICA is to solve BSS [Blind Source Separation] problems which arise from a linear mixture. In fact, the prototypical problem addressed by ICA is the cocktail party problem"

A critique of ICA could be its assumption of isolated sound sources, which is often not the case in the real world. Sounds frequently overlap and share similar frequency ranges. Even so, ICA is a sensible method that computes with reasonable efficiency. Its strength lies in where it has been applied to well-understood problems with known statistics. Additional techniques may be needed to expand ICA's realm of application toward more ill-defined problems such as real-world surveillance. In parallel, many neural network systems in other domains make use of principal component analysis (PCA) for dimensionality reduction and, more recently, federated learning frameworks for distributed or privacy-preserving model training; however, these approaches are beyond the scope of the present work. Figure 3.1 shows an adaptation of a schematic drawing of the CPP and ICA being used to solve the problem (Hwang *et al.*, 2019).

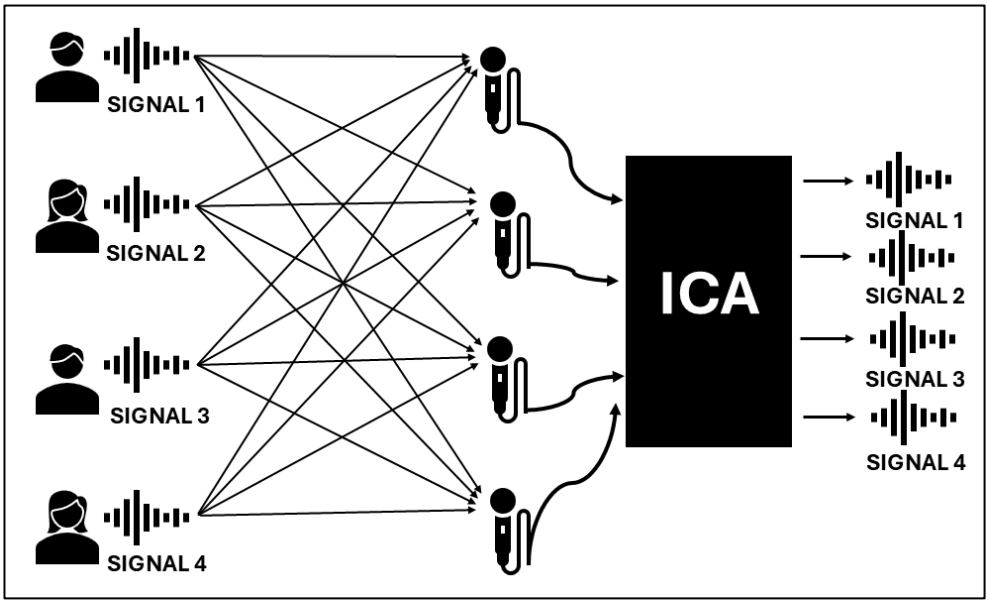

Figure 3.1. Cocktail Party Problem and a proposed ICA solution. (Adapted from Hwang *et al*., 2019).

The blind source separation framework proposed by Wang and Cavallaro (2020) illustrates how these ideas can be applied directly to drone-recorded audio. In their work, the rotor egonoise produced by a drone is treated as a dominant but unknown source, and BSS is used to separate and attenuate this component, thereby improving the target signals. This represents an example where source separation is explicitly tailored to UAV conditions, and it demonstrates that BSS can be effective even when the interference is broadband and highly non-stationary. However, as discussed in Section 3.3, their evaluation focuses on controlled scenarios and generic speech or environmental sounds rather than weak, far-field conversational speech in dense urban scenes, and the framework is not developed with forensic transparency or evidential reporting in mind.

3.5.3 Supervised Deep Learning

Kavalerov *et al*. (2019) developed a deep-learning method for speech separation and enhancement. They trained their models using a large Dataset of mixtures. Using an Iterative Time-Dilated Convolutional Network (TDCN), they produced an average Scale-invariant signal-to-distortion ratio (SI-SDR), achieving nearly 10dB enhancement in sound separation and more than a 13dB enhancement in speech versus non-speech separation. The authors felt that this result indicated that universal sound source separation would soon be within reach.

There are limitations to the method of Kavalerov *et al*. despite the promising nature of the results. Large training datasets are required, and this raises concerns about resource scalability. There would be a need to curate and clean the kinds of datasets TDCN, and other deep learning methods require to be truly diverse and mostly free of the noise that real-world datasets are usually rife with. It would be expected that the popular deep-learning models will be usable under a wide range of real-world conditions. Working in laboratory conditions the results are promising, but can they be expected to work under real-world conditions?

Tzinis *et al*. (2019) attempted Unsupervised Deep Clustering for Source Separation (only exposing a machine to input data). Their proposed solution was a monophonic source separation system, trained by observing mixtures of sounds. They employed a deep clustering approach that did not use spatial information, pre-learned dictionaries, or neural network models to identify sound sources. They reasoned that humans evolved by listening to 'mixtures' of sounds and identifying the individual sources, even without any spatial cues, which echoed the conclusions of Yost *et al*. (1996). Once trained, the process aimed to separate signals within a single channel. (i.e., a mixture of sounds within a mono audio file). The goal was to develop a system that could form source models in an unsupervised manner instead of requiring training data designed by their users.

This method is conceptually impressive but still has some limitations. It assumes that the training of models without human help will occur, but this would demand sound algorithms that consistently cluster meaningful features from sound mixtures. However, their performance may be poor in some contexts, such as noisy environments, or when the sounds we want to recognise are too similar. Since the system is monophonic, it cannot take advantage of directional information (TDOA). It is well established that our biological spatial hearing mechanisms work well even in very challenging auditory scenes. This method would not take advantage of this fact.

As this system uses an unsupervised method, the performance of unsupervised approaches often lags behind that of supervised methods, especially when high accuracy is required.

Tzinis *et al*. (2020a) proposed a more efficient neural network for audio source separation. The model was based on "Successful down-sampling and resampling of multi-resolution features" (SUDO RM-RF). Their concept was that SUDO RM-RF could eventually be installed on devices with limited resources and possibly trained quicker than earlier models. The approach of achieving such a system was more efficient and could be employed on an unmanned aerial vehicle such as a drone.

Although the SUDO RM-RF framework lowers the overall computational burden, the tight energy budget, limited memory, and modest clock rates typical of unmanned aerial vehicles can still throttle performance, especially in highly dynamic or unpredictable sound fields. Robustness across a broad set of acoustic environments remains ambitious because such systems are difficult to pre-train effectively (Amiriparian *et al.*, 2020).

3.5.4 Deep Learning and Universal Source Separation

Building on the audio-visual correspondence work of Owens and Efros (2018), Tzinis *et al*. (2021) introduced Audioscope, which separates on-screen sound sources without supervision by means of a method called Mixture-Invariant Training (MixIT). An advantage of MixIT is that it allows for the separation of synthetic Mixtures of Mixtures (MOM) into individual sources.

Similar to other proposed methods, MixIT faces challenges. Its performance tends to decrease in real-world settings, which are far more complex and variable than ideal synthetic soundscapes. Hence, the method tends to be less effective when it is needed the most: in uncertain, dynamic environments where the signals of interest are likely to be accompanied by a broad range of other competing sounds. Meanwhile, the computational cost of training the model to work well with vision built into the sound source is hardly trivial; real diversity in sound and vision, represented in a dataset for training in such a model, is a considerable challenge.

The authors of the Audioscope paper (2021) had previously researched improving separation performance by utilising conditional information regarding sound classes (Tzinis *et al.*, 2020b). Earlier efforts to achieve 'Universal Sound Separation' demonstrated that the task is feasible when working with a set number of sounds (Kavalerov *et al.*, 2019). All these previous methods needed well-structured datasets comprising isolated sounds for training purposes. The Audioscope paper, on the other hand, managed to accomplish sound separation without the aid of well-structured

datasets comprising well-defined sound classes, either isolated or in the wild. The lack of structured training data may hinder its consistency and accuracy in exceedingly intricate or noisy situations. Still, further research is essential if this technology is to work reliably in varied and complex, real life auditory environments.

In recent years, Liaquat *et al*. (2021) used ICA to extract the localised source signals on Ad-Hoc Microphone Arrays. This research was informative when planning the project's potential microphone array designs. The dependence on ICA, again, assumes that the sources are statistically independent. This may not be the case in the real world, where it seems more plausible to have overlapping or even correlated sound sources.

3.5.5 Limitations for Drone-Mounted Forensic Audio

Zhang *et al*. (2022) proposed a multi-speaker ASR model, denoising and separating speech from noises using a neural beamformer that performed well in noisy and reverberant conditions in their tests. Still, neural beamformers are computationally intensive and need a lot of training data, sometimes even more than what is usually required for neural networks (Chen *et al.*, 2024). This may limit where and how quickly they can be applied.

The source separation literature spans classical statistical methods such as ICA and BSS, mask-based approaches and powerful deep learning architectures that achieve impressive SI-SDR gains on benchmark datasets. However, these systems typically rely on large, curated training data, assume predictable mixtures and often target close-miked or studio-style recordings, rather than weak, far-field speech embedded in outdoor drone recordings. In addition, many learning-based methods are computationally heavy and difficult to validate within a forensic chain of custody, where transparency and reproducibility of the processing steps are essential. In this thesis, separation concepts are used in a more constrained way: as lightweight post-filters and masks applied to beamformed signals to improve SNR and intelligibility, trading some absolute performance for robustness, explainability and suitability for deployment on or alongside police drone systems.

3.6 Noise Reduction

3.6.1 Array Based Noise Reduction

In any research concerning drones, a significant obstacle will be the noise from the motors and the blades. If sufficient noise reduction could be applied, it could theoretically make any SSL and BSS easier to accomplish. Kaneda and Ohga (1986) presented a small-size microphone-array

system called AMNOR (Adaptive Microphone-array for Noise Reduction). The process was designed to help improve the accepted compromises of the time between noise reduction and frequency degradation of a desired audio signal. Though a significant advance in its time, AMNOR faces difficulties with real-time adaptability and robustness in drone flight environments. AMNOR's design was bound by the era's computational constraints and microphone technology.

The reduction was achieved by introducing a fictitious desired signal into the chain. The AMNOR system improved SNR by more than 15dB in the 300Hz - 3.2kHz frequency range. 300Hz - 3.2kHz being roughly the same bandwidth provided by telephones, the minimum which is required to decode speech accurately. Kataoka and Ichinose (1990) expanded upon this research in 1990 by demonstrating that spatial aliasing was the most critical factor in AMNOR performance.

Kataoka and Ichinose discovered that the spatial limitations of small microphone arrays affect their handling of higher-frequency components. This, in turn, impacts the performance of such arrays in modern applications that are attempting to deal with complex soundscapes.

3.6.2 Intelligibility Measures

Legibility of noisy speech would be a primary goal of any audio zoom related noise reduction. Traditional intelligibility assessment relied on human listeners transcribing sentence lists (Kalikow, Stevens and Elliott, 1977). While somewhat accurate, those tests are slow and costly, which drove the search for computer-based predictors. The Short-Time Objective Intelligibility (STOI) metric was introduced by Taal *et al.* (2010) as an intrusive measure for predicting how intelligible processed speech will be to a normal-hearing listener. In STOI, the clean reference and processed signals are first time-aligned and resampled, then passed through a one-third-octave filterbank covering the main speech band. Short-time temporal envelopes are extracted in each band using 384 ms windows with 50% overlap. These clean and processed envelopes are then normalised to remove overall level differences, truncated to limit the influence of very low local SNRs and compared by computing a linear correlation for each time–frequency segment. The final STOI score is obtained by averaging these local correlation values over time and frequency, giving a single number between 0 and 1, where higher values indicate higher predicted speech intelligibility.. Across three listening experiments it achieved correlations up to $\rho = 0.96$ with word-correct percentages, outperforming five prior metrics (Taal *et al.*, 2011).

The downside of the STOI method is that it can only deliver an accurate intelligibility grade if you give the algorithm a clean reference file of the targeted speech. In most field recordings such as surveillance captures, a clean reference is not available, but for laboratory work, it could prove to be an extremely useful tool.

Prior work on noise reduction provides a rich toolbox, including spectral subtraction, Wiener filtering, adaptive filtering and beamforming-based suppression, and more recent learning-based denoisers. Yet only a subset of this work explicitly tackles drone rotor noise, wind and traffic noise acting together on low-weight microphone arrays in open, reverberant spaces, and even fewer studies report upon intelligibility under such conditions. There is also limited discussion of how to balance algorithmic complexity, latency and power consumption when noise reduction is part of a wider drone surveillance and forensic workflow. The noise-control strategy adopted in this thesis therefore combines MVDR beamforming with computationally modest post-filters that specifically target rotor and environmental noise, and is evaluated for both SNR improvement and intelligibility in the simulated and real-world scenarios described in Chapters 4–7.

3.6.3 Speech Enhancement and Noise Reduction Using Drone Microphone Arrays

A growing subset of the speech enhancement literature focuses specifically on recordings made with microphone arrays mounted on multirotor drones. In these systems, the dominant challenge is rotor egonoise: broadband, non-stationary noise generated by the motors and propellers that can drive the input signal-to-noise ratio (SNR) to extremely low values. Classical signal-processing approaches adapt familiar techniques such as beamforming, spectral subtraction, multichannel Wiener filtering and adaptive filtering to this setting. For example, delay-and-sum and MVDR-type beamformers are used to steer spatial nulls towards the rotor plane while maintaining gain in ground-facing directions, and simple spectral subtraction stages can further attenuate harmonic components associated with the blade-passing frequency and its multiples. In other cases, multichannel Wiener filters are designed using noise covariance estimates obtained during hover or ascent phases, effectively extending single-channel enhancement concepts to the array domain under UAV-specific noise statistics.

More recent work has combined these classical front-ends with tailored rotor-noise models and ego-noise references. Wang and Cavallaro (2020) proposed a blind source separation framework that treats this egonoise as a dominant but unknown source in a multichannel mixture and uses BSS techniques to suppress it, thereby improving the residual target signals at very low SNRs.

Their experiments demonstrate that, when the mixture is rotor-noise dominated, multichannel separation can substantially improve the quality of the remaining audio, even though the interference is highly non-stationary.

Manamperi *et al.* (2024) extended this idea by presenting an audio signal-enhancement pipeline for drone-embedded microphones that combines array processing with post-filtering. Their system operates directly on noisy on-board recordings and aims to recover ground-level sources in the presence of strong rotor noise, showing further gains when the processing is tailored to the UAV geometry and flight conditions. Wang, Clayton and Rossberg (2023) reviewed such methods within complete drone-audition pipelines, emphasising that effective enhancement requires consideration of microphone placement, egonoise characteristics, platform motion and the intended analysis.

3.7 Theoretical Principles of Audio Zooming

This chapter explores the foundational mathematical equations behind the acoustic properties and spatial concepts that underlie sound propagation and, in the case of this project, the reception of sound by systems such as microphone arrays and the signal-processing techniques used to steer sound and reduce noise. The essential theoretical aspects required to understand why sound does what it does are reviewed comprehensively. Along with other foundational acoustic equations, the introduction of the speed of sound in air is covered. Several advanced concepts for removing noise are presented, such as beamforming separating sources via independent component analysis. Furthermore, spectral subtraction is covered. These principles constitute a sound base for further investigation of the real-world applications of auditory processing systems. In the following equations, t denotes continuous time (acoustic propagation models) and n denotes discrete time (sampled signals used in the implemented algorithms). Scalars are italic, vectors are bold lowercase, matrices are bold uppercase; frequency-domain quantities are complex unless stated.

3.7.1 Acoustic Properties and Spatial Considerations

The formula for the speed of sound in air is given as Equation 3.2. It expresses the speed of sound in that medium as a function of the fundamental physical properties that determine sound propagation in air.

$$c = \sqrt{\frac{\gamma RT}{M}} \tag{3.2}$$

It is known from fundamental physics that sound travels faster in some materials than in others. If any of the physical properties of air were changed, while keeping the others constant, the speed of sound in air would also change. The speed of sound in air, then, is determined by the adiabatic index, γ (which is 1.4 for air), the gas constant R (which has a value of 8.314 J/mol·K in this case), in Kelvin (which can vary considerably), and M, the molar mass of air (which can be taken as 0.029 kg/mol) (Kinsler *et al.*, 2000).

The wavelength of a sound wave, given in Equation 3.3, is determined by, the speed of sound (*c*), and the frequency (*f*).

$$\lambda = \frac{c}{f} \tag{3.3}$$

Specifically, the equation suggests that one can easily calculate the third if one knows two of these quantities.

The sound pressure level (SPL) equation, as seen in Equation 3.4 shows, *SPL* is the sound pressure level denoted in decibels (dB) and is defined as 20 times the logarithm to the base 10 of the ratio of the measured sound pressure (p) to the reference sound pressure, which is 20 microPascals (p_0).

$$SPL = 20 \log_{10}\left(\frac{p}{p_0}\right) \tag{3.4}$$

In simpler terms, this means that the measurement is expressed in Pascals, compared to an established known reference, which is the threshold of human hearing, and then this information is used to express the sound level in a way that is more meaningful to both scientists and laypeople (Kinsler *et al.*, 2000).

The intensity of a sound wave can be described by Equation 3.5, where one can determine the intensity by using the sound pressure and knowing something about the medium through which the sound travels.

$$I = \frac{p^2}{\rho c} \tag{3.5}$$

The medium's density and the speed of sound in that medium are necessary to find the intensity of the sound wave from its pressure. Hence, *I* can be equated to intensity, *p* is the sound pressure, ρ is the air density, and *c* is the speed of sound (Meyer, Neumann and Taylor, 1972).

Equation 3.6 displays the decibel (dB) power ratio formula as follows: L, which stands for level difference, denotes how many decibels one signal is from another and is usually expressed in a logarithmic form. P1 and P2 are the two power levels being compared.

$$L = 10 \log_{10} \left(\frac{P_1}{P_2} \right) \tag{3.6}$$

When working with power levels, it is generally expressed in terms of 10, despite the common binary expression (using log base 2) to describe two signals in terms of power. In short, dB is used for power level comparisons (Kinsler *et al.*, 2000).

The Inverse-square law is stated in Equation 3.7 and can be simply presented as $I = 1/r^2$. In this formula, "*I*" stands for intensity, and "*r*" represents distance. The law expresses a mathematical relationship between intensity and distance.

$$\frac{I_1}{I_2} = \frac{r_2^2}{r_1^2} \tag{3.7}$$

Both values and their ratio have physical significance. Intensity has a straightforward meaning, as it is measuring how much something is occurring in a given way over a given area (Palacios, 1964).

In acoustics, sound propagation can be described in terms of wavelength, propagation speed, and frequency, which together determine the phase relationships observed at spatially separated sensors (Thompson, 1997). In a microphone array, these phase differences mean that the same source arrives at each microphone with a different delay, depending on the source direction and the array geometry. A generic linear beamformer exploits this by applying a set of weights to the microphone signals and summing them to form a single output as in Equation 3.8 (Veen and Buckley, 1988).

$$\mathrm{y}[n] = \sum_{i=1}^{M} w_i x_i[n] \tag{3.8}$$

Equation 3.9 shows acoustic impedance, represented by the symbol Z and defined as the ratio of sound pressure to particle velocity. Sound pressure is denoted by p, and particle velocity is denoted by v.

$$Z = \frac{p}{v} \tag{3.9}$$

The relationship between these fundamental quantities can be expressed as Z = p/v. An analogy may be drawn between a sound wave propagating through a medium and an electric wave propagating through a conductor. Just as the quantity of electric current is related to the potential (or "pressure") of electricity via Ohm's law (Halliday, Resnick and Walker, 2014), the above expression is essentially a statement of Ohm's law for acoustics (Pierce and Beyer, 1989).

The formula for the general Euclidean distance in Equation 3.10 gives the way to determine the distance between two points in n-dimensional space. The formula shows that with two points, x and y, to find the distance between them in Euclidean terms, this formula can be used.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3.10}$$

It essentially states that to find the distance between two points, the squared difference between their corresponding coordinates would be found, and then these differences would be summed, and then the square root would be taken (Cheng, 2024).

The formula for the Euclidean distance in a 3D space, as seen in Equation 3.11, expresses the relationship among the coordinates of several points in a three-dimensional space. In this case, the points are the coordinates of two objects: a sound source and a microphone.

$$D(i, j) = \sqrt{(x_\mathrm{i} - x_\mathrm{j})^2 + (y_\mathrm{i} - y_\mathrm{j})^2 + (z_\mathrm{i} - z_\mathrm{j})^2} \tag{3.11}$$

The coordinates of the sound source and the microphone can be plugged into the equation in place of the "i" and "j" variables to compute the distance between the two in a direct line through space (Strang, 2006).

3.7.2 Theoretical Basis for 3D Scene Creation

Equation 3.12 calculates the azimuth angle for the ith microphone. The angle is wrapped to fall between −180 to 180 degrees, and the dimensionless angle is converted to degrees for use.

$$\text{azimuth}_i = \text{wrapTo180}\left(\arctan\left(\frac{y_i}{x_i}\right) \cdot \frac{180}{\pi} - \text{ref\_azimuth}\right) \quad (3.12)$$

The reference angle is then subtracted to get the angle of the ith microphone relative to that set for alignment. The variables used in the equation have the following meanings: azimuth_i is the calculated azimuth angle for the ith microphone; wrapTo180 is a function that wraps the angle; arctan is the arctangent function, calculating the angle in radians; y_i is the y coordinate of the ith microphone; x_i is the x-coordinate of the i^{th} microphone; $\frac{180}{\pi}$ Converts the angle from radians to degrees, and ref_azimuth is the reference angle to which the equation is aligned (Weisstein, 2024).

Equation 3.13 is used to calculate the elevation of microphones. It works with several factors, including the microphone position, to produce a result.

$$\text{elevation}_i = \text{wrapTo180}\left(\arctan\left(\frac{z_i}{\sqrt{x_i^2 + y_i^2 + z_i^2}}\right) \cdot \frac{180}{\pi}\right) \quad (3.13)$$

The microphone must be in a known position in 3D space to do its job correctly; if the sound is coming from everywhere, it must pick up that sound in a way that the ear, the brain, or some other means can interpret it as the sound coming from a specific direction. To determine that direction, the angles can be expressed in radians. Where azimuth_i is the calculated azimuth angle for microphone *i.,* wrapTo180 is a function that wraps the angle to a range of −180° to 180°. Arctan is again the arctangent function, which calculates the angle in radians. The coordinates of microphone *i* are x_i, y_i, z_i, and $\sqrt{x_i^2 + y_i^2 + z_i^2}$: displays the Euclidean distance from the origin to the microphone *i*., and finally, $\frac{180}{\pi}$: Converts the angle from radians to degrees (MathWorks, 2024).

Equation 3.14 describes the setup of the azimuth grids. These are a set of azimuth angles, the beamforming angles, in a sense; the beamformer uses them to focus on sound sources at a certain angle.

$$\textbf{azimuth\_grids} = [45{,}0, -45{,}90{,}0, -90{,}122.5{,}180, -122.5]^T \quad (3.14)$$

The azimuth_grids are a set of azimuth angles used in beamforming equations, with the values given in degrees (Veen and Buckley, 1988).

3.7.3 Microphone Array Configuration and Beamforming Principles

The time delay between two microphones can be calculated with Equation 3.15, where τ_{ij} represents the time delay between microphones i and j, d_{ij} is the distance between microphones i and j, and c is the speed of sound, approximately 343 m/s in air.

$$\tau_{ij} = \frac{d_{ij}}{c} \quad (3.15)$$

This foundational measurement is critical to determining the delay introduced by the various electronic components of a microphone array. It is, therefore, essential for understanding the corresponding signal alteration (Knapp and Carter, 1976).

Equation 3.16 is the classical delay-and-sum (DAS) beamformer. The delays τ_i are the steering delays applied to each microphone channel so that signals arriving from the chosen look direction are time-aligned before summation. In this work, τ_i is obtained from the propagation time implied by the array geometry and the speed of sound (cf. Eq. (3.2)), and uniform weights $1/N$ are used unless stated otherwise.

$$S_{\text{Out}}(t) = \frac{1}{N}\sum_{i=1}^{N} S_{\text{In},i}(t - \tau_i) \quad (3.16)$$

For signals known as plane waves, there is a simple relationship between the time delay needed to align the signals and the wave's angle of incidence. For more complicated signals or situations where the wavefront is not planar, the problem can quickly become exceedingly complex. Nonetheless, the basic principle of beamforming is given where: $S_{out}(t)$ is the beamformed output signal, $S_{in}(t)$ is the input signal at microphone i, w_i is the weight applied to each microphone, τ_i is the time delay for microphone i, and N is the number of microphones in the array (Bell, Ephraim and Van Trees, 2000).

The MVDR beamformer is described by equation 3.17. $w_{\mathrm{MVDR}}(\Omega)$ is the weight vector for the beamformer. It is a minimum output power weight vector, which means it yields the least power at the array's output when the waveform of interest is not present at the array while also ensuring that the array has a distortionless response in the desired direction. Here, $\mathrm{a}(\Omega)$ is the steering vector for look direction Ω, and $\mathrm{R}_{xx}$is the spatial covariance matrix of the microphone signals. The MVDR weights are then given by

$$\boldsymbol{w}_{\mathrm{MVDR}}(\Omega) = \frac{\boldsymbol{R}_{xx}^{-1}\,\boldsymbol{a}\,(\Omega)}{\boldsymbol{a}^{H}(\Omega)\,\boldsymbol{R}_{xx}^{-1}\boldsymbol{a}(\Omega)}. \tag{3.17}$$

The solution depends on the inverse of $\mathrm{R}_{xx}$and the steering vector $\mathrm{a}(\Omega)$, and uses the Hermitian transpose to enforce the distortionless constraint (Veen and Buckley, 1988).

3.7.4 Noise Reduction Techniques in Array Design

Equation 3.18 expresses the Adaptive microphone array for Noise Reduction (AMNOR) output y(t) as the difference between the directional microphone signal $x_{dir}(t)$ and a weighted sum of the omnidirectional reference signals $r_{\mathrm{m}}(t)$. where $x_{\mathrm{dir}}(t)$is the directional (primary) signal, $r_m(t)$ are the reference channels, $M_r$is the number of reference microphones, and $h_m[l,t]$are time-varying FIR filter coefficients, following the method of Kaneda and Ohga (1986).

$$y(t) = x_{dir}(t) - \sum_{m=1}^{M_r}\sum_{l=0}^{L-1} h_m\,[l,t]r_m(t-l) \tag{3.18}$$

As shown in Equation 3.19, the spectral subtraction process can be summarised in three steps. First, the noisy time domain signal is transformed to the frequency domain to obtain Y(f). Then, the clean and noisy signal power spectrums are compared to estimate the noise power spectrum. Finally, the clean signal and noise estimates are taken, and an inverse operation is performed to yield the reconstructed clean signal.

$$\widehat{P_s}(f) = |Y(f)|^2 - |\widehat{N}(f)|^2 \tag{3.19}$$

Where $\widehat{P_s}(f)$ is the estimated clean signal in the frequency domain, $|Y(f)|^2$ is the power spectrum of the noisy signal, and $\left|\widehat{N}(f)\right|^2$ is the estimated noise power spectrum (Boll, 1979).

The Wiener filter can be described in terms of power spectral density (PSD) relations. Let $H(f,t)$ denote the Wiener filter gain at frequency f and $S(f,\,t)$ and $N(f,t)$ denote the PSDs of the desired

signal and noise, respectively. Then, the relationship between these quantities can be expressed in a deterministic sense using Equation 3.20 (Wiener, 1949)

$$H(f,t) = \frac{S(f,t)}{S(f,t) + N(f,t)} \tag{3.20}$$

The primary components of the Wiener filter's spectral mask are the power spectral densities of the desired signal and the noise. They inform which parts of the spectrum have useful signal energy and which have noise. For an ideal filter with a mask that perfectly delineates the two, then all of the filter's output must have useful energy (Cohen, 2003).

In Equation 3.21, the time-frequency mask is shown as $M(f,t)$. The estimated clean signal in the time-frequency domain, $\hat{X}(f,t)$ is derived from the observed (noisy) signal, $X(f,t)$.

$$\hat{X}(f,t) = M(f,t) \cdot X(f,t) \tag{3.21}$$

The mask is applied such that its values range from 0 to 1, indicating the degree to which the mask allows the signal to pass or not. An ideal mask would yield a clean signal. The process of applying the mask to estimate the clean signal can be described by Equation 3.21 (Boll, 1979), then later (Ephraim and Malah, 1984).

3.7.5 Theoretical Principles of Source Separation

Independent Component Analysis (ICA) for Sound Source Separation separates a mixed signal into its independent source components. In Equation 3.22, the observed mixed signal, X, is represented as a function of the mixing matrix, A, and the independent source components, S.

$$\mathbf{X} = \mathbf{A} \cdot \boldsymbol{S} \tag{3.22}$$

In more practical terms, if an audio signal in a room (where multiple sound sources exist) is "mixed" together in a digital way, then ICA allows you to separate the sound back into its original independent components using some additional "mixing" assumptions about the nature of the sounds (and perhaps the nature of the "room") in which the signals were initially emitted (Hyvärinen and Oja, 2000).

In Equation 3.23, source separation is accomplished using a blind SIMO system. The equation defines several signals of interest. The signal received at each microphone (indexed by i) is represented by $y_i(t)$.

$$y_i(t) = h_i(t) * x(t) + n_i(t) \tag{3.23}$$

The channel's impulse response between the source and the *i*th microphone is represented by $h_i(t)$. The source signal is represented by $x(t)$. Finally, the equation also includes the noise at each microphone, written as $n_i(t)$ (Parra and Spence, 2000).

Equation 3.24 gives the Generalised Cross-Correlation using Phase Transform (GCC-PHAT). In that expression, $G_{xy}(f)$ is the cross-power spectral density of microphone signals *x* and *y*, while τ is the time delay separating their arrivals at the two sensor positions.

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} G_{xy}(f) e^{j2\pi f \tau} df \tag{3.24}$$

Combining the cross-power spectrum with cross-correlation measurements markedly improves DOA estimates (Knapp and Carter, 1976).

Equation 3.25 presents the Griffin–Lim algorithm (Griffin and Lim, 1983), which reconstructs a time-domain waveform from its magnitude spectrogram. The magnitude spectrogram is usually obtained by applying a Short-Time Fourier Transform (STFT) to the original signal. At each iteration, the algorithm refines the missing phase by re-estimating it from the previous reconstruction and then inverts the transform to return to the time domain. Such phase interpolation is widely used in audio restoration and speech-synthesis tasks.

$$x^{(k+1)}(t) = ISTFT\left(|X(\omega, t)| e^{j\theta^{(k)}(\omega,t)}\right) \tag{3.25}$$

3.7.6 Summary of Theory

This chapter reviewed the theoretical tools that underpin modern audio processing, starting with basic sound-propagation relations and acoustic impedance and moving to advanced concepts such as beamforming, noise reduction, and Independent Component Analysis (ICA). Together, these formulas illustrate how physics, mathematics, and engineering combine to handle complex, auditory scenes, precisely the situations that challenge the audio algorithms discussed earlier in Chapter 2. The same principles form the computational core of the MATLAB simulator introduced in Section 5.3.

3.8 Summary

This chapter has reviewed the main strands of work relevant to audio zooming for forensic surveillance: audio zoom and the Cocktail Party Problem, sound source localisation, source separation and noise reduction. The literature demonstrates that there are many mature techniques for directional capture and speech enhancement, and that recent machine-learning approaches can deliver strong separation and denoising performance under controlled conditions. The review also highlights several gaps when these methods are viewed through the lens of creating a lightweight, drone-mounted microphone array operating in a real-world urban environment and subject to evidential constraints.

Across all strands of the literature, a pattern emerges. Classical statistical signal-processing methods such as MVDR beamforming, MUSIC, ICA and multichannel Wiener filtering provide physically interpretable baselines that are still widely used in practice. Machine-learning-based methods have advanced the field by delivering stronger separation, denoising and localisation performance on benchmark tasks and by coping better with highly complex mixtures. At the same time, they introduce new challenges in terms of training data requirements, computational cost, robustness to distribution shifts and the difficulty of explaining or reproducing their behaviour in a forensic setting. The approach taken in this thesis is therefore to build on the insights and performance gains of the learning-based literature, while centring the proposed system on transparent, geometry-aware beamforming and post-filtering that can be implemented and audited in real-world police-drone deployments.

The key research gaps identified are:

• Drone-specific audio zooming systems: Relatively few studies design and evaluate complete audio-zooming pipelines for drone-mounted arrays in noisy outdoor scenes, particularly in policing or forensic contexts.

• Integrated array design, localisation and beamforming under UAV constraints: There is limited guidance on how to select array geometries, sensor types and beamforming strategies that respect weight, power and motion constraints while still delivering useful directional gain on a moving drone platform.

• Noise reduction for weak, far-field speech: Existing noise-reduction methods are often tested on higher-SNR or laboratory data, with less emphasis on weak, far-field speech masked by crowd, traffic and rotor noise as in the Ava White case style scenario.

• Forensic-grade, reproducible processing chains: Many learning-based enhancement and separation methods are difficult to audit and reproduce in a legal setting, and there is little work on audio-zoom systems that explicitly support evidential logging, transparency and repeatability.

The remainder of the thesis addresses these gaps as follows. Chapter 4 designs and characterises microphone arrays, custom hardware and controlled booth experiments under realistic weight and deployment constraints, addressing Gaps 1 and 2. Chapter 5 develops a MATLAB-based simulation framework that embeds array geometry, beamforming and noise-reduction theory into configurable 3D scenes, addressing Gaps 2 and 4. Chapter 6 extends this framework to the Exemplar Houses and drone-relevant field scenarios, linking simulated and real-world results to the kind of urban environment seen in the Ava White case, further addressing Gaps 1–3. Chapter 7 then quantifies the performance of the proposed audio-zooming and noise-reduction chain using SNR and intelligibility metrics, demonstrating how the system can improve the quality and usefulness of audio evidence captured from drone-mounted arrays.

Chapter 4: Initial Experimental Approach

4.1 Overview

This chapter presents the initial experimental approach used to establish a controlled baseline for the proposed audio-zoom system in a professional sound-booth environment. The main contribution of this chapter is a reproducible experimental framework and multichannel dataset for evaluating drone-motivated audio zooming under practical constraints (limited channel count, low-mass sensors, repeatable geometry, and controllable playback conditions). Standard acoustic concepts and manufacturer specifications are included only to justify design choices; the novel element is the end-to-end measurement workflow, including custom playback hardware, repeatable calibration procedures, and the progression from a four-sensor prototype to a sixteen-sensor low-mass array that is carried forward into later chapters.

Section 4.2 details the equipment, signal chain and physical layouts used for multi-loudspeaker playback and synchronous multichannel recording. Section 4.3 introduces the problem formulation and then describes the core studio experiments, including setup, calibration and controlled playback conditions, and defines the datasets used in later analysis. Section 4.4 reports additional experiments that refine the initial setup and extend the recordings under further conditions. Section 4.5 summarises the key outcomes and explicitly links the resulting measurements and datasets to the simulation and field evaluations presented in subsequent chapters. The outputs of this chapter are calibrated multichannel recordings, validated array/build choices under UAV constraints, and the parameter values and procedures used to configure and validate the later simulations and evaluations. The results are analysed and reported in Section 7.2.

4.2 Equipment Used

The experimental work relied on carefully selected equipment to ensure accurate audio capture and reliable system performance. The necessary equipment included:

- A multi-channel digital audio recorder that is capable of 24-bit audio capture at a 48kHz sampling rate.
- A microphone array: wooden baffle with 25 potential mounting positions, of which 16 omnidirectional microphones were populated and wired, arranged in a calibrated array for spatial audio capture. The choice of 16 active microphones was constrained by the

number of available XLR inputs in the studio; the remaining positions were left unpopulated to allow for future reconfiguration.

- Preamplifiers and Audio Interfaces: High-quality interfaces (e.g., Focusrite 18i20, Tascam 24) ensuring minimal signal distortion.
- Computer: High-performance computer running MATLAB and a DAW for playback, simulation, real-time processing, and data analysis.
- Acoustically Treated Environment: A recording studio designed to minimise noise, with provisions for these realistic field tests and simulations.

Table 4.1 summarises the hardware used in the initial experimental sessions. Each specification is chosen to meet the minimum signal-quality and synchronisation constraints imposed by audio processing algorithms.

Table 4.1. List of Equipment and justification of choices.

Equipment	Specification	Rationale	Model
Audio Interface	24-bit / 48kHz, 16+ tracks	24-bit gives 144dB dynamic range headroom for loud speech bursts and low-level ambience without clipping or dithering artefacts. 48kHz meets the Nyquist criterion for all speech-band (> 0–22kHz) content and is the de-facto standard in pro DAWs, avoiding resampling steps.	Antelope Orion 32 +
Microphone Array	4 -16 omni capsules, calibrated positions (±2 mm)	16 channels let you realise a 3-D cardioid/beamformer with sufficient spatial degrees-of-freedom to test zoom algorithms; matches the 16 XLR inputs available, so no sub-banking. Omnis are phase-coherent and flat below 100Hz, which is important for array processing.	Custom
Mixer	EIN ≤ –128dBu, THD < 0.001 %, 16 balanced inputs	Low Equivalent-Input-Noise keeps pre-amp hiss below room noise, so SNR is dominated by scene content, not hardware. Balanced lines reject studio ground loops. Interfaces like 18i20 offer word-clock sync to align all channel samples accurately.	Audient ASP 4816

Equipment	Specification	Rationale	Model
Computer	≥ 8-core CPU, 32GB RAM, NVMe SSD	Real-time array processing in MATLAB plus DAW playback can peak at > 400 GFLOPS. 32GB buffers the 16-track WAV stems (~ 600 MB/min) and lets you run large STFT windows without disk thrash.	Mac Studio
Acoustic space	RT60 < 0.3 s, NC ≤ 20dB	Short RT60 keeps early reflections from masking direct-path localisation cues; NC-20 ensures self-noise sits 15dB below soft speech (~45dB SPL), so captured noise floor is scene-limited.	Treated isolation booth

4.2.1 Microphones

Multiple types of microphones were employed throughout these experiments to ensure comprehensive audio capture under diverse acoustic conditions and to facilitate the spatial analysis essential for audio zooming. The microphone selection process considered factors such as polar patterns, frequency responses, sensitivity, portability, and practical deployment on drones. The following types were specifically chosen:

- Ambisonic Microphones: Zoom H3-VR ambisonic microphones capture full-sphere surround audio, offering a detailed three-dimensional acoustic environment useful for spatial audio analysis and providing realistic auditory scene reproduction (Zoom Corporation, 2022).
- Shotgun Microphones: Rode® NTG-2 directional condenser microphones, featuring highly directional polar patterns, shotgun microphones were used primarily to isolate sounds from a particular direction or source at a distance, thus effectively testing the directional capabilities of the beamforming algorithms in real-world conditions (Rode, 2025b).
- Lavalier Microphones: (Ruochen, Yuhong and Wei, 2014)Rode Lavalier microphones. Small and discreet, these microphones enabled the close capturing of speech and detailed audio signals from specific targets, useful for evaluating the effectiveness of source isolation techniques and post-processing clarity (Rode, 2025a).

- Omnidirectional Microphone Capsules: MCE-400 electret microphone cartridges. These provided uniform sound capture from all directions, supporting comprehensive spatial audio acquisition. Owing to their low cost, small size, and weight, these were particularly suited for drone deployment, making them an integral part of the microphone array design (CPC, 2025b).

The combination of microphones available during the initial experimental phase allowed for robust data collection and analysis, and helped to characterise how different capsule types respond under forensic surveillance conditions. Table 4.2 lists all microphones that were available and evaluated during this phase; individual experiments used specific subsets of these sensors, rather than all microphones simultaneously. Table 4.3 justifies each sensor's role in the project.

Table 4.2. Microphones available and evaluated during the initial experiments

Microphone Model	Polar Pattern	Type	Amount
Zoom Ambisonic H3 VR	Cardioid (× 4 matched heads)	Condenser	4
Rode NTG-2	Super Cardioid	Condenser	4
Rode Lavalier	Omnidirectional	Condenser	4
CPC Omnidirectional Electret	Omnidirectional	Electret	25
Shure SM57	Cardioid	Dynamic	4

Table 4.3. Comparative summary of microphone types and relevance to drone audio zooming

Microphone	Role	Advantage	Limitation	Practicality
Zoom Ambisonic H3 VR	Capturing a 3D sound field for reference	Provides full-sphere capture useful for qualitative spatial analysis	Additional payload, processing, and calibration complexity	Low
Rode NTG-2	Directional baseline	Strong directivity provides an intuitive comparison point for array steering/beamforming	Requires power, bulkier mounting, wind susceptibility and weight.	Medium-low
Rode Lavalier	Close-speech reference	High SNR near-field speech reference helps verify intelligibility and processing behaviour	Not applicable to standoff surveillance; not a realistic drone payload for remote capture	Low
CPC Omnidirectional Electret	Candidate array sensor for UAV deployment	Low mass, low cost, uniform directivity; supports consistent multi-channel array construction	Requires careful matching/calibration; susceptible to wind without shielding	High
Shure SM57	Rugged baseline microphone for comparison	Robust, no power required, predictable behaviour	Heavy relative to capsules; poor choice for multi-channel UAV payload	Low-medium

In terms of practical deployment considerations, while multiple microphone types were evaluated during the initial phase, a mixed-microphone payload is not proposed as a practical end-state for a police drone. In real deployment, heterogeneous microphones introduce additional mass and cabling, increase aerodynamic drag and wind susceptibility, and can reduce flight endurance. They also complicate the power budget (e.g., condenser microphones requiring bias/phantom power) and increase calibration complexity because channels may differ in sensitivity, self-noise, frequency response and phase, which can bias spatial processing unless carefully equalised and level-matched. For these reasons, the mixed-microphone tests in this chapter are used primarily to benchmark directional characteristics and recording quality under controlled conditions and to inform sensor selection. The array configuration taken forward for the main audio-zoom experiments therefore prioritises a uniform set of low-mass omnidirectional capsules; specifically, the CPC electret cartridges are used to populate a 16-microphone circular array matched to the available 16 XLR inputs, providing consistent channel characteristics and a more realistic pathway to UAV deployment.

The Zoom H3-VR, Rode NTG-2, Rode lavalier and Shure SM57 microphones were used in small test configurations to compare directional characteristics and recording quality. The CPC omnidirectional electret capsules were later mounted on the circular wooden baffle, and only 16 of these capsules were populated and wired to individual channels for the main array experiments, matching the 16 available XLR inputs. Thus, all key results presented in Chapters 5–7 are based on a 16-microphone array, not a 41-microphone system.

The microphone types listed in Tables 4.2–4.3 were evaluated in small test configurations to benchmark directional behaviour, noise floor and capture quality under controlled conditions. These comparisons informed sensor selection, but the deployable pathway prioritised a uniform, low-mass capsule array. For this reason, the main multichannel datasets used in Chapters 5–7 are based on 16 CPC electret capsules as shown in Figure 4.1, each wired to an individual channel, mounted on the circular baffle Additional reference microphone specifications and manufacturer response plots are provided in Appendix C.

Figure 4.1. CPC Omnidirectional electret capsule MCE-400. (Taken from CPC (2025b)).

4.3 Studio Experiments in Sound Booth Environment

To establish an acoustic reference for the beamforming experiments that would follow, a series of baseline tests was run in a purpose-built, reflection-controlled sound booth. The booth measures 4.56 m × 3.42 m × 2.24 m and contains two glass windows (0.86 × 0.64 m each) surrounded by an acoustic treatment package.

Thirty absorber panels (0.91 × 0.91 m, 50 mm pyramid foam) cover ~70 % of the combined wall-and-ceiling area. One-third-octave sine-sweep measurements (ISO 3382) gave an RT_{60} of 140 ± 10 ms above 250Hz**,** rising smoothly to 190 ms at 125Hz. The floor was left untreated, apart from a regular carpet. This helps retain low-frequency energy for array calibration. A Brüel & Kjær 2236 sound pressure level meter, A-weighted, was placed at the array position and logged for 15 minutes with the ventilation fan cycling normally. Table 4.4 summarises background noise measurements in the sound booth. The Aeq,15 min values represent the A-weighted equivalent continuous sound pressure level over a 15-minutes.

Table 4.4. Background Noise Benchmark in Sound Booth.

Condition	Aeq 15 mins [dB(A)]	NC Curve	Dominant bands
Fan off	24dB(A)	NC-10	160–200Hz hum at 27dB
Fan on	31dB(A)	NC-15	+5dB boost from 63Hz–1kHz

Both states are comfortably below the NC-20 criterion commonly adopted for critical listening spaces (Sylvestre-Williams, 2020), confirming that low-level algorithm tests would be limited to microphones rather than room noise. All eight loudspeakers were cabled with 4.5m low-capacitance twin leads and soldered connectors to minimise phase skew between channels. Figure 4.6 shows a CAD drawing of the sound booth setup with the four-sensor array and nine speakers, illustrating the options for different geometric layouts of the speakers and the resulting fundamental direction of sonic travel.

The studio experiments are presented in the same order as the practical workflow used to generate the datasets. First, the problem is formulated mathematically in section 4.3.1 to define the input–output relationship and processing. The experimental method then proceeds from system setup and calibration to controlled signal playback and multichannel recording, and finally to structured variations such as distance, interference, and wind/absorption conditions. This ordering allows later chapters to link measured results directly to the modelling and beamforming framework.

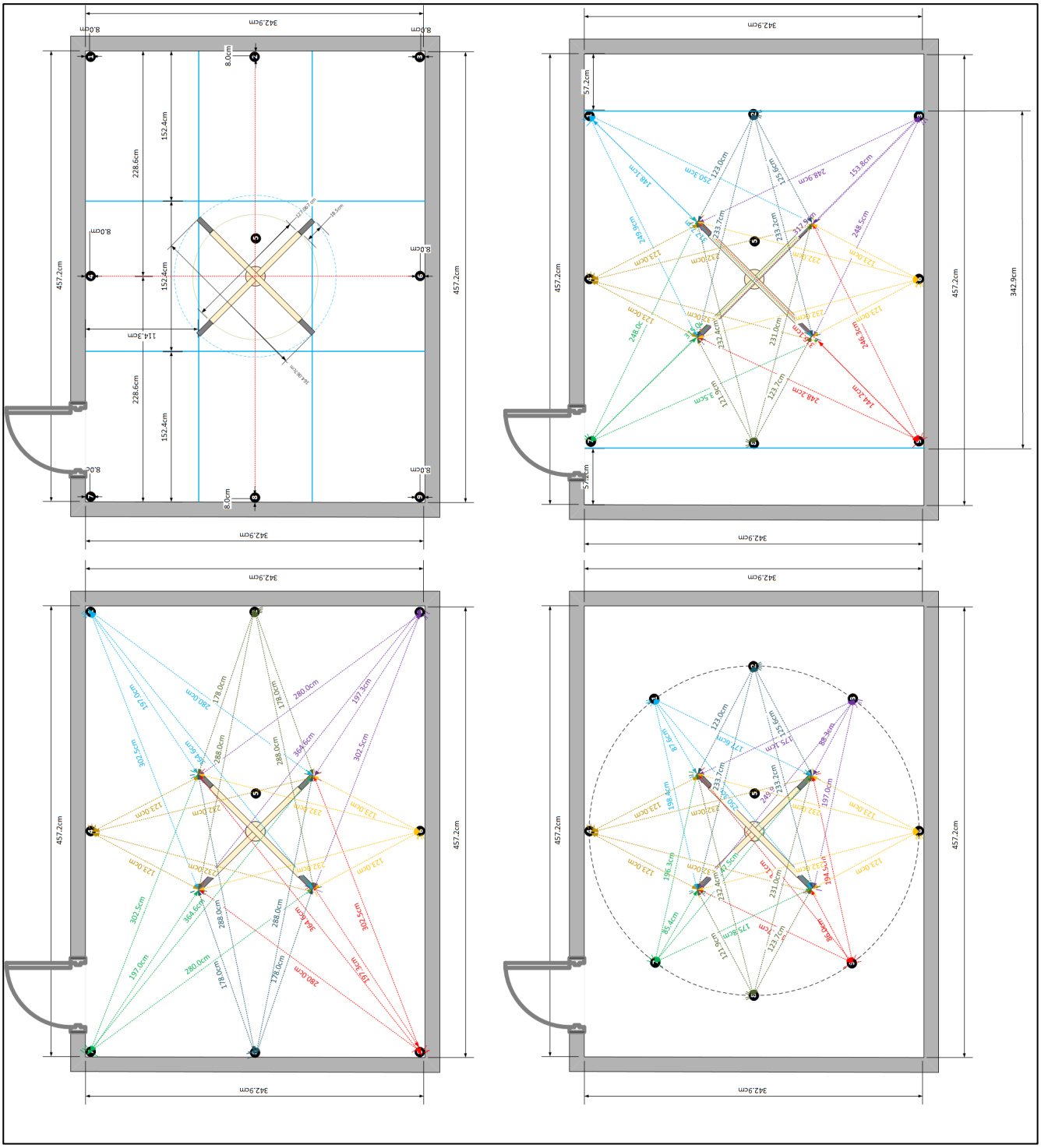

Figure 4.2. Illustration of multiple studio setups, showing simulated microphone and source placements.

4.3.1 Problem Formulation

This section approaches audio zooming for surveillance as a multichannel array signal-processing problem. The aim is to recover an intelligible result from a target source from within a noisy acoustic scene using a compact microphone array and a reproducible processing chain, under practical constraints relevant to drone deployment. The studio experiments in Sections 4.3 and 4.4 establish controlled multichannel recordings that match the modelling assumptions used later for beamforming and post-filtering.

The mathematical formulations that follow aim to give a better understanding of the problem being addressed.

An array of M microphones recording discrete-time signals and then collecting them into a vector.

$$\mathrm{x}[n] = [x_1[n] \quad x_2[n] \quad \dots \quad x_M[n]]^T. \tag{4.1}$$

A target source $s[n]$ arriving from a direction of interest Ω, where $\mathrm{a}(\Omega)$ is the array steering vector and $\mathrm{u}[n]$ represents interference and noise.

$$\mathrm{x}[n] = \mathrm{a}(\Omega)\, s[n] + \mathrm{u}[n], \tag{4.2}$$

In the short-time Fourier transform (STFT) domain, the multichannel mixture is $\mathrm{X}(k, \ell)$. A linear beamformer produces an audio-zoom output.

$$Y(k, \ell) = \mathrm{w}^H(k)\, \mathrm{X}(k, \ell), \tag{4.3}$$

$\mathrm{w}(k)$are frequency-dependent complex weights. In this thesis, the primary spatial filter is based on the minimum variance distortionless response (MVDR) criterion (Capon, 1969), which preserves the target direction while minimising output power:

$$\min_{\mathrm{w}(k)} \quad \mathrm{w}^H(k)\, \mathrm{R_u}(k)\, \mathrm{w}(k) \text{s.t.} \mathrm{w}^H(k)\, \mathrm{a}(k, \Omega) = 1, \tag{4.4}$$

$\mathrm{R_u}(k)$is the spatial covariance matrix of the interference-plus-noise. The closed-form solution is

$$\mathrm{w}_{\mathrm{MVDR}}(k) = \frac{\mathrm{R}_{\mathrm{u}}^{-1}(k)\,\mathrm{a}(k,\Omega)}{\mathrm{a}^{H}(k,\Omega)\,\mathrm{R}_{\mathrm{u}}^{-1}(k)\,\mathrm{a}(k,\Omega)}. \tag{4.5}$$

To further suppress residual interference after beamforming, a post-filter $G(k,\ell)$can be applied. $G(k,\ell)$ can be derived from an estimate of the SNR. The time-domain output $\hat{s}[n]$ is obtained via inverse STFT.

$$\hat{S}(k,\ell) = G(k,\ell)\,Y(k,\ell), \tag{4.6}$$

The experiments that follow are designed to validate the practical assumptions behind these formulations.

4.3.2 Amplifier Build

A custom-built 9-channel audio amplifier was constructed specifically for use in the experimental phase of the research, enabling simultaneous delivery of multiple audio signals to distinct speaker outputs within the sound booth. The amplifier design focused on achieving low distortion and a high-power output tailored to the project's specific audio reproduction requirements. Components were carefully selected based on targeted gain specifications and desired frequency response characteristics. Repeated bench testing and thermal monitoring accompanied every stage of construction to guarantee that the amplifier would run reliably during lengthy experiments. The development process began with a stripped-down two-channel circuit to confirm core operation, then expanded to the full nine-channel design once the initial checks had been passed. All semiconductors, passives, and ancillary parts were ordered from specialist suppliers, most notably CPC Farnell.

The assembly involved PCB-mounted 6.35mm stereo jack sockets, WHADDA WSAH4001 mono amplifier kits, a stable IDEAL POWER 12V, 5A desktop power supply, PCB terminal blocks, solderless breadboards with jumper wires, and compact 3-inch, 8Ω speakers. Additional hardware such as a power-distribution board, locking DC sockets, and a rugged, plastic 19-inch rack case, protected the finished electronics. Figure 4.3 shows the single-channel amplifier circuit replicated across all nine outputs. Each module was built and verified first in a two-channel prototype to confirm stability, noise floor and thermal behaviour before scaling to the full unit.

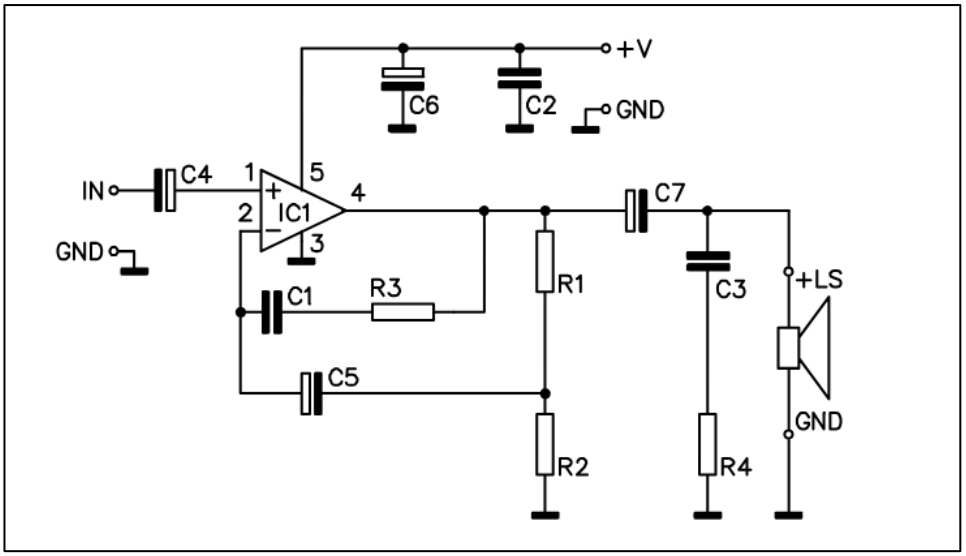

Figure 4.3. Single-channel amplifier circuit (Taken from CPC (2025a)).

After the two-channel prototype had met its performance targets, the nine-channel amplifier was assembled inside the rack. Each channel used its own amplifier module and dedicated wiring loom, keeping the signal paths isolated from one another. Careful attention was given to internal wiring and soldering practices to guarantee electrical reliability. This approach ensured each speaker received a clear, isolated audio signal, crucial for precise acoustic measurements during testing.

All channels demonstrated effective functionality, achieving robust output levels, as shown in Table 4.5.

Table 4.5 Output levels of the 9-channel amplifier.

Test	Method	Result
Max. continuous power	1kHz sine, 8Ω, THD ≤ 1 %	2.5W rms per channel
THD +N@ 1W	1kHz, 8Ω	0.12%
Signal to noise (A)	20Hz-20kHz, ref. 1W	84dB
Frequency response	20Hz-20kHz sweep, ref. 1W	+0.4
Channel crosstalk	1kHz, drive adjacent channel	-55dB
Idle noise (all channels active)	Microphone at 1m	29dB(A)

A low-level hiss was detected when all nine channels were operated simultaneously. This noise was primarily attributed to cumulative amplifier circuit noise and potential electromagnetic interference due to the proximity of internal components. Future versions of the amplifier could incorporate additional shielding and enhanced power filtering techniques to further improve audio signal clarity. Figure 4.4 shows the assembled nine-channel amplifier interfaced to the multichannel audio interface using ¼-inch TRS patch leads, providing one discrete playback feed per loudspeaker channel. Figure 4.4 also documents the internal wiring layout and channel mapping used to keep signal paths separated and to minimise crosstalk. During all multichannel loudspeaker tests, each speaker was driven from a dedicated amplifier channel so that individual sources could be activated, muted or level-matched without affecting adjacent outputs. This architecture is central to replication because it defines the exact routing from DAW buses to physical loudspeaker positions, and ensures that any measured differences in array response arise from geometry and processing rather than accidental summing or shared wiring.

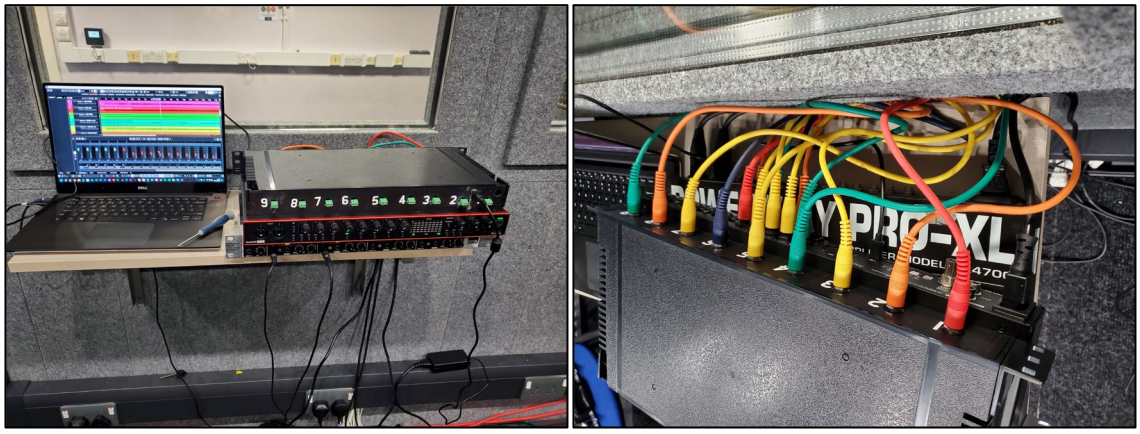

Figure 4.4. Photographs of the nine-channel amplifier connected to the audio interface.

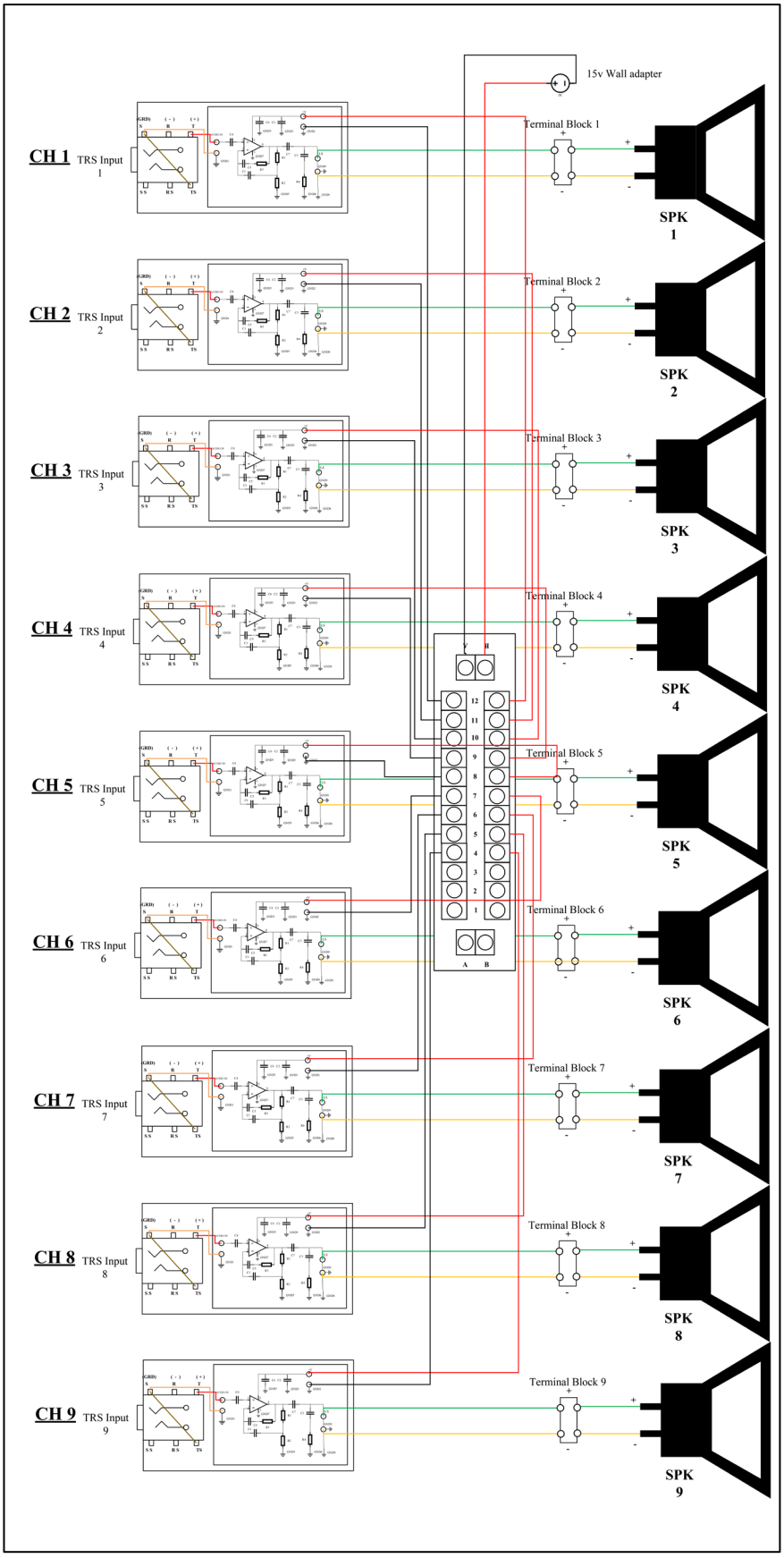

Figure 4.5. The Nine Channel Amplifier wiring diagram.

4.3.3 Speakers

Constructing the speakers was another critical element of the experimental setup. The speakers were required to have a flat frequency response and reproduce sounds from within the 100Hz to 15 kHz spectrum. The chosen driver units were Visaton FR10 HM 3-inch drivers, selected specifically for their reasonably flat frequency response of 88dB SPL/1 W/1 m and within ±4dB from 150Hz to 12kHz with a full range of 95Hz to 22kHz (Visaton GmbH & Co., 2023), as illustrated in Figure 4.6. Figure 4.6 justifies the selection of the Visaton FR10 HM drivers by showing that their response is sufficiently flat across the main speech band for controlled playback. Using a consistent driver model across all loudspeaker positions reduces source-dependent coloration, which is important when later interpreting beamforming performance and intelligibility metrics. These speakers were lightweight and portable, which was essential for the multi speaker experimental setups and would deliver accurate audio playback during the experiments.

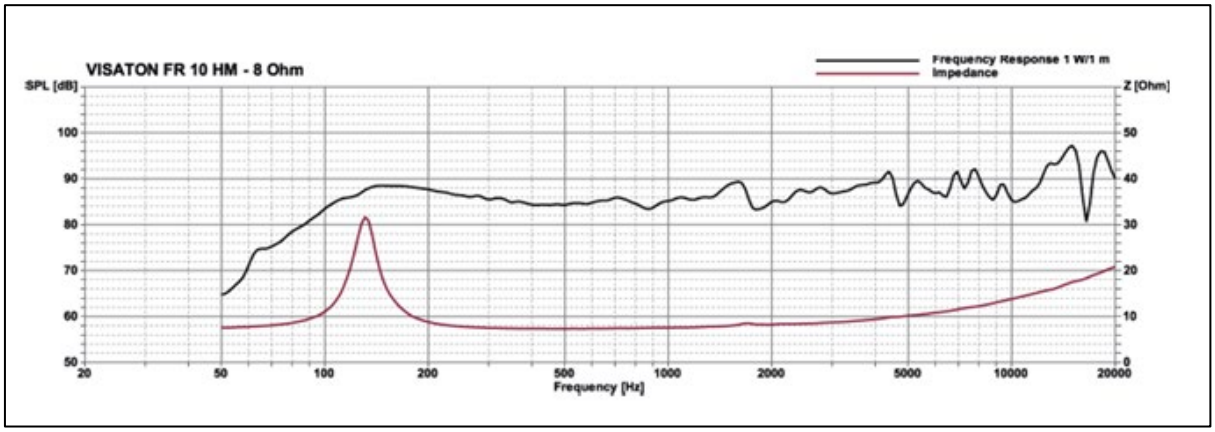

Figure 4.6. Frequency response of the Vistaton FR 10 HM – 8Ω Speaker. (Taken from Visaton GmbH & Co. (2023)).

Enclosures were designed to optimise acoustic performance and minimise unwanted resonance, ensuring consistent and reliable sound reproduction. Although full-range drivers were employed, attention was given to detailed characterisation of speaker performance through impedance, sensitivity, and frequency response measurements to validate their suitability for the experimental objectives.

Speaker leads required significant extensions to reach the desired positions within the sound booth, each extended to over 15 feet. This was achieved by manually joining additional cable lengths using solder connections and protective heat shrink tubing, a method chosen for its

reliability and signal integrity. All 4.6 m extensions used 2 × 1 mm² oxygen-free copper twisted pair to keep loop resistance below 0.3 Ω, thereby minimising voltage drop and preserving amplifier damping factor. Care was taken to ensure correct polarity during connections, with the black-marked lead consistently routed to the negative terminal on each speaker and connected to the left slot of the green terminal blocks on the amplifier.

Upon testing with all nine speakers simultaneously active, a noticeable amount of background hiss was present. This issue, similar to the amplifier noise previously described, was likely caused by cumulative circuit noise or electromagnetic interference from closely grouped amplifier channels. Despite this, the overall speaker system provided a stable and uniform audio playback environment, which would be important for trustworthy results and allow for an accurate evaluation of the audio zooming algorithms.

4.3.4 United Kingdom Police Drone Research

The aim of this research was to create a drone-mounted microphone array that could be carried as a payload by a commercial drone. The research utilised the DJI Matrice 300 RTK drone platform, a robust and heavy-duty aircraft frequently employed by United Kingdom Police forces, such as North Wales Police. The drone, measuring 810 mm in length, 670 mm in width, and 430 mm in height (excluding propellers), features a diagonal wheelbase of 895 mm. The M300 RTK supports a substantial maximum payload of 2.7 kg and offers an operational flight time of up to 31 minutes when fully loaded, ensuring ample time for extended surveillance operations (DJI, 2024).

The acoustic array must ride on the DJI Matrice 300 RTK without degrading flight safety, endurance, or image-stabiliser clearance. Table 4.6 shows the hard limits derived from the airframe manual and in-house weight tests.

Table 4.6. Police drone dimensions and limits (DJI, 2024).

Parameter	DJI Matrice 300 Spec	Design target for the array and mount
Spare payload mass	2.7 kg(max) – 1.95 kg (H20T camera) = 0.75 kg	≤ 0.60 kg (allows 0.15 kg margin)
Overall Footprint	810 x 670 mm	Tip to Tip

Parameter	DJI Matrice 300 Spec	Design target for the array and mount
Landing Gear Footprint	310 x 260 mm	Array ring Ø 300 mm (fits inside)
Ground clearance	130 mm	0 ± 3 mm after mount tuning

The M300 RTK integrates seamlessly with a variety of high-performance camera systems, including DJI's own H20 series and the specialised L1 and P1 surveying models. Additionally, the drone can support third-party payloads such as the Wingsland Z15 bright spotlight for enhanced night-time operations, the U10 methane detector for hazardous environment assessments, and the FLIR Vue TZ20 dual thermal zoom camera, which provides versatile imaging capabilities crucial for law enforcement applications.

For scenarios requiring portability and rapid deployment, a secondary option the DJI Mavic 2 Enterprise Advanced was also considered. Significantly smaller at dimensions of 322 mm by 242 mm by 84 mm when unfolded, this drone model offers a lightweight and easily transportable solution while retaining advanced imaging capabilities. The obvious downside would be that the payload limit would be reduced from 2.7 kg for the M300 to only 190g (0.19 kg) for the DJI Mavic 2 Enterprise Advanced. The reduced capacity would make the carrying of a full-sized microphone array challenging.

Detailed technical literature and additional insights into the drone platforms were provided directly by Heliguy, a specialist supplier, facilitating informed decision-making and comprehensive equipment assessment (Heliguy Ltd, 2025).

An accurate beamformer evaluation demands that the sonic profile and motor harmonics of the host aircraft be present in the virtual simulation. Therefore, a clean noise-signature was obtained from audio recordings online with the following setup: M300 RTK hover at 15m AGL, no payload, calm evening (LAeq = 26dB(A) at mic). A Brüel & Kjær 4193 half-inch reference microphone was placed on a 4 m pole directly beneath the hover point; 96kHz / 24-bit capture on Sound Devices MixPre-6 II.

The drone-noise profile used in this study was obtained as a baseline acoustic signature for the DJI Matrice 300 RTK during steady hover with no external payload (Ramirez, 2021). This choice

provides a repeatable reference condition for capturing the dominant tonal components (motor harmonics/blade-passing-related tones) and the broadband rotor noise bed that drives the low-SNR problem. It is acknowledged that the noise spectrum can change when payloads are attached because added mass and aerodynamic drag alter rotor speed control and loading. To account for this, the simulation treats the Ramirez recording as the minimum-noise case and evaluates robustness by applying a conservative noise margin (overall level increase) and by allowing small shifts in the dominant harmonic structure during analysis, so that the beamforming and post-filter stages are not tuned to a single, idealised spectrum. In practice, this means the proposed processing is assessed against rotor noise conditions that are equal to or worse than the no-payload baseline.

4.3.5 Sensor Array Build (4 Microphones)

A four-microphone plywood frame was built to prove the spacing strategy before committing to the 16-channel flight array. The mics sit 70 mm apart close enough to avoid spatial aliasing across the speech band (0.5–4kHz) yet compact enough to hang under a Police drone. Each position is a 10 mm laser slot that takes a standard 3⁄8-in bolt, so off-the-shelf shock mounts drop straight in. A cold-shoe glued at the centre lets the rig snap onto a tripod or the Matrice landing cage. Two laser-cut versions were made from FSC birch: a 9mm plate (485g) that proved too heavy, and a 6mm plate (325g) that met the payload limit without altering the 70mm geometry. CAD-to-laser workflow kept manufacturing error below 0.1mm.

Bench tests uncovered another compromise. Full-size shotgun mics, chosen for their 12dBA self-noise, bent the thin plate by 1.5mm at the tips inside the ±2mm tolerance but risky for repeatable flight data. Swapping to 22g MEMS capsules removed the sag and, according to FE analysis, lifted the array's first resonance an octave clear of rotor harmonics. Plywood is therefore good enough for laboratory work, but it can bend once heavier equipment goes on. Future builds could use a material such as carbon fibre to reduce weight and boost stiffness, or alternatively reduce the weight of the sensors themselves. Figure 4.7 places the four-microphone prototype array within the physical envelope of the DJI Matrice 300 platform so the reader can see the intended mounting scale and clearance constraints. The array geometry (70 mm spacing) was fixed at design time and then held constant across the early trials to isolate the effect of signal-processing choices from mechanical changes. The array centre was treated as the reference point for all distance and steering calculations in the subsequent analysis, and the array orientation was kept consistent relative to the loudspeaker grid during booth captures.

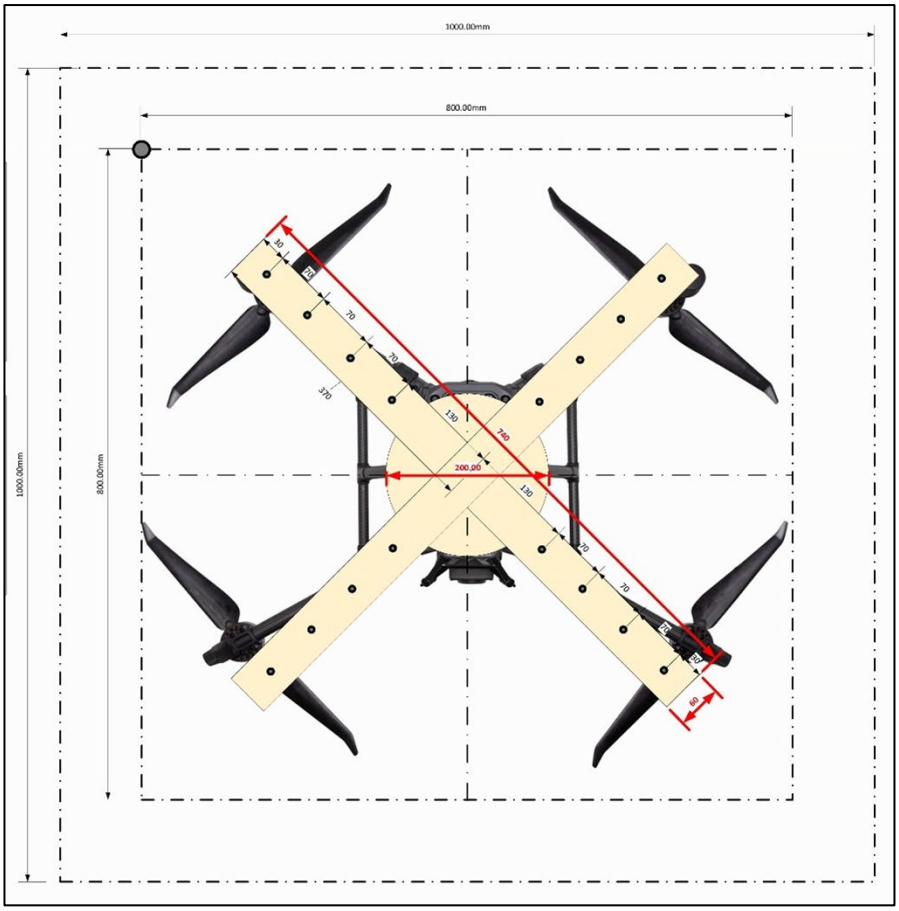

Figure 4.7. Array design, modelled on the dimensions of a United Kingdom Police Drone.

Figure 4.8 shows the realised laser-cut array mounted on a stand with the shoe connector and 3/8″ adapters used to achieve repeatable positioning. In practice, the stand height and array orientation were fixed for a given session so that all recordings share a common coordinate frame. This reduces between-trial variability and allows the multichannel recordings to be compared directly under changes in source position or playback content.

Figure 4.8. Laser Cut Physical Microphone Array

The drone airframe can scatter sound through reflection and diffraction, altering the amplitude and phase at each microphone relative to a free-field situation. This is a known concern for body-mounted arrays, where curved/edged structures introduce propagation-path perturbations that can bias direction-of-arrival estimation and beamformer steering (Guo, 2024)

In this experiment, the array has an externally mounted payload rather than integrated sensors. The intended mounting strategy places the microphone ring below the main body, so that ground-originating sound arrives with minimal occlusion and with more symmetric geometry around the array. This approach also reduces asymmetry caused by panned mounting and helps keep any residual scattering minimal across the microphones, which in turn, is less disruptive to beamforming. Related drone-audition systems likewise emphasise that microphone placement relative to propellers and the body materially affects recorded mixtures and enhancement performance, highlighting the importance of careful placement and robustness testing (Clayton *et al.*, 2023)

The booth experiments characterise the array and processing chain without airframe scattering; airframe effects are treated as a deployment consideration and robustness factor rather than being embedded in the experiment's geometry.

4.3.6 Sensitivity and Frequency Sweep Tests

Audio tests were designed to investigate the sensitivity and frequency response of the prototype microphone array. Initially, tones at predetermined frequencies over two musical octaves were generated and passed through the audio chain. The initial test was a speech band-limited check. Pure tones spanning two octaves (C4 to B5, 261.63Hz–493.88Hz) were injected at –12dBFS to confirm that the signal chain was recording as expected.
Then, a full-frequency sweep (20Hz to 20kHz) was performed. Analysis would focus on measuring harmonic distortion levels and amplitude consistency across the frequency range. This helps with establishing baseline performance levels for when more complex tests were performed. Table 4.7 shows the frequencies and associated harmonics generated in the sensitivity tests.

Table 4.7. Frequencies and harmonics of the sensitivity tests.

Fundamental f_1	$2\times f_1$ (2nd harm.)	$3\times f_1$ (3rd harm.)	Notes
20Hz	40Hz	60Hz	LF limit
25Hz	50Hz	75Hz	
31.5Hz	63Hz	94.5Hz	
40Hz	80Hz	120Hz	
50Hz	100Hz	150Hz	
63Hz	126Hz	189Hz	
80Hz	160Hz	240Hz	
100Hz	200Hz	300Hz	
125Hz	250Hz	375Hz	
160Hz	320Hz	480Hz	
200Hz	400Hz	600Hz	
250Hz	500Hz	750Hz	
315Hz	630Hz	945Hz	
400Hz	800Hz	1.2kHz	
500Hz	1kHz	1.5kHz	
630Hz	1.26kHz	1.89kHz	

Fundamental f_1	2× f_1 (2nd harm.)	3× f_1 (3rd harm.)	Notes
800Hz	1.6kHz	2.4kHz	
1kHz	2kHz	3kHz	Speech upper octave
1.25kHz	2.5kHz	3.75kHz	
1.6kHz	3.2kHz	4.8kHz	
2kHz	4kHz	6kHz	
2.5kHz	5kHz	7.5kHz	
3.15kHz	6.3kHz	9.45kHz	
4kHz	8kHz	12kHz	
5kHz	10kHz	15kHz	
6.3kHz	12.6kHz	18.9kHz	
8kHz	16kHz	(>20kHz)	3rd Harmonic ignored
10kHz	20kHz	(>20kHz)	2nd Harmonic 20k
12.5kHz	(>20kHz)	(>20kHz)	Harmonics ignored
16kHz	(>20kHz)	(>20kHz)	Harmonics ignored
20kHz	(>20kHz)	(>20kHz)	Fundamental at HF limit

4.3.7 Inverse Square Law Test

To validate the relationship between sound intensity and distance, and the performance of the different kinds of sensors, essential to the experiment, an inverse square law test was performed. The experiment involved placing a calibrated sound source at known distances from a sensor and measuring the decay in intensity to confirm adherence to the inverse square law. Theory predicts (Palacios, 1964) that the array's level readings fall off at 6dB per doubling of distance. A pink-noise loudspeaker (calibrated to 94dB at 1 m) was moved distances of 1 meter at a time up to 98 meters.[3] The distances were measured out and marked, and the height of the sensors was kept constant to avoid ground-reflection gain.

For each position, a 20-s sample was recorded simultaneously on four different microphones through an audio interface into a DAW at 48kHz, 24-bit. The microphones used were a Shure

[3] 98 meters was the length of the corridor at 5th Floor, Byrom Street, Liverpool John Moores University.

SM57 dynamic, an Audio Technica 2050 condenser, a Rode NTG-2 Condenser shotgun and a small capsule electret sensor.

In an ideal case, the inverse square law test would be carried out in an unobstructed outdoor space so that reflections are minimised. In the present study, a long internal corridor was chosen instead because it provided a straight, line-of-sight measurement path approaching 100 m, with stable environmental conditions, controlled background noise and safe access for repeated measurements.

Although the corridor walls introduce reflections and therefore deviate from a perfect free field, the primary aim of this test was to verify that the different sensors behaved consistently with the expected 6 dB per distance doubling trend under realistic, reverberant conditions, and to obtain practical calibration data for subsequent experiments. These measurements therefore serve as a relevant check on microphone performance rather than a strict laboratory validation of the inverse square law. The results would be used to calibrate the system for real-world distance variations.

4.3.8 Wind Pressure Test

The wind pressure test assessed the impact of airflow on microphone performance. A controlled fan setup generated a steady, repeatable airflow across the front of the sensor array, to probe its susceptibility to wind-induced self-noise under laboratory conditions, rather than to replicate full outdoor wind fields. Measurements focused on recording pressure fluctuations induced by wind and its effect on signal integrity. This would help greatly with informing design adjustments for wind noise mitigation.

Wind noise can raise broadband levels by 20–30dB under propeller (Prinz and Ewert, 2020) and is therefore treated as a design consideration for the drone array experiments. A bench-top axial-flow fan (Ø 300mm, nominal velocity 7m s^{-1} at centreline) was mounted 0.45m above a rotatable microphone fixture. Two test conditions were recorded: a dynamic microphone (Shure SM57) and the same microphone fitted with a 25 mm open-cell foam windshield. The capsule axis was aligned with the fan hub to reproduce the low-pressure vortex described in Alkmim *et al*. (2022).

Sound-pressure levels were logged for 60 s at 48kHz/24-bit using a Focusrite Scarlett 18i20 interface; calibration to 94dB SPL at 1kHz was performed before each run. Flow velocity was verified with a TSI hot-wire anemometer (±0.1m s^{-1}). Audio data was high-pass-filtered at 20Hz to remove interface rumble, then analysed in MATLAB where a 1-s Hann-windowed spectra

were averaged to obtain the wind-noise profile, and the intelligibility-weighted SNR (IEC 60268-16) was computed for both conditions.

The measurement protocol follows the wind-screen evaluation procedure outlined in Hosier & Donavan (1979) and Lyons *et al.* (2021), ensuring comparability with reference datasets.

4.3.9 Absorption Test

An absorption test was carried out to determine the acoustic absorption characteristics within the sound booth. This involved placing material (carpet tiles) with known absorption coefficients and measuring changes in reverberation time and frequency response with the aim of validating the sound booth's suitability as a controlled experimental environment.

The airborne sound attenuation of a sound booth carpet tile (50cm × 50cm × 5mm) was measured in LJMU's 4.6m × 3.4m × 2.9m sound-isolation booth (Room 5.12, Byrom Street). Test practice followed the diffuse-field approach set out in ISO 354:2003 **(**ISO, 2003**)** and ASTM C423-17 (ASTM International, 2017), scaled to a small enclosure in line with Cox & D'Antonio's design guidance (Cox and D'Antonio, 2017).

A Dell XPS-15 laptop running Cubase® [4]generated WAV files containing 30-s segments of white noise, pink noise, and pure-tone bursts at 250Hz, 500Hz, and 2kHz. The file was replayed via USB to a Focusrite Scarlett 18i20 interface and then reproduced with an ADAM A7 powered loudspeaker (balanced TRS). A class-1 sound-level meter (Eurisem EP-626) was calibrated at 94dB SPL, then positioned 1 m on-axis to the loudspeaker. The interface gain was adjusted until the broadband stimulus registered 100dB C at that point, ensuring repeatable source level across all trials.

A Behringer ECM-8000 omnidirectional reference microphone (±2dB, 15Hz–20kHz) was fixed in a floor stand 1 m from the loudspeaker cone. Phantom-power supply and digitisation (48kHz/24-bit) were provided by the Scarlett interface. To minimise early reflections, three 600mm × 1200mm mineral-fibre baffles were hung along the booth's glazed wall, and two additional panels enclosed the microphone.

The room background noise was recorded for 30 seconds prior to each signal playback. Every stimulus segment was captured twice (baseline and carpeted), generating matched WAV pairs

[4] Cubase is a registered trademark of Steinberg Media Technologies GmbH.

for later comparison. Waveforms were imported into MATLAB and then windowed with 1-s Hann functions. One-third-octave spectra were averaged over each 30-s segment. Following ISO 354, the random-incidence absorption coefficient of the carpet was derived from the difference between baseline and carpeted spectra:

$$\alpha(f) = \frac{0.161V}{S}\left[T_{carpet}^{-1}(f) - T_{ref}^{-1}(f)\right] \tag{4.7}$$

where V=54.2 m^3 (booth volume) and S=0.25 m^2 (specimen area). The conversion from ΔL to reverberation time ratio employed the room constant method (Jeong and Lee, 2011).

Figure 4.9 summarises the geometry used for the absorption measurements, including specimen placement and the relative positions of source and microphone. For replication, the key point is that the same loudspeaker–microphone spacing, source level calibration and capture format were maintained between baseline and treated conditions, with the only intended change being the introduction of the carpet specimen. The resulting matched WAV pairs (baseline vs carpeted) were then analysed using consistent windowing and one-third-octave averaging so that differences can be attributed to absorption effects rather than to altered gain or placement.

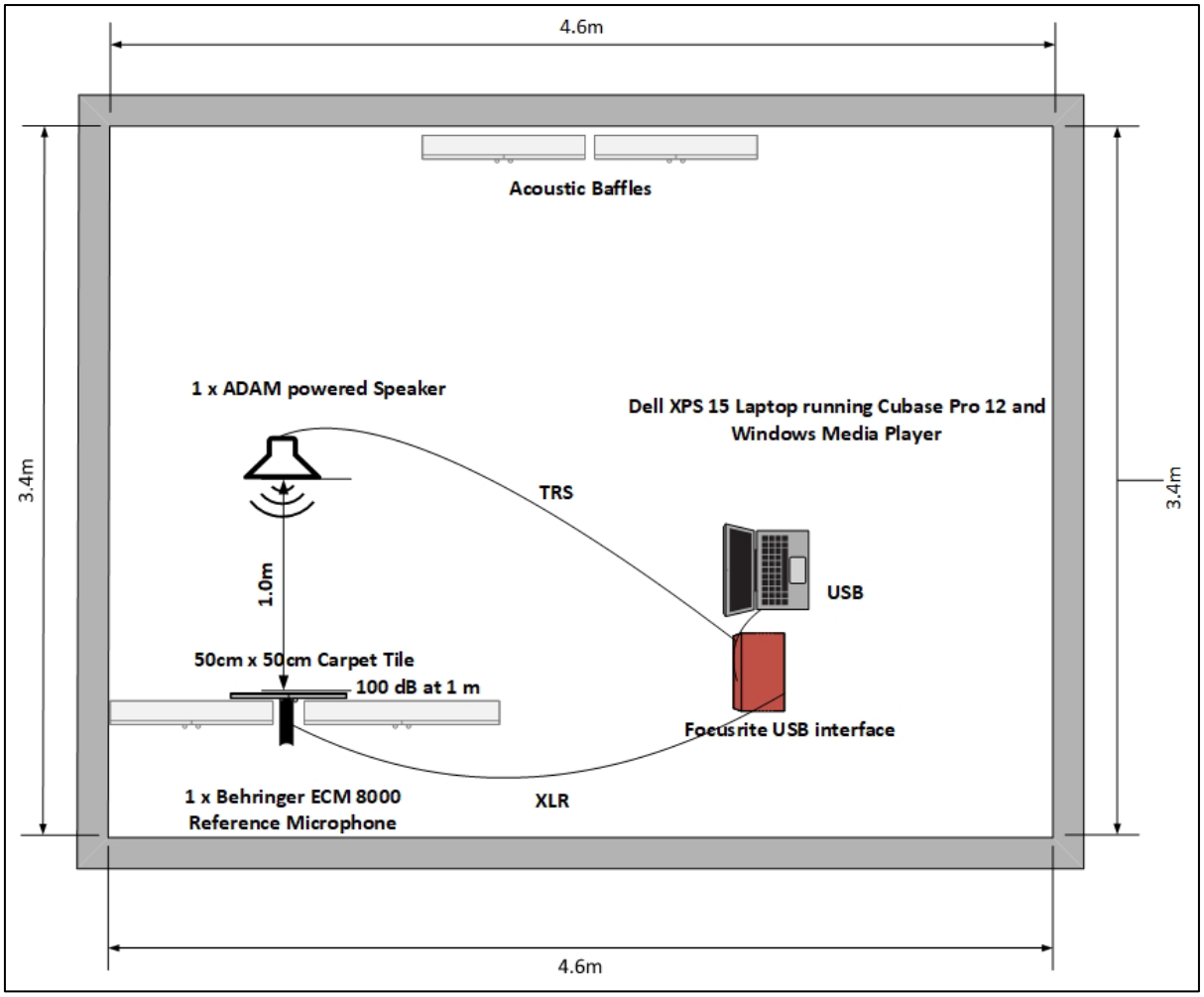

Figure 4.9. Dimensions of the carpet absorption test.

4.3.10 Multichannel Recording Session with Four Microphone Array

A multi-channel recording session was conducted in the sound booth to capture a reference data set using the complete signal chain developed in sections 4.2.1 – 4.2.8. The recording session included multiple runs with varied configurations of speakers, sensors, and test signals that were completed with the four-sensor array before moving onto a more complex sensor array.

Figure 4.10 defines the spatial layout for the initial multichannel session, including the microphone array position and loudspeaker locations used to generate repeatable datasets for later beamforming tests. The array centre was used as the coordinate origin, and loudspeaker positions were fixed to the 3 × 3 grid so that each recording corresponds to a known direction-of-arrival. All microphones were recorded simultaneously at 48 kHz/24-bit with the same gain settings once calibrated, preserving phase relationships required for steering-vector processing. The session template (routing, channel order and naming) was kept unchanged across runs to

support reproducibility and reduce bookkeeping errors when analysing the multichannel WAV files.

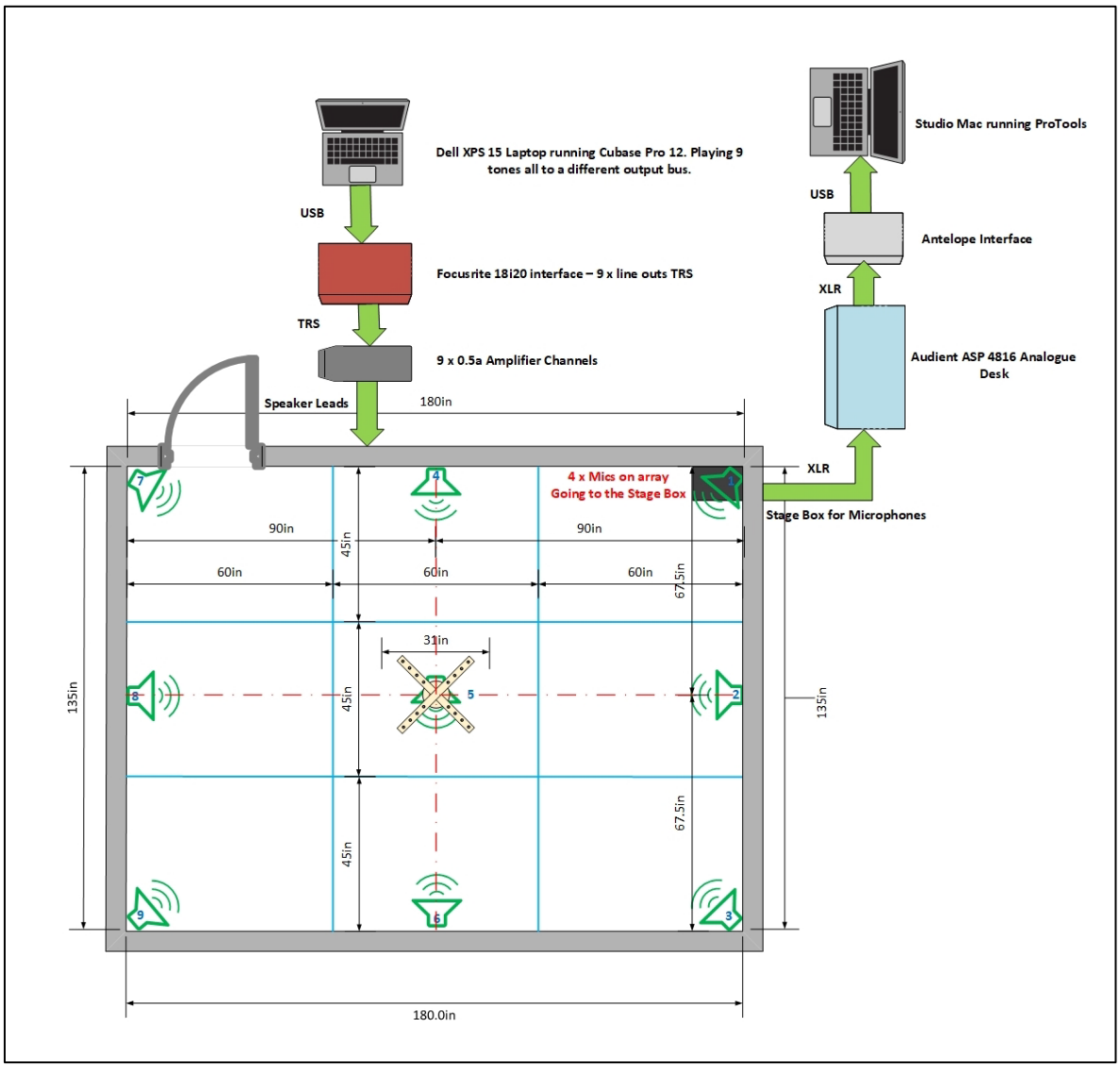

Figure 4.10. Sound-booth geometry for the four-microphone sessions with loudspeaker locations.

Table 4.8 lists the parameters of the multichannel recording session with the four-microphone setup and Figure 4.11 is a photograph of the experimental setup with the four Rode shotgun microphones.

Table 4.8. Parameters of the multichannel recording sessions.

Parameter	Details
Dimensions	4.57 m × 3.43 m × 2.90 m (l × w × h); walls and floor carpeted as characterised in Chapter 4.2.8
Grid	3 × 3 lattice (610 mm module); nine speakers as characterised in Chapter 5.2.2, centred in each cell, tweeter height 0.10 m
Signals	Nine logarithmic sine sweeps (50Hz → 22kHz, 30s) routed to discrete output buses.
Playback chain	Cubase Pro 12 → Dell XPS-15™ → USB → Focusrite 18i20 (nine TRS line-outs) → custom 9 × 0.5 A amplifier as characterised in Chapter 5.2.1
Receiver	Cross-shaped array (arm length 310 mm) carrying four sensors centred in grid cell 5
Cabling Chain	Star-quad XLR → floor stage-box → Audient ASP-4816 desk → Antelope Orion (USB)
Record format	48kHz / 24-bit WAV; synchronous capture in Pro Tools® [5](Mac Studio)

Figure 4.11 provides a visual confirmation of the physical booth arrangement during the four-microphone recordings, including stand placement and cable routing. The photograph is included to support replication of the experimental geometry and to show the practical constraints that influenced array mounting.

[5] Pro Tools is a registered trademark of Avid Technology, Inc.

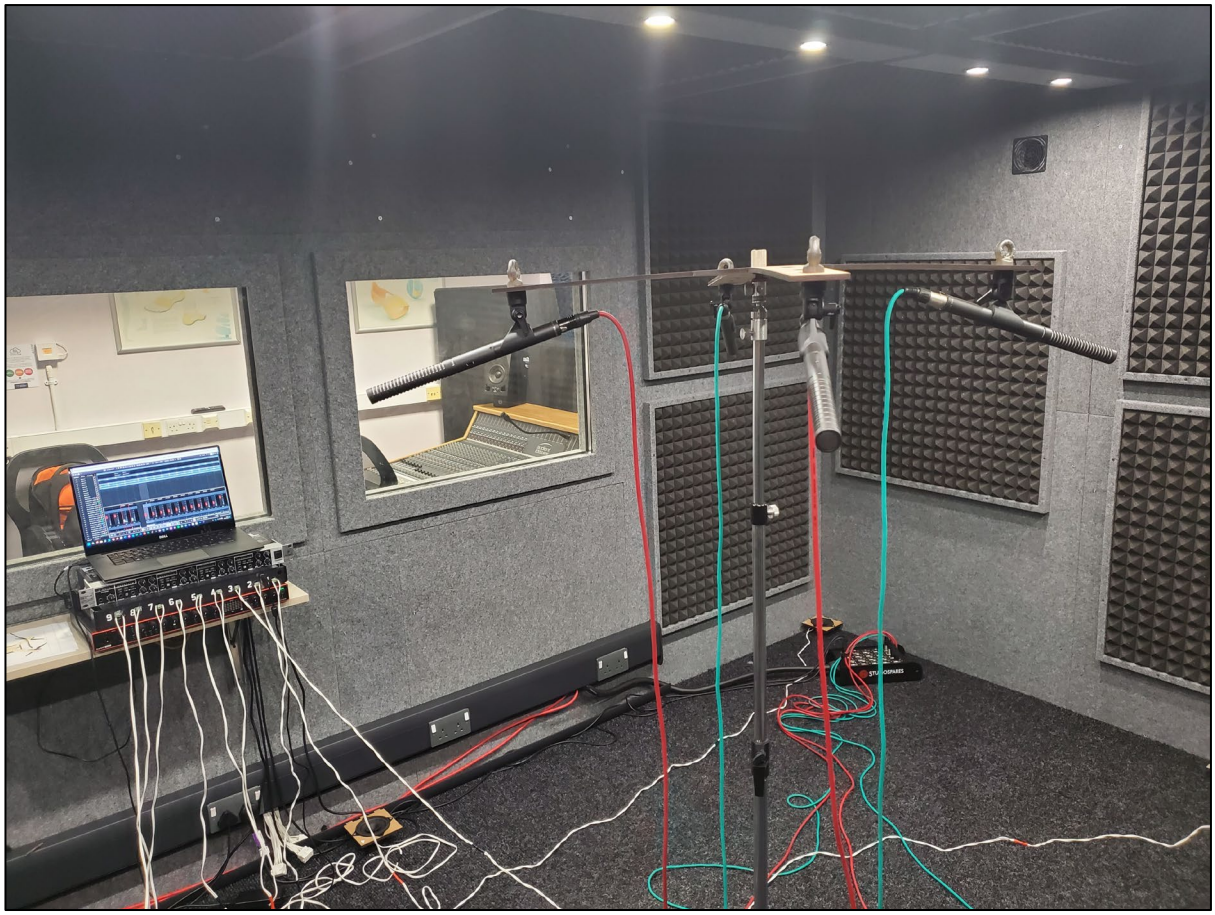

Figure 4.11. Recording session with Rode NTG-2 Shotgun Microphones.

4.4 Further Studio Experiments in Sound Booth Environment

After establishing baseline performance, further experiments were designed to test advanced configurations and to refine the system.

As weight was a significant issue with the sensor array, a sensible course of action seemed to be to design and build small capsule condenser microphones to replace the heavier microphones used in the initial testing stages. Using delay-and-sum as a baseline, the literature shows that small arrays (2–4 microphones) exhibit limited directivity and higher sidelobes at higher frequencies, suggesting a higher microphone count would be sensible. This is consistent with the results of (Qualcomm Technologies, 2021).

"Simulations show that two-, three- and four-microphone arrays exhibit "significant off-axis lobing" at 4kHz, which lowers array-gain and increases recognition errors, whereas arrays with more microphones maintain a tight main-lobe and higher SNR" (Qualcomm Technologies, 2021).

This section introduces delay-and-sum beamforming as a baseline method and motivates the move to higher microphone counts using theoretical beam-pattern behaviour and prior literature; quantitative performance results from recordings and simulations are reported later in Chapter 7.

For an M-element array with multichannel input vector, the delay-and-sum (DS) beamformer output for a look direction Ω is

$$y[n] = \mathrm{w}^{H}(\Omega)\mathrm{x}[n], \tag{4.8}$$

Here, $\Omega = (\text{azimuth } \phi, \text{elevation } \theta)$ which would be the target signal direction.

Where the DS steering weights apply phase shifts (or equivalent time delays) that align the wavefront at the array. In the frequency domain, the m-th weight is

$$w_m(\Omega, \omega) = \frac{1}{M} e^{-j\omega\tau_m(\Omega)}, \tag{4.9}$$

with $\tau_m(\Omega)$ defined by the array geometry and assumed plane-wave propagation. DS provides a computationally simple baseline against which MVDR-based steering and post-filtering are compared later (Capon, 1969; Veen and Buckley, 1988).

In light of this, a decision was made to increase the number of microphones on the array, another reason to keep the weight of the capsules down. The maximum number of simultaneous inputs in the studio environment was 16, so a 16-microphone array with low weight sensors was proposed. Figure 4.12 and 4.13 illustrates the designs of the 16-microphone arrays prior to laser cutting.

4.4.1 Sensor Array Build (16 Microphones)

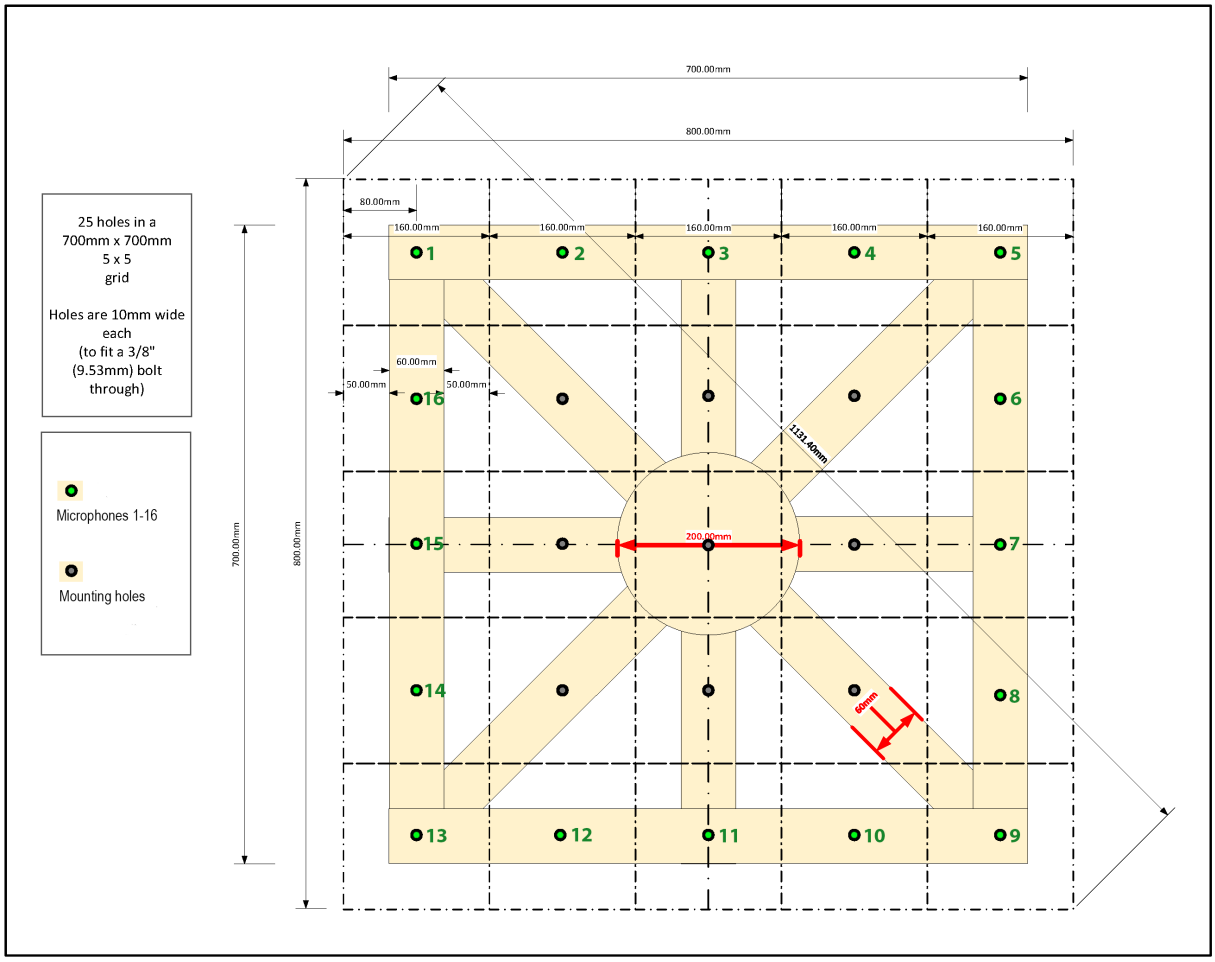

Figure 4.12. Design schematic of the square 16-microphone array. Coloured markers 1-16 indicate the microphone positions used in experiments. The remaining holes are structural mounting and fixing points and do not contain microphones.

Building on the shortcomings identified with the four-element prototype, two sixteen-microphone arrays, one square and one circular, were fabricated to achieve finer spatial resolution. The higher sensor number delivers a wider effective aperture and built-in redundancy, while the contrasting geometries allow direct comparison of beam patterns and overall coverage. Each microphone was individually calibrated, and the assembled arrays were aligned to minimise systematic bias, ensuring that any directional response reflects the intended design rather than construction tolerances. The intention for these enhanced microphone arrays was to provide a cleaner capture of directional sound than the earlier configuration. Figure 4.13 shows the circular 16-microphone array design prior to laser cutting in the workshop.

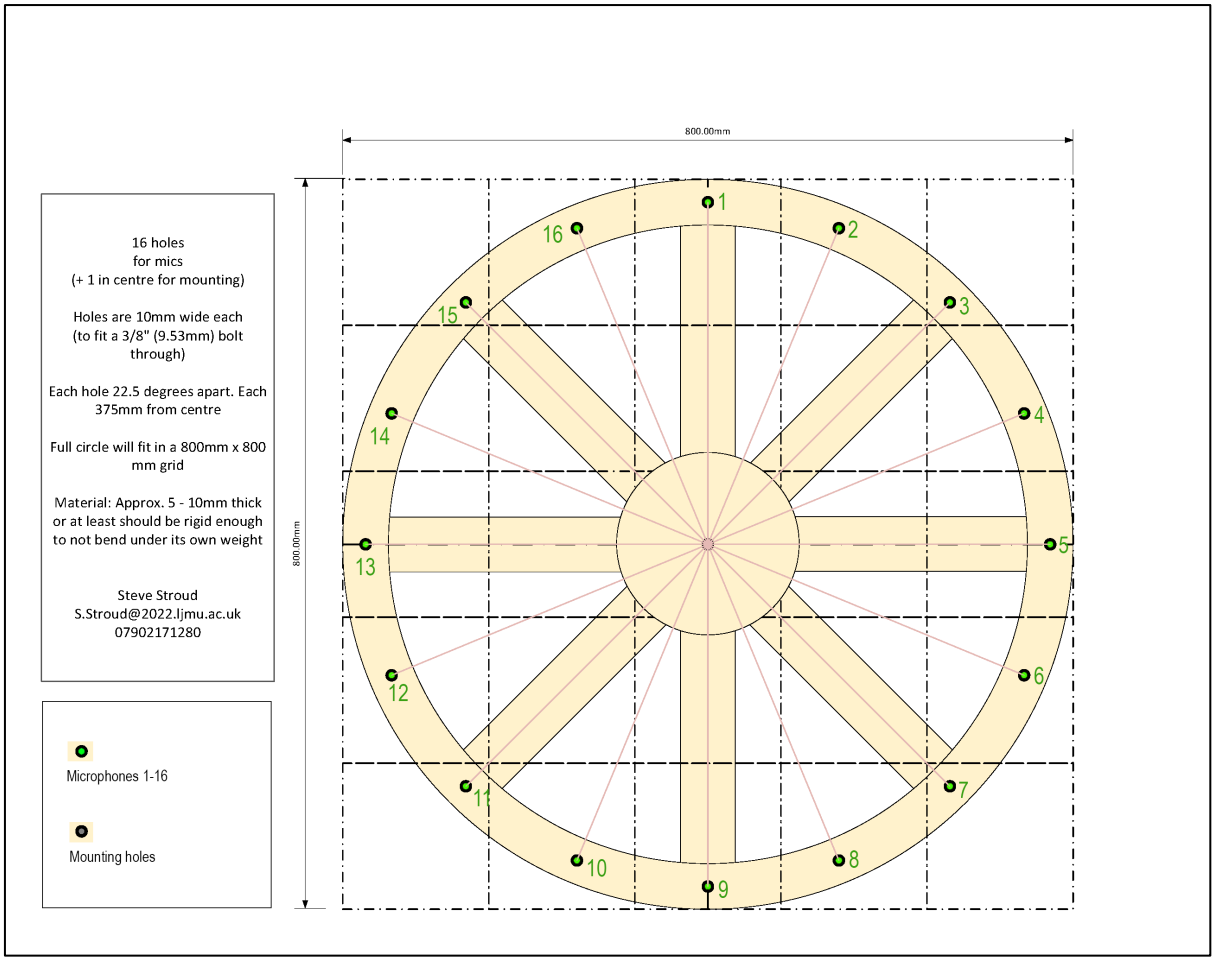

Figure 4.13. Design schematic of the circular 16-microphone array.

Figures 4.12 and 4.13 show the final CAD layouts used to fabricate the two 16-element array frames, enabling the geometry to be reproduced exactly in future builds. The purpose of presenting both geometries is to allow direct comparison of beam patterns and coverage under identical recording conditions, while keeping sensor type and channel count fixed. The labelled microphone indices define the channel ordering used during recording and analysis, which is necessary for steering, plotting polar responses and interpreting any asymmetry in the beam patterns. Both were produced on the same 80 W CO_2 laser cutter used for the initial four-sensor prototype, ensuring identical kerf width and edge quality across all builds.

The square array is shown in Figure 4.12. An 800 mm × 800 mm lattice carries 16 mounting apertures (Ø 10 mm) arranged on a 70 cm pitch. Integral radial and perimeter stiffeners maintain rigidity while keeping the overall mass within the drone payload envelope.

The circular array is shown in Figure 4.13. A 16-hole ring (Ø 800 mm) with 22.5° angular spacing provides an equivalent aperture in a rotationally symmetric form factor. Eight radial struts tie the

ring to a central boss that houses the tripod/drone shoe, yielding comparable stiffness to the square version. Both panels were cut from 6 mm plywood. The laser-cut process guarantees accuracy for every microphone seat, minimising array-shape errors that would otherwise degrade beam-forming performance.

4.4.2 Microphone Build

To populate the 16-channel array frame, low-mass omnidirectional back-electret capsules (50Hz – 16kHz; operating voltage 1.5–10 V; ϕ 9.7 mm × 6.7 mm) were selected. The capsules provide a nominally uniform polar response essential for maintaining identical gain and phase across the array and are light enough to meet the drone payload budget.

Each element was wired to a 6 m balanced lead terminated in a standard XLR to mate with the stage-box loom (maximum capacity: 16 returns). The wiring diagram for a single electret microphone attached to an XLR lead.

Figure 4.14 documents the exact wiring used to integrate each electret capsule into a balanced XLR workflow, which is critical for replication and troubleshooting. Each capsule was treated as an independent channel with consistent biasing and coupling so that channel-to-channel response differences are minimised. The intent of this design is not to exploit XLR symmetry electrically (since the capsule is single-ended), but to standardise cabling, shielding and connectivity within a 16-channel recording environment.

Figure 4.15 shows the physical arrangement of the capsule, strain relief and insulation used to ensure mechanical robustness and consistent proximity between the capsule and the bias network. This matters because movement at the capsule or solder joints can introduce handling artefacts or intermittent faults that would corrupt multichannel phase relationships. The build shown here therefore supports repeatable recording quality when the array is repositioned across sessions.

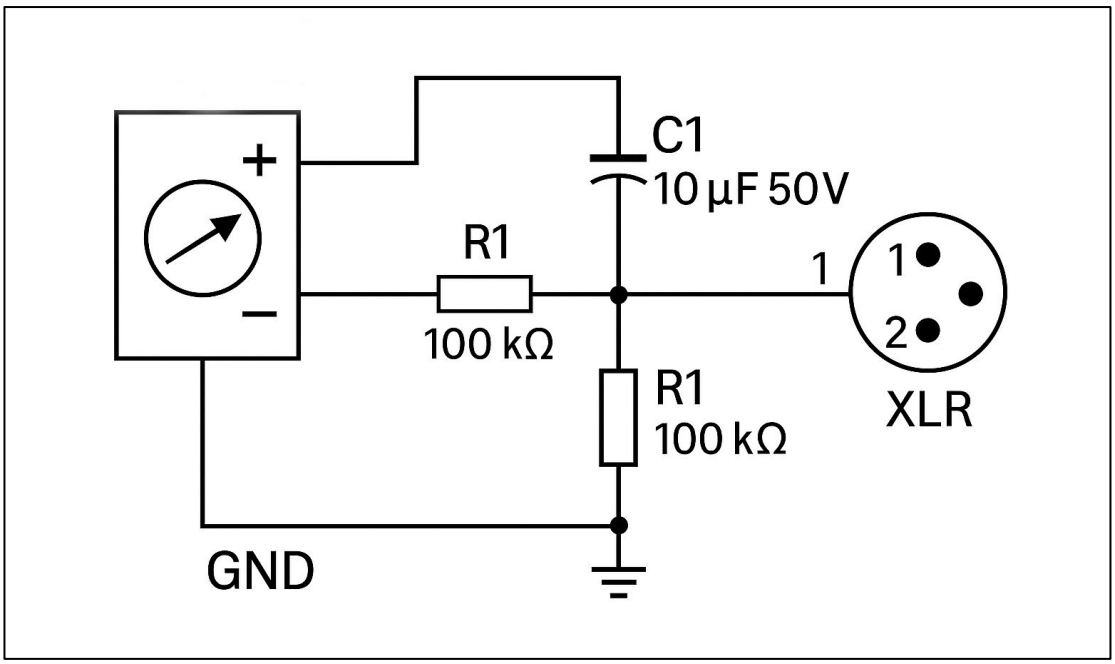

Figure 4.14. Wiring diagram of a single electret microphone.

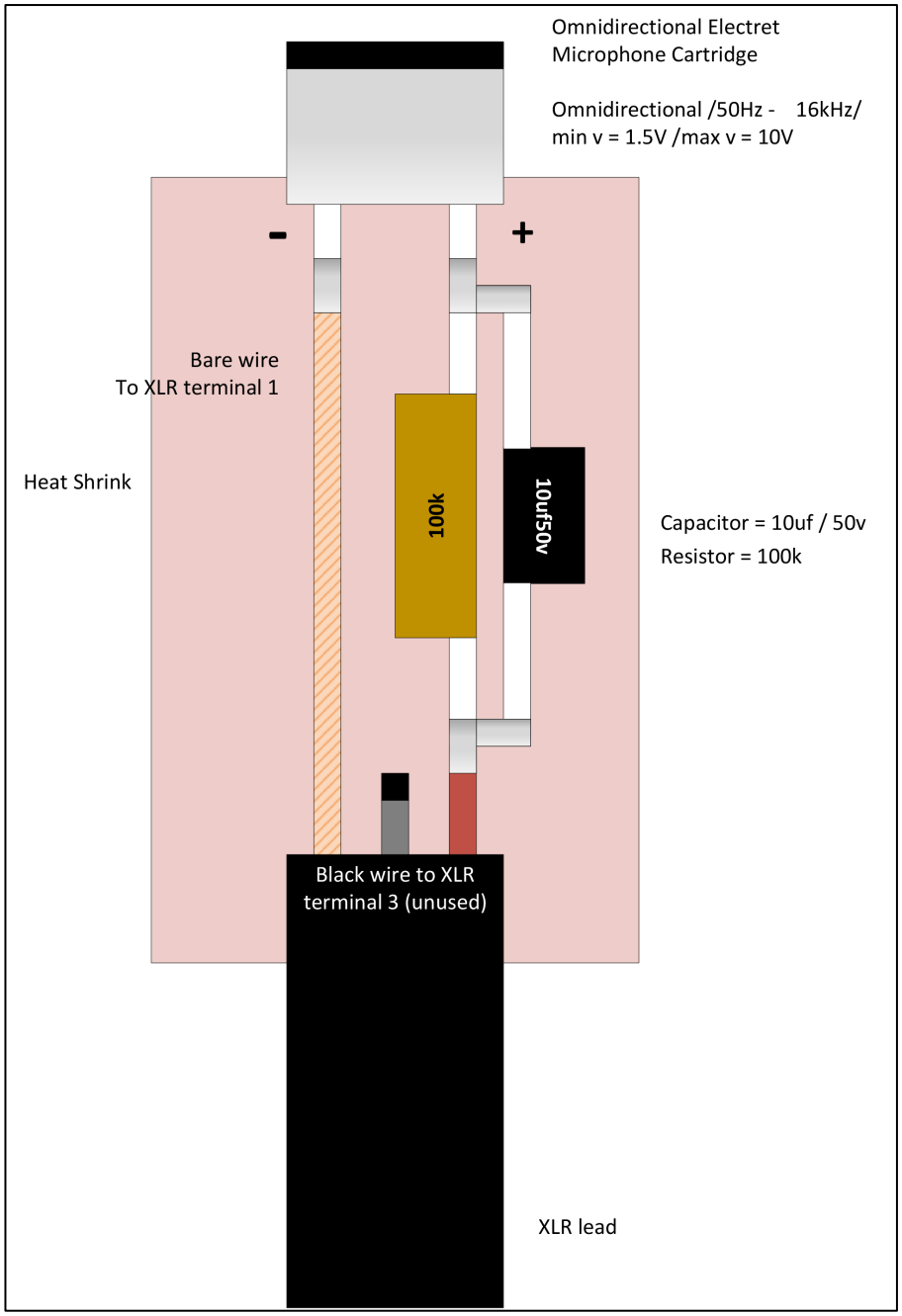

Figure 4.15. Cross-section of the custom electret microphone insert and XLR connection. The electret capsule used is a miniature type with a diameter comparable to the outer diameter of the XLR microphone cable.

Each cartridge is biased by a 100 kΩ axial carbon-film resistor and AC-coupled through a 10 µF, 50 V electrolytic capacitor. The capsule's negative lead is tied to XLR pin 1 (shield/ground), while the positive lead is routed via the coupling capacitor to XLR pin 2; pin 3 remains unconnected. The resistor, capacitor, and solder joints are consolidated in a short length of heat-shrink tubing positioned immediately behind the capsule to provide both strain relief and electrical insulation. This network presents a high-impedance pull-up for the electret FET while

blocking phantom power and preventing low-frequency offset at the pre-amplifier. Component tolerances were chosen to keep the high-pass corner below 2Hz, well outside the analysis band.

4.4.3 Custom Microphones Recording Sessions

A second recording session was carried out using the more advanced sensor array and custom microphones. Sessions were undertaken with the same setup as in Figure 4.11, with the exception of the new array frame and the custom electret microphones. The goal was to test the bandwidth and dynamic range, verify the mechanical and electrical integration of the bespoke capsules, and produce a richer dataset for subsequent optimisation of the audio-zoom and beam-forming algorithms. Figure 4.16 shows the labelled booth setup used for the 16-microphone session, including array placement relative to the loudspeaker grid and the fixed reference orientation used for subsequent steering. The recording procedure matched the earlier four-microphone session (sample format, routing strategy and fixed gain after calibration), with the key change being the higher channel count and the custom electret sensors. This figure, together with the session parameters table, allows the reader to reconstruct the full experimental geometry and reproduce the multichannel dataset generation step prior to any beamforming or post-filtering.

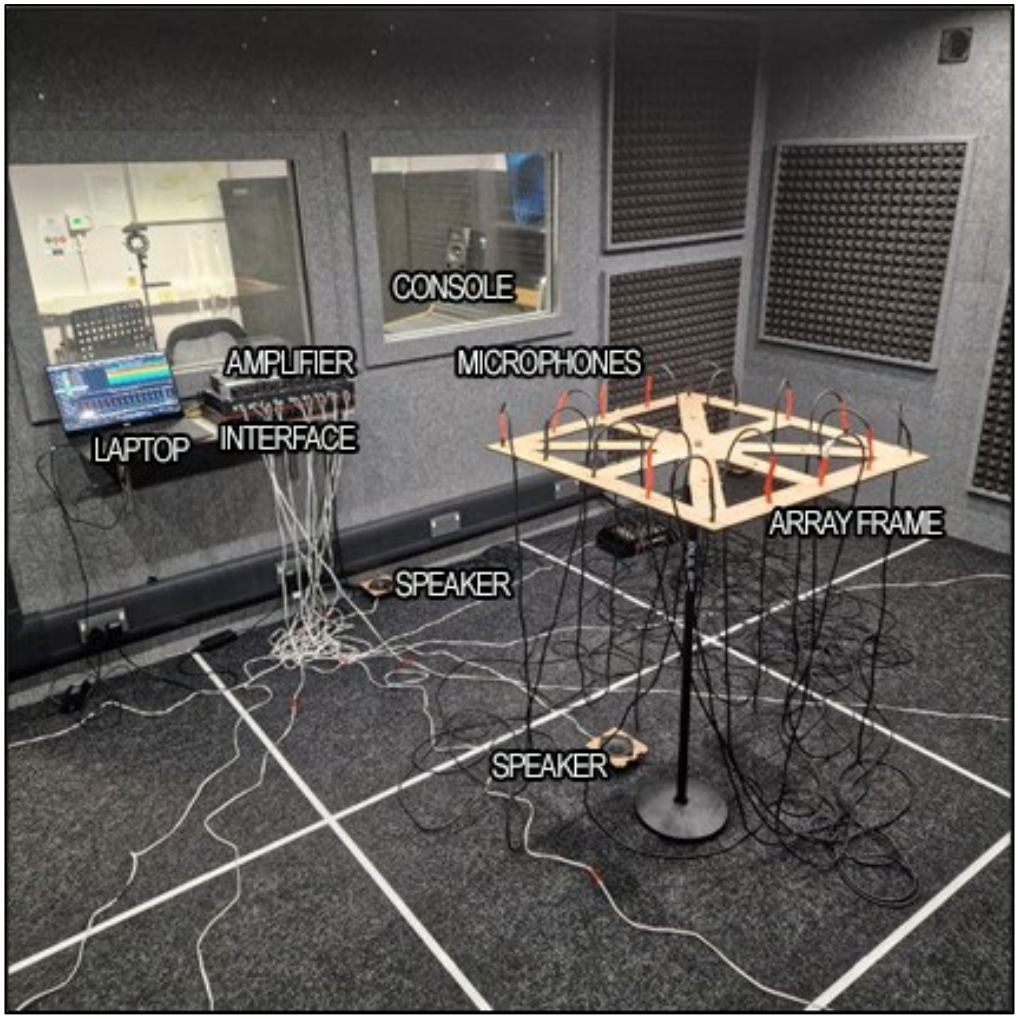

Figure 4.16. Labelled experimental setup with the custom square array.

4.5 Summary

This chapter has set out how the proposed audio-zoom system was designed, tested and validated, moving from tightly controlled studio work, through physics-based simulations to proof-of-concept trials with real-world environments.

Section 4.1 catalogued the core equipment, four classes of microphones, purpose-built amplifiers, multichannel recorders, and modular loudspeaker rigs. These components were chosen with the constraints of a Police drone payload in mind.

Sections 4.2 and 4.3 documented two phases of sound-booth evaluation. The first used a four-microphone array to establish baselines for wind loading and absorption and the verification of inverse-square experimental behaviour; the second scaled the array to sixteen elements and introduced custom capsules, generating a reference dataset for later algorithm tuning.

These initial experimental approaches allowed the next stage acoustic physics-based simulation research to be calibrated and validated. Studio measurements informed the acoustic parameters of the model, and then the calibrated measurements obtained in these experiments fed directly into the physics-based simulation approach that will be introduced in Chapter 5.

Chapter 5: Simulation Approach

5.1 Overview

Chapter 5 provides a bridge between the initial experimental work of Chapter 4 and the Exemplar House simulations and field demonstrations in Chapter 6. It first defines the simulation framework used to generate multichannel microphone signals for controlled evaluation of the audio-zoom pipeline. The framework is implemented in MATLAB and encodes the microphone geometry, drone-relevant constraints and the rotor-noise characteristics used in the thesis. The simulator is then applied to two environments: a reflection-controlled sound booth and a reconstruction of the Ava White crime scene in Liverpool city centre. These simulations provide repeatable test conditions that allow direct comparison with the laboratory measurements reported in Chapter 4 and the field results reported later.

5.2 Simulation Framework

This chapter evaluates the proposed audio-zoom pipeline in controlled virtual environments before field deployment. Each simulation is defined by (i) a 3D scene geometry and material assumptions, (ii) source positions and input signals, and (iii) the microphone array geometry and sampling parameters. The simulator generates multichannel microphone signals by modelling direct-path propagation and reflections consistent with the chosen environment, and then adds interference components including drone egonoise and ambient noise at controlled levels.

The resulting microphone signals are processed using the same beamforming and post-filtering pipeline used for the real recordings, and performance is logged using SNR and intelligibility metrics. The purpose is to provide repeatable test conditions that complement the sound-booth and field experiments reported later. Beamforming and post-filtering follow the formulations defined in Section 3.7; the present chapter focuses on how the acoustic scenes and multichannel signals are generated and evaluated.

5.3 Simulations of Real World Environments

This section describes the models and simulation setup used to emulate real-world environments. Results obtained from these simulations are presented in Chapter 7. Using MATLAB as the platform, simulations of real-world environments were chosen to be the next logical step in the process. Three environments were chosen; the sound booth, a city centre crime scene location and buildings on site at Liverpool John Moores University. The sound booth was chosen as the

first location to simulate, since the data from the initial physical experiments could be compared and developed to create a virtual testing environment for the audio zooming algorithm. This simulation would provide a more consistent experimental setup and enable controlled manipulation of all acoustic variables.

5.3.1 Implementation in MATLAB

The simulation framework in this chapter is implemented in MATLAB primarily as a rapid-prototyping environment for multichannel array processing and acoustic scene modelling. MATLAB in this project is used for matrix operations, FFT/STFT, covariance estimation and visualisation, and also for custom-developed functions written for this thesis (array geometry generation, steering-vector construction, beam scanning, MVDR and delay-and-sum weight computation, and post-filtering/noise-suppression stages. Where MATLAB built-in functions are used (e.g., FFT, linear algebra solvers), they serve as standard interchangeable numerical tools rather than novel custom algorithms.

The equations in Section 3.7 define the proposed theoretical methods and are therefore not MATLAB dependent. The same processing chain can be reproduced in other programming environments that support equivalent numerical operations, such as Python or C++. Porting requires only matrix arithmetic, covariance estimation, constrained beamformer weight computation, and time-to-frequency transformations. To support reproducibility, the implementation is organised as modular functions with defined inputs and outputs.

5.3.2 Simulation of the Sound Booth

To verify that the audio-zoom algorithm behaved as expected in a space with controlled reflections, the first virtual scene recreated the Sound Booth used as the location for the initial experiments described in Chapter 4. The booth is represented in accurate dimensions as a 4.56m by 3.42m by 2.24m, space, giving a volume of 35m³. Acoustic treatment follows the measured room: pyramid foam 50mm thick covers seventy per cent of the wall and ceiling area with a Sabine absorption coefficient of about 0.68 above 250Hz, while the exposed paint-finished MDF panels retain a coefficient close to 0.07. Background noise recorded during the experiments in Chapter 4 was measured at 23dBA.

Scene construction begins with the import of the booth geometry and the associated absorption coefficients. Fifth-order image-source models then yield an impulse response for every loudspeaker microphone pair. Each response is convolved with clean speech, and a measured

rotor-noise spectrum is mixed in at a level ten decibels below the target voice. Delay-and-sum beamforming was applied to the resulting multichannel signal using the classical formulation in Eq. 3.16, with steering delays computed from the array geometry (Eq. 3.15), after which signal-to-noise ratio and short-time objective intelligibility are logged.

The processing chain is deliberately sparse, matching the hardware available when the physical booth tests were run. A three-by-three loudspeaker grid is captured by a square sixteen-microphone array whose beam steering relies on the classical delay-and-sum approach as shown in Figure 5.1. The results that follow, therefore, act as a reference set against which later improvements can be judged.

Chapter 6 re-uses this code base but introduces three key upgrades. Firstly, the microphone aperture becomes circular, bringing even azimuthal coverage and lower sidelobe energy. Second, delay-and-sum steering is replaced by a minimum-variance distortionless response beamformer, which affords deeper nulls for off-axis interferers. Finally, a post-processor that combines Wiener filtering with independent component analysis is then added to isolate the target waveform. The stripped-back data presented below thus provide the calibration point for the full audio-zoom algorithm examined in Chapter 6.

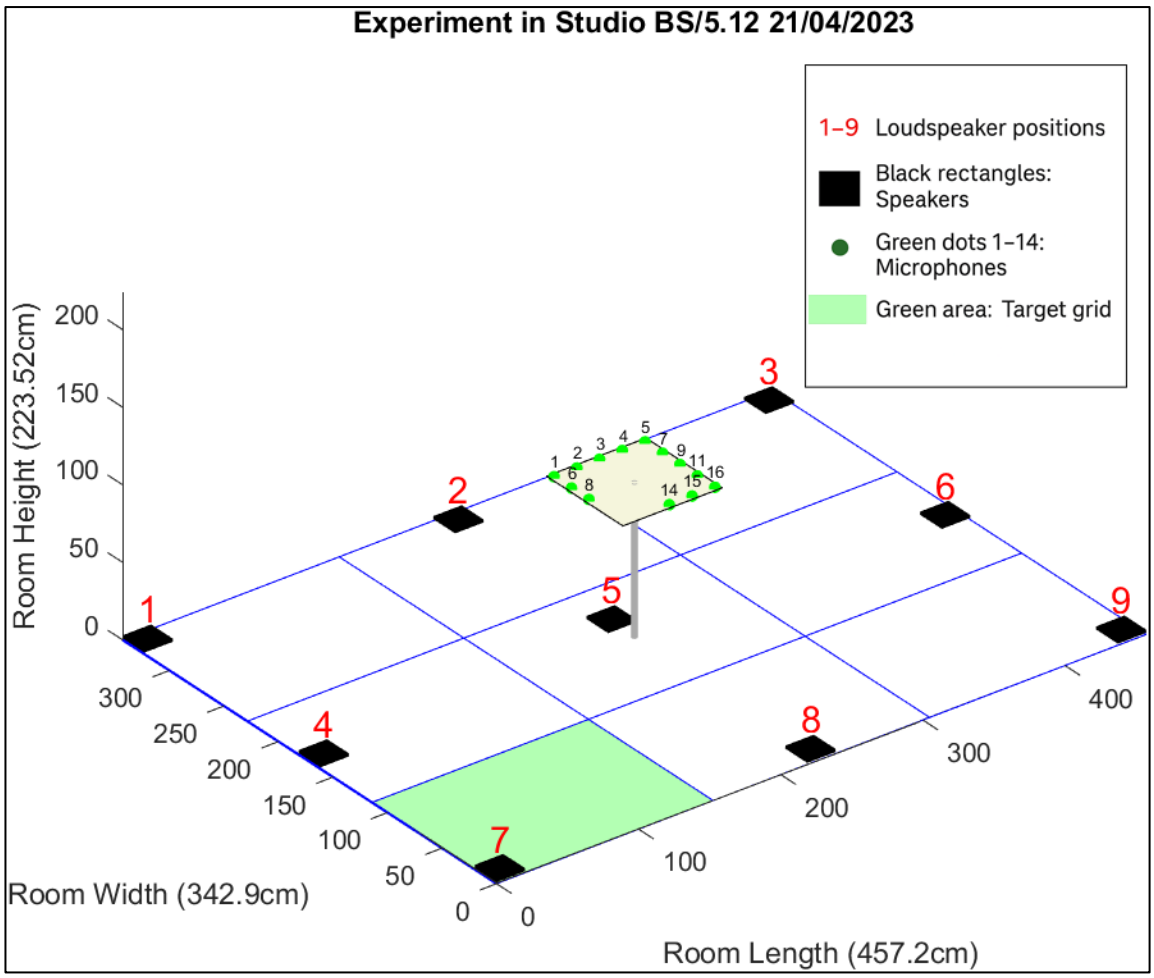

Figure 5.1. Layout of the Studio BS/5.12 robust zoom experiment. Red numbers 1–9 mark the loudspeaker positions at floor level. The central beige plate shows the 16-microphone array mounted on a stand; green dots mark individual microphone capsules, with three intentionally omitted to simulate sensor failure. The shaded green area indicates the target grid for the audio zoom focus.

5.3.3 Simulation of a Crime Scene

A second simulation was created to explore how the algorithm copes with the less predictable acoustics of an outdoor crime scene. The model represents the courtyard in Liverpool city centre where the Ava White murder occurred. Irregular boundaries, hard reflecting façades and multiple competing noise sources create a demanding testbed before any field deployment.

The courtyard is approximated as a rectangle thirty metres long and fifteen metres wide, bordered on two sides by brick and glass walls ten metres high. Asphalt, with an absorption coefficient close to 0.03, forms the ground plane. The brick sections of the façades follow frequency-

dependent coefficients between 0.02 and 0.05; glazing is set to an amount of 0.07. The remaining sides of the rectangle are open to adjoining streets, allowing lateral traffic noise to enter the model. Three persistent background layers are added: a diesel-engine rumble concentrated between 100Hz and 400Hz, crowd noise that dominates the 500Hz to 3kHz band, and a gusting wind spectrum based on a speed of four metres per second.

Scene construction starts with a satellite-derived footprint that is extruded to create the façades. All surfaces are discretised, after which a fifth-order image-source routine generates impulse responses that incorporate the frequency-dependent absorption. Ambient noise and the same rotor-noise spectrum used in the booth study are included, then a speech source at a height of one point six metres is stepped through nine grid locations labelled G1 to G9.

The recorded multichannel signal is processed by a delay-and-sum beamformer whose output feeds an MVDR stage. The resulting signal then passes through an adaptive Wiener filter. For each grid position, the simulation records signal-to-noise ratio, speech clarity (STOI) and any localisation errors.

Chapter 7 compares these simulated metrics with the results of playback field tests conducted on site. Figures 5.2 and 5.3 give, respectively, the MATLAB reconstruction of the courtyard and the satellite view from which the geometry was derived.

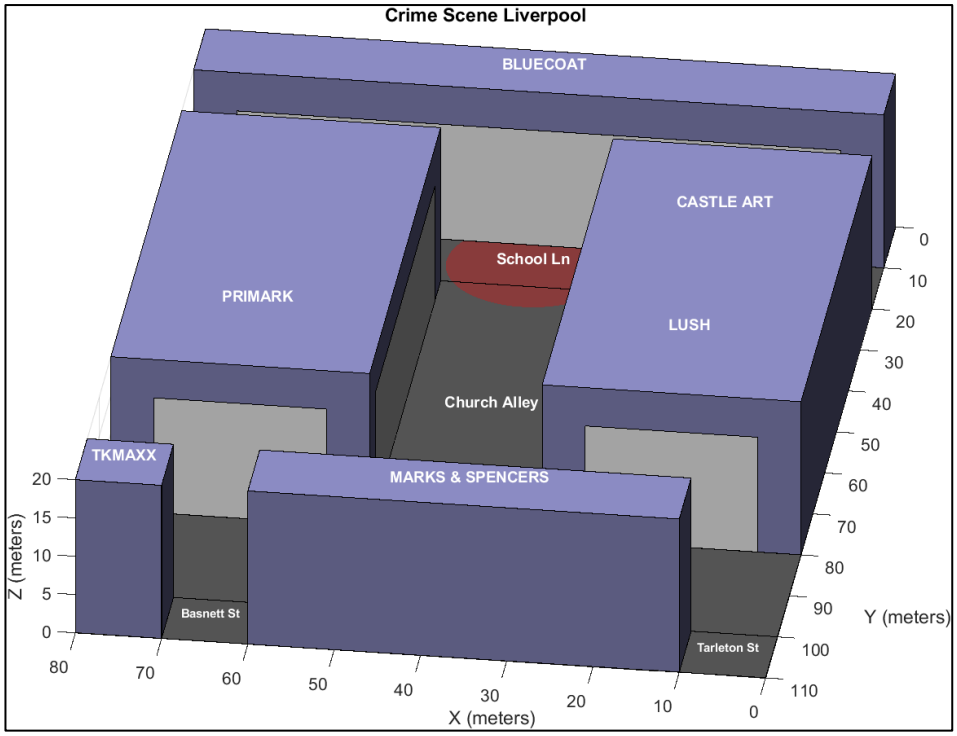

Figure 5.2. MATLAB Simulation of the Liverpool Crime Scene.

Figure 5.3. Google Earth view of the murder scene (53.40452, -2.98377) (Google Earth, 2025).

5.4 Summary

Chapter 5 explained how the project moves from laboratory measurements to fully controlled numerical experiments. It began by outlining the practical limits that a drone payload imposes lightweight omni-directional microphones, custom low-power amplifiers, a MATLAB modelling suite and a fixed library of reference signals, and showed how these boundaries informed both the physical prototypes and their virtual counterparts.

The preceding sections revisited the core acoustic principles underpinning the model, restated the governing equations for wave propagation and geometric spreading, described the image-source method used for reflections, and sets out the two principal beam-forming options, delay-and-sum and minimum-variance distortionless response. The same section reviewed the noise-reduction and source-separation tools, spectral subtraction, Wiener and sub-space filters, and optional ICA that are applied to the beamformed output.

Finally, Section 5.3 explains how theory turned into practice. Three increasingly demanding digital twins are created: a reflection-controlled sound booth, an urban courtyard that reproduce the Ava White crime scene, and the Exemplar Houses complex on the university campus. Each scene was rendered in three dimensions, given frequency-dependent surface coefficients and loaded with calibrated loudspeaker material plus measured drone self-noise. The model steers beams across a nine-point grid, tests resilience to microphone drop-out and logs signal-to-noise and intelligibility metrics before the most promising configurations are reproduced in physical trials. The methods described here set the stage for Chapters 6 and 7, which will tie the bespoke hardware and software to the validation tests that demonstrate how well the system performs.

Chapter 6: Exemplar Houses

6.1 Overview

There was a realisation that real-world testing would be impractical in the Church Street, Liverpool City Centre crime scene location. An alternative idea was to simulate buildings close to the Byrom Street University campus, where access to testing would be more accessible. The decision was then made to create another physics-based simulation of Liverpool John Moores University's Exemplar Houses, located on the Byrom Street Campus. The buildings would be easier to model and then validate with physical tests.

A 3D model in MATLAB reproduced the geometry and surface finishes of the houses, applying Sabine absorption coefficients to walls, windows and furnishings. The scene incorporated a configurable microphone array plus a noise-reduction sub-array. Nine virtual loudspeakers were placed, eight around the scene and one beneath the array to emulate environmental sound sources; the SPL, azimuth, elevation, and beam-width values were defined and customisable. The model propagated sound waves, calculated reflections, and then applied MVDR beamforming and post-filtering so that predicted gains in signal-to-noise ratio (SNR) could be logged for each grid cell in simulation.

To validate the simulation, field experiments at the physical houses were undertaken. A self-contained power pack, mixer and recorder made the rig independent of mains supply. Tests began with a single reference loudspeaker and then replicated the full nine-speaker configuration used in the simulation. Together, these processes allowed verification that the work developed in Chapters 4 and 5 scales from idealised simulated spaces to a realistic outdoor setting.

Although the Exemplar Houses do not reproduce every aspect of a busy city-centre crime scene, such as dense crowds, irregular source movements and complex background noise, they provide a controlled but representative built environment in which key propagation effects and array behaviour can be realistically assessed.

6.2 Simulation of Exemplar Houses

This section discusses utilising MATLAB for creating and applying a beamforming-based audio zooming algorithm, specifically for use with drones in a surveillance capacity. It uses the foundation of sound beamforming technology to inform this work, offering a theoretical

underpinning that the reader must understand in order to comprehend how the ‘zooming’ algorithm works and why the decision to use it in combination with a 16-microphone array was reached. The chapter explains how the Minimum Variance Distortionless Response (MVDR) beamformer was selected as the most appropriate method for achieving satisfactory results with a mobile platform, with an array of lightweight microphones, and in a simulation of real-world scenarios.

Figure 6.1. Photograph of Exemplar Houses at Liverpool John Moores University.

Figure 6.2. Google Maps view of the Exemplar Houses, houses outlined in yellow. (53.41163, -2.98122) (Google Maps, 2025b).

Chapter 5 brought to light the difficulty of extracting certain audio signals in a drone's typical noisy environment. Wind, engine noise, and various environmental sounds conspire to obscure the desired target audio signals. To address this dilemma, Chapter 3 touched upon possible solutions, with an emphasis on spatial filtering techniques that might enhance the signal quality by eliminating the undesired sounds. The choice of beamformer was also a critical consideration. In early experiments, a delay-and-sum beamformer was used as an initial baseline; however, all beamforming results presented in this chapter use the MVDR beamformer (see Eq. 3.17), unless explicitly stated otherwise.

The MVDR beamformer (Capon, 1969), favoured in this work, optimally adapts to the trade-offs in mic-array configurations that the authors face when trying to maintain computational efficiency. This is especially important in applications such as audio zooming for Police surveillance. The audio zooming algorithm chosen for the simulation, showing every major step of the computational process in MATLAB is laid out in Figure 6.3, while figure 6.4 focuses on the audio zooming part of the algorithm in greater detail. For clarity, the governing mathematical relationships are summarised first (propagation delay, steering, beamforming, and enhancement),

and the primary MATLAB implementation is provided in Appendix B (Figures B.1–B.2, Tables B.1 – B.11). The theoretical foundations that underpin the processing stages used here such as beamforming, noise reduction, and source separation are established in Chapter 3. This chapter focuses on how those methods are configured for the Exemplar Houses geometry and how parameter choices affect practical performance.

6.2.1 Exemplar Houses Simulation Workflow Summary

The Exemplar Houses workflow used in this chapter is summarised as pseudocode in Algorithm 6.1 (with full MATLAB implementation details provided in Appendix B). References to the governing equations are provided in Section 6.2.2 and Chapter 3 (e.g., MVDR and post-filter formulations).

```
Inputs:
  Scene dimensions and 3D geometry definition
  Material absorption coefficients / reflection settings
  Array options (shape, mic count, spacing, array position)
  Noise-reference (NR) sensor option
  Source definitions (positions, SPL, azimuth/elevation, beamwidth,
signals)
  Grid definition (grid size, grid index/choice)
  Processing options (MVDR parameters; post-filter on/off;
imported/virtual signals)

Outputs:
  Multichannel simulated microphone signals
  Beamformed and enhanced outputs (audio + metrics + figures)
1. Start; set global parameters (fs, c, scene dimensions).
2. Create 3D scene model from geometry and surface/material definitions.
3. Select array location (default or user-defined x,y,z).
4. Select array shape (square/circle/octagon/cross/star):
     If selection changes: update array geometry and redraw scene.
5. Configure noise-reference (NR) sensor / sub-array (if enabled).
6. Select grid size:
     If grid size changes: update grid layout and redraw scene.
7. Define sound sources:
     Choose default or user-defined source positions (x,y,z).
     Define SPL, azimuth, elevation, beamwidth, and signal assignment
for each source.
8. Generate sound waves / signals:
     Choose (a) virtual signals or (b) imported audio signals.
     If parameters change: regenerate source signals.
9. Enable/disable reflections:
     If enabled: compute reflection paths using scene surfaces and
chosen reflection order.
10. Create/update 3D visualisation of the scene (for figures/validation
views).
11. Select grid choice (target evaluation cell/direction):
     Update interactive legend / display selections as required.
12. Distance and delay calculations:
```

```
     For each source s and microphone m:
        compute distance d(s,m), delay τ(s,m) = d(s,m)/c, sample delay
n(s,m) = round(τ fs).
13. Signal synthesis (per microphone channel):
     Sum direct-path contributions (delayed/attenuated).
     Add reflections (if enabled).
     Add noise sources (if enabled).
14. Beamforming (MVDR):
     Apply MVDR to the multichannel signals using the selected steering
direction/grid.
15. Post-filter (optional):
     Apply spectral subtraction / Wiener or TF masking using configured
parameters.
16. Produce outputs:
     Export WAV files; generate plots/figures; log SNR (or other
metrics) per grid cell.
17. End.
```

Algorithm 6.1: Exemplar Houses MATLAB workflow (pseudocode, as illustrated in Figure. 6.3).

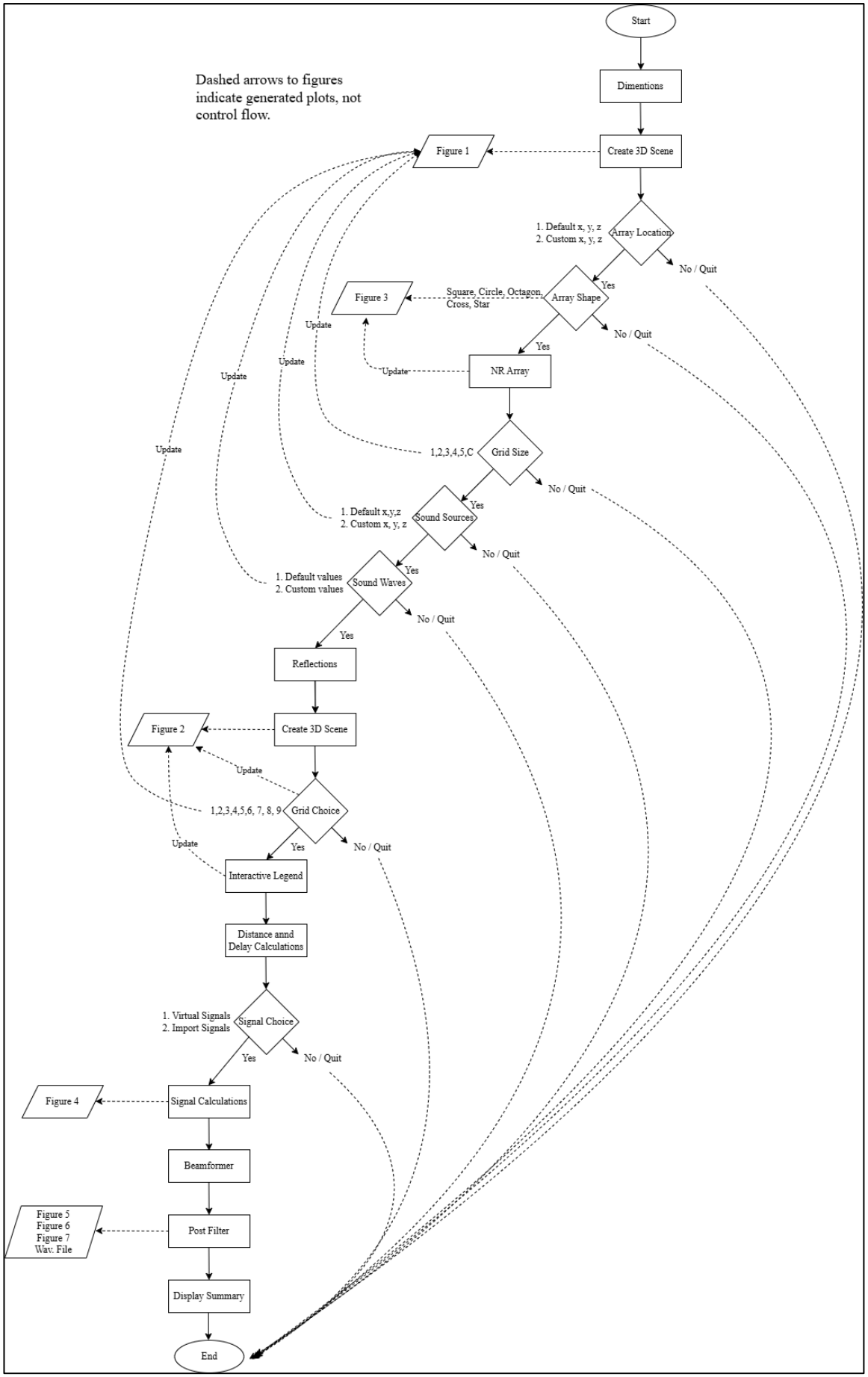

Figure 6.3. Flowchart of the MATLAB algorithm.

```
Inputs:
  fs, array_geometry, mic_coordinates, speaker_coordinates
  steering_grid (azimuth/elevation candidates)
  audio_multichannel x_m[n] for m = 1..M
  MVDR parameters (covariance window/averaging, diagonal loading ε)
  Post-filter parameters (mask type, smoothing, exponent, thresholds)
Outputs:
  y_pre[n]   (summed pre-beamformed signal)
  y_bf[n]    (MVDR beamformed signal)
  y_filt[n]  (post-filtered / separated signal)
1. Read input parameters (fs, grid, array, source coords, audio).
2. Compute microphone azimuth/elevation angles from mic_coordinates and
array reference.
3. Compute beam steering grids (candidate look directions).
4. Configure MVDR beamformer:
     4.1 Compute steering vectors for each look direction.
     4.2 Select the desired look direction (grid index / azimuth-
elevation pair).
5. Apply MVDR beamformer to multichannel audio:
     5.1 Convert x_m[n] to STFT X_m(f,t).
     5.2 Estimate spatial covariance R(f) over time frames (optionally
apply diagonal loading ε).
     5.3 Compute MVDR weights w(f) for selected direction.
     5.4 Form beamformed spectrum Y(f,t) = w(f)^H X(f,t).
     5.5 Invert STFT to obtain y_bf[n].
6. Generate baseline reference signal:
     y_pre[n] = sum_m x_m[n]  (or specified pre-beamformed reference)
7. Post-process beamformed output (optional):
     7.1 Apply spectral subtraction or Wiener/TF masking as configured.
     7.2 If separation is enabled, apply the configured separation
stage to y_bf[n].
     7.3 Output y_filt[n].
8. Normalise y_filt[n] using the chosen level reference (peak or RMS).
9. Export WAV files: y_pre[n], y_bf[n], y_filt[n].
```

Algorithm 6.2: MVDR-based audio zoom workflow (pseudocode, as illustrated in Figure 6.4).

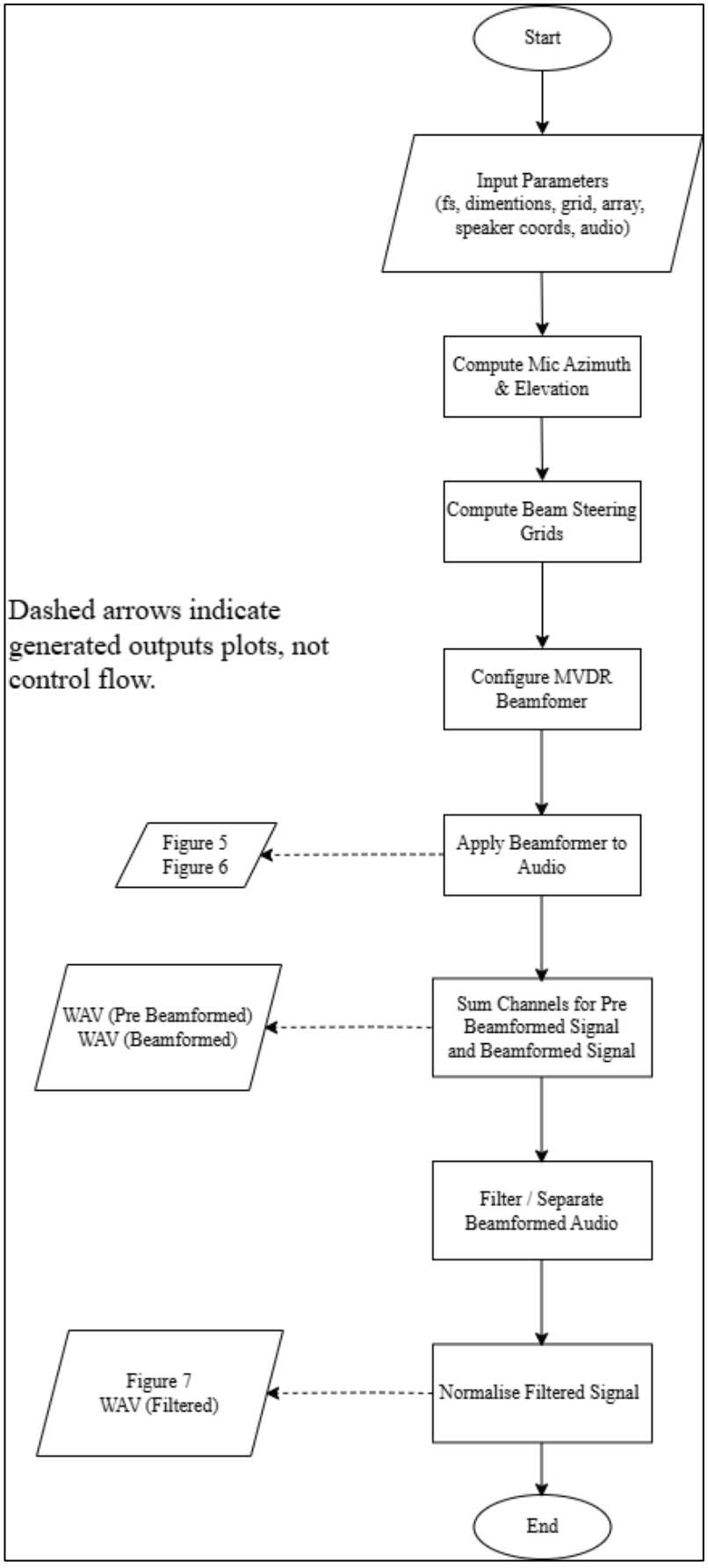

Figure 6.4. Flowchart of the audio zooming aspect of the MATLAB algorithm.

6.2.2 Mathematical formulation used in the Exemplar Houses simulation

The Exemplar Houses simulation models propagation from each source position to each microphone position using a distance-to-delay relationship. For a source–microphone separation distance d, the propagation delay is $\tau = d/c$, and the corresponding discrete delay in samples is $n = \mathrm{round}(\tau f_s)$. (These relationships are used throughout the simulation when generating and aligning multichannel signals.)

Beamforming is then applied using the MVDR formulation established in Chapter 3 (Eq. 3.17). In brief, MVDR computes a weight vector that minimises the array output power subject to a distortionless constraint in the desired look direction; this produces a steered output that suppresses off-axis interference while preserving the target direction.

Post-processing is applied to the MVDR beamformed output to reduce residual interference. The mathematical definitions for spectral subtraction (Eq. 3.19) and Wiener filtering / time–frequency masking (Eqs. 3.20–3.21) are given in Chapter 3. Where enabled, source separation follows the mixture model in Eq. 3.22. Implementation details are provided in Appendix B and referenced from the main text.

6.2.3 Introduction to Beamforming

Beamforming is a critical technology in various applications such as radar, telecommunications, and especially in audio signal processing. It uses sensor arrays to filter spatial information from signals, improving the signal-to-noise ratio by focusing on specific directions or regions while attenuating signals from unwanted directions. The fundamental idea behind beamforming is to apply delays to signals received by individual microphones, aligning the signals coming from the direction of interest and reinforcing them. In contrast, signals from other directions remain misaligned and are suppressed.

Regarding audio zooming, beamforming could assist in isolating audio sources in a particular location within a scene. This process is similar to the directional microphone principle but is more flexible as it employs multiple sensor elements rather than a single one. The dynamic environment where drones are used for surveillance demands the best possible audio performance. It can be a target-rich or a noise-rich environment. The unwanted sounds that might interfere with the sounds of interest can come from either side of a dynamic drone platform. Those sounds can be ubiquitous: wind, engines, traffic, and similar. Beamforming is an especially effective technology for this problem. The downside is that beamforming, like other forms of

spatial audio, is only as good as the array of microphones and the spatial filtering algorithms used to process the signals coming from the array.

6.2.4 Virtual Environment Setup and Dimension Definition

Creating an accurate simulation environment for testing the beamforming algorithms is crucial. The virtual environment is intended to accurately reflect not only the basic properties of the physical world, such as room dimensions, building forms, and street layouts but also precisely portray the spatial relationships important to the project. For instance, special care was taken to ensure that the distances between the virtual sound sources (or "targets"), the microphone array, and any other objects in the scene were consistent with what you would find in the real-world version of the same scene. Using government blueprints and real-world measurements taken on site, the scene was simulated in MATLAB in great detail, down to the acoustic Sabine coefficient properties of the materials used in the construction, so that sonic reflections would be accurately modelled.

Precise dimensions influence how accurate the simulation is. For instance, the outdoor dimensions of a crime scene are such a large space that an accurate simulation of how sound waves might behave cannot be done without considering the many kinds of surfaces that sound waves might hit. In addition, the height, length, and width values must be used in the scene to accurately render a realistic sound interaction with outdoor surfaces. The simulation allows the user to scale up or down the size of any sound environment and repeat the process for any set of sounds the user may wish to reproduce.

For this work, it was decided that the simulation space would measure 50 meters long, 80 meters wide, and 50 meters high since these dimensions create a large, open space that is well-suited for investigating the use of audio zooming and beamforming in applications where the goal is to isolate sounds from within a significant area of space.

6.2.5 Microphone Array Configuration

How well a microphone array configuration performs beamforming directly results from the user's design choice. This simulator allows users select from an array of options for the kind of microphone array to use, such as linear, circular, or custom configurations. These designs have different advantages and are more efficacious in various contexts. For example, a linear array is efficient and presents an uncomplicated computational challenge. A circular array, on the other hand, performs well, but it is more complicated to deal with mathematically and is preferable to

any geometrically asymmetrical array, which would be impractical for any 360-degree coverage, such as drone surveillance. The beamforming performance is highly dependent on accurate microphone placement, which will enhance the chances of accurate sound localisation.

The developed simulation can show how various placements improve or impair beamforming, at least in optimal conditions. It can be used to determine which positions work best in a room or at a crime scene, moving the array to where the user thinks that it will be most effective. Each microphone in an array records sound at slightly different times, with the difference in distance to the sound source accounting for most of that variation. The array of signals is used to focus on the source of interest while disregarding other sounds that might also be present.

The interactive definition of these geometries allows for the storage of the coordinates for each microphone in matrix form in the array. This matrix is then used in the calculations relevant to the beamforming process.

6.2.6 Microphone Array Design

The end user has a choice of a square, circular, octagonal, cross or eight-pointed star-shaped microphone array in the computer simulation. The choice allows for a direct comparison between the array shape and how this impacts the array's combined polar response and the recording of the soundfield. Figure 6.5 shows the five microphone array geometries that were tested in the Exemplar Houses simulations: square, circular, octagonal, cross and a symmetric star. In each case the upper plot is an overhead view of the array, with the array centre in blue and the individual microphones in green. The lower plot shows the corresponding MVDR beam pattern in azimuth when the array is in omnidirectional mode.

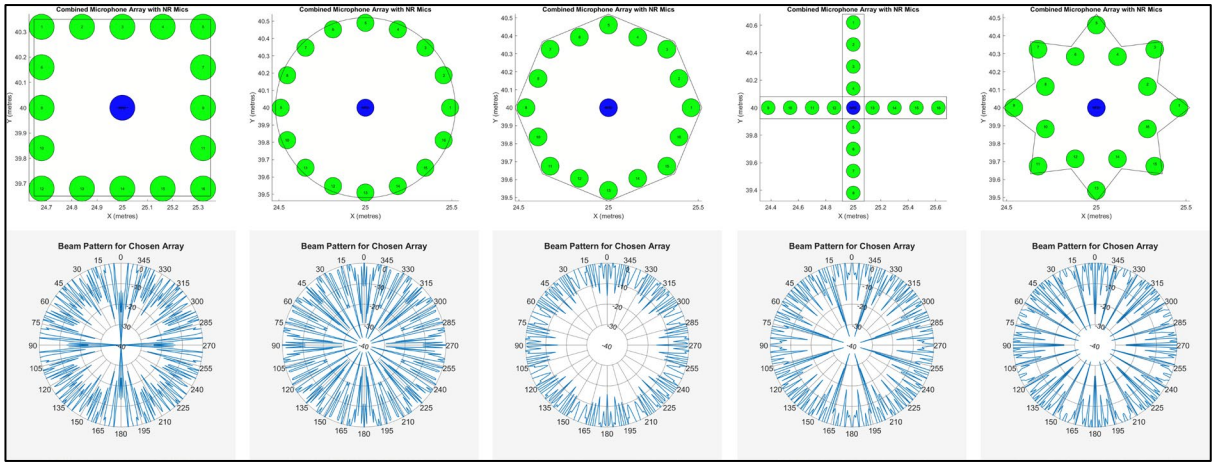

Figure 6.5. Geometric array layouts (top row) and their corresponding MVDR azimuthal beam patterns (bottom row).

The result of the different microphone positions on each geometric array provides a different polar response, which would impact the resulting beamforming performance. The geometric arrays on offer to the end user were designed and tested to be robust in the event of microphone failure on the array. Figure 6.6 reproduces the square 16-microphone array used in the robustness experiments reported in Stroud *et al*., (2023). The two panels show the microphone positions and numbering around the perimeter of the square frame, which are used later when simulating element failures and comparing beamformer performance. This numbering scheme allows specific microphones to be 'removed' in the simulations while keeping the underlying geometry fixed, so that the effect of sensor loss on the array's directivity pattern can be quantified. A conference paper, 'Robust Audio Zoom for Surveillance Systems: A Beamforming Approach with Reduced Microphone Array' was published by the authors at the IEEE 37th International Conference on Information Technologies (InfoTech-2023), highlighting this design feature

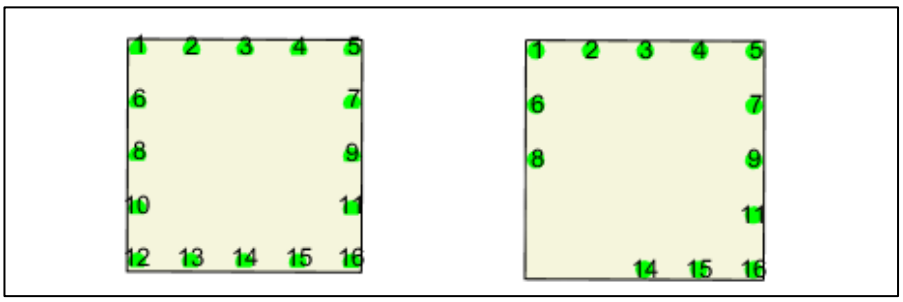

Figure 6.6. Microphone positions and numbering for the square 16-element array used in the robustness array experiments from (Stroud *et al.*, 2023).

In that study, the array geometries were evaluated under simulated microphone loss by deliberately disabling one or more sensors and observing the resulting degradation in the combined polar response and beamformed directivity. The key finding was that beamforming remained feasible with reduced microphone counts, with only a limited loss of focus for moderate sensor loss. This directly motivates the inclusion of multiple geometric array options in the simulator and the ability to enable/disable microphones, since a drone-deployed system must tolerate individual sensor failures while still providing a usable steered output.

6.2.7 Sound Source Simulation and Wave Propagation

A crucial feature of the simulation is its capacity to portray several sound sources dispersed throughout the environment. These sources could be anything from human dialogue to noise that one might encounter in the background while trying to hear a conversation, such as the sound of a motor running or the wind outdoors. The coordinates of these sound sources are vital for performing calculations that determine how well the microphones pick up the sounds in the environment based on the relative positioning of the microphones to the sound sources.

After the number and location of the sound sources are specified, the simulator forms virtual sound waves that move through the virtual environment. The sound waves will eventually be absorbed and reflected off the scene's surfaces. Every building material in the simulation is programmed with accurate acoustic Sabine coefficient properties to better simulate a realistic sonic environment. Less acoustic energy is absorbed by a smooth, hard surface like glass; therefore, more of the original signal is reflected into the environment (Kuttruff, 2016). The opposite is generally true of softer surfaces. Since sound does not have a visual form, to allow the end user to easily understand how the virtual sound waves behave in the virtual environment, a short-form visual indicator of six straight lines per waveform is created to display the direction, power and reflection angles of the sound sources' sound waves.

The simulator allows investigators to see how various sonic situations, such as having several speakers at various powers and angles, interfering with an environment, affect beamforming.

Figure 6.7 illustrates the Exemplar Houses simulation scene and the 3×3 grid used for beam steering. The terrace of houses and surrounding garden walls form a small residential street, with the microphone array positioned in grid 5, the centre of the scene, in between the houses. Nine potential sound-source locations (blue pillars labelled 1–9 in red) are distributed around the scene. In this example, the operator had selected grid 3 as the region of interest, so this cell is

highlighted in yellow, and the beamformer is steered towards it. Red rays show the direct and reflected propagation paths from the sound sources, and the yellow markers denote the corresponding reflection or collision points on obstacles and walls. These pre-computed paths are later used to derive the time delays and amplitudes that form the steering vectors for the beamformer.

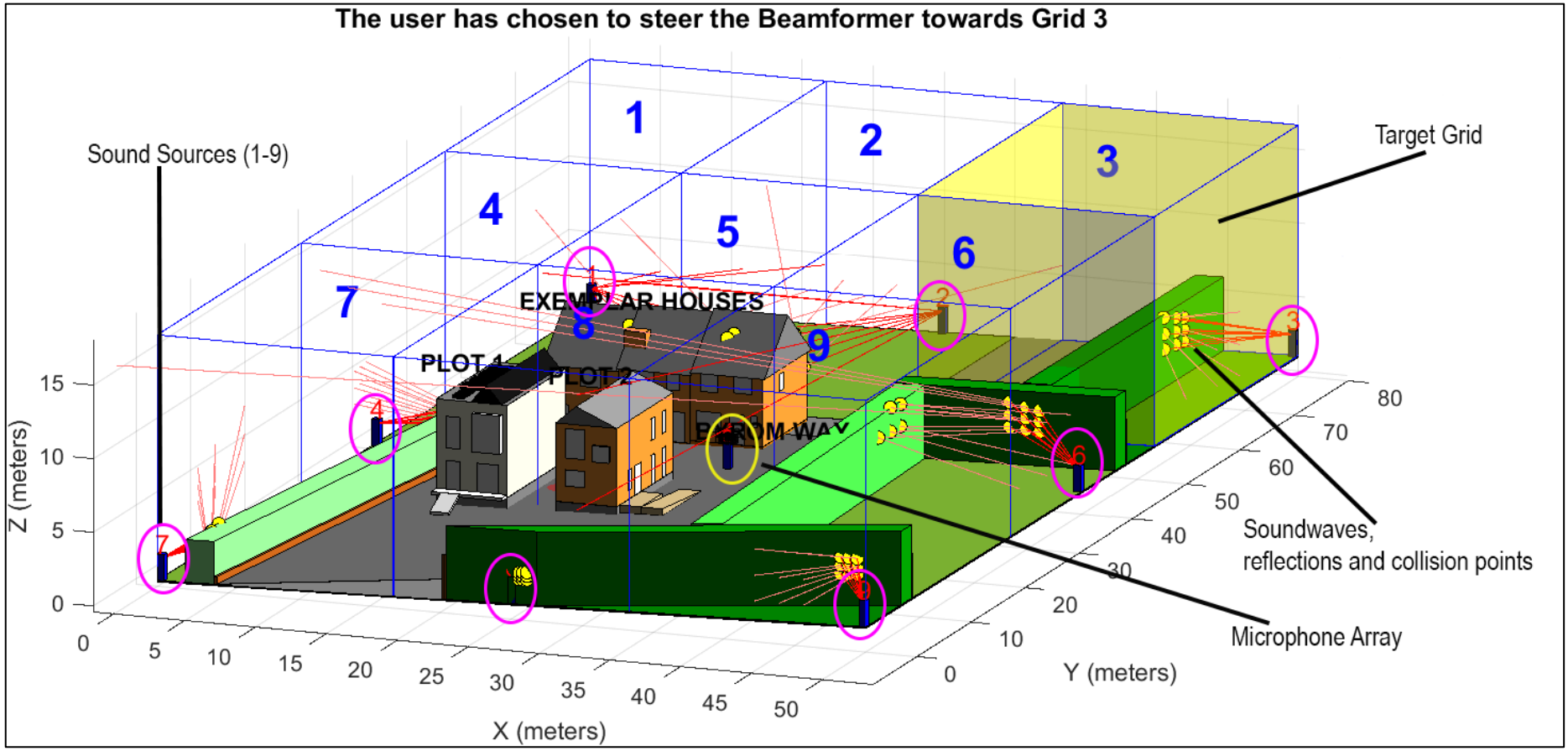

Figure 6.7. Exemplar Houses simulation scene showing the microphone array, source locations, the 3×3 grid (cells 1–9) and the propagation paths with the beamformer steered towards grid 3.

6.2.8 Delay Calculations and Distance Estimation

Delays in time are crucial to the beamforming process. When a sound wave strikes a microphone array, it reaches each individual sensor at slightly different points in time, contingent upon the line-of-sight distance from the sound source to each microphone. The beamformer's first task is to determine these momentary differences; doing so permits the multi-sensor array to act like a single microphone with adjustable polarity pickup patterns. This grants the array enhanced directional properties (Brandstein and Ward, 2001).

To ascertain the time delay, the distance is divided by the speed of sound. Next, the delay between each microphone is determined and forms the basis of a simple time delay algorithm used for beamforming. With these known values, the simulation can calculate the path for sound travelling through the array of microphones and determine from which direction the sound is coming. This forms the foundation of the delay-and-sum beamforming technique, which time shifts the signal from each microphone in an array to synchronise with the other signals. Once all the audio signals are synced, they can be summed to produce an output that resembles the sound captured by a single microphone placed at the target location (Veen and Buckley, 1988).

6.2.9 Beamforming Process and Audio Zooming

Beamforming is a central part of the audio zooming algorithm. The simulation is capable of performing multiple beamforming methods. It can be operated via time delay using a delay-and-sum approach. The recorded time delays are used to align the signals from the microphones and stored in a matrix. After the sound from the target direction has been recorded, stored and aligned, they are summed. The result is enhanced sound from the target direction, while sounds from unwanted directions are attenuated.

The microphone signals can also be processed using a Minimum Variance Distortionless Response (MVDR) technique. This permits an audio effect similar to "zooming in" on a sound source, amplifying it to the desired level while keeping other surrounding sounds in the equation at a level that makes them irrelevant.

This MVDR method substantially enhances the signal-to-noise ratio relative to the array's original output (Capon, 1969). The MVDR beamformer can be directed at a sound source and reduce noise from other directions, including the spaces around the intended target. This works by considering unwanted noise in the signal as interference and then reducing it. This is partly achieved by giving the proper weight to each microphone in the array. The MVDR beamformer

attempts to achieve the desired outcome of steering the audio using a small amount of energy, making it a highly economical beamforming method.

An MVDR beamformer employs these directional cues, moving from a basic delay-and-sum method to an enhanced interference rejection approach. The microphone array input audio is processed using the MVDR beamformer to create a beamformed audio output. This output is then normalised in preparation for noise reduction and visualisation. This function has advantages for a certain kind of application involving audio zooming, in which concentrating on a particular sound source in a dynamic environment is crucial. Working in conjunction with the user's chosen target grid, the MVDR beamformer will steer the array towards the desired location using an azimuth value. The focused beam pattern after beamforming to Grid 7 (122.5°) can be seen in Figure 6.8.

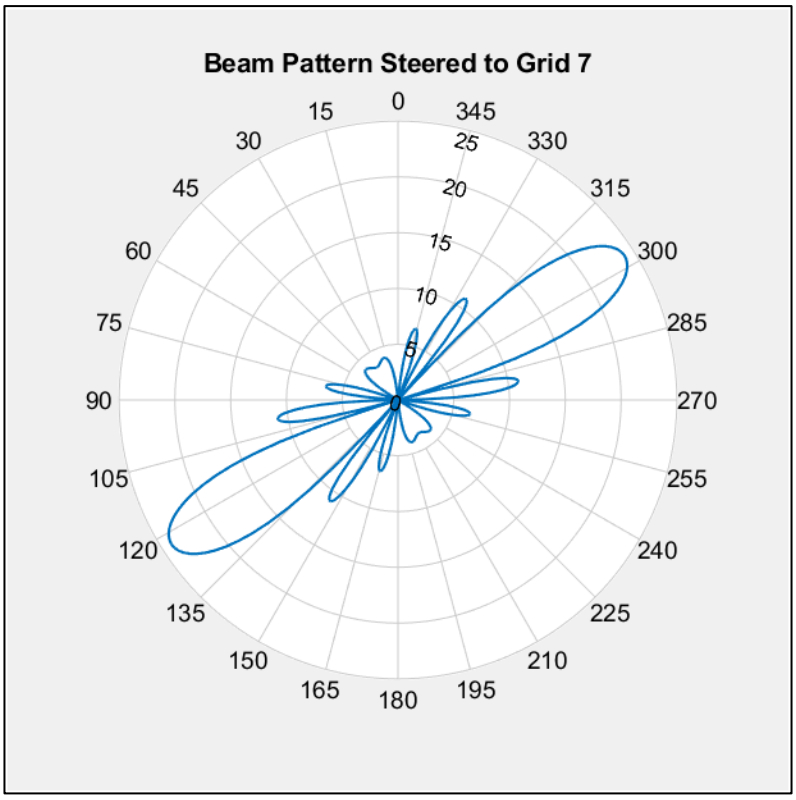

Figure 6.8. Polar response of the array after beamforming to a chosen direction.

6.2.10 Noise Reduction and Signal Enhancement

In the environments in which drone surveillance operates, there are likely to be many different types of unwanted sounds. To carry out its mission, the drone must be capable of distinguishing between the sounds it is interested in and the "noise" that is not relevant to its task. For this reason, noise reduction is a significant part of the design of the audio simulator.

The first step is to apply the MVDR beamforming algorithm to the various microphone signals. "Beamforming" in this context means creating a virtual microphone with an audio pickup pattern similar to an actual microphone that would be used locally in the test. After the signal has been "beamformed" or made more directional, "noise reduction" can be applied to clean up the signal further. The simulator is capable of multiple methods of noise reduction. One way, used after beamforming, is called subtractive noise reduction. First, a noise estimate is made from the beamformed signal. This estimate can be used in the next step of subtractive noise reduction. Using the beamformed signal to estimate noise makes the noise reduction more accurate.

After the MVDR beamforming stage is completed, the drone noise is treated as the interference or distortion it is, and a spectral mask is created, which performs spectral subtraction to carve out a target signal by dramatically attenuating unwanted frequencies. Overly dominant frequencies within the target audio are adjusted to improve intelligibility.

A Wiener filter (Wiener, 1949) is then deployed to clean up the signal further. The Wiener filter offers several advantages over standard high or low-pass filtering. Firstly, it gives a reasonable estimate of the unwanted interference when a well-defined target signal is available and then provides the user more options for adjustment. This improves the likelihood of clearer target audio, which would benefit Police using a drone surveillance system. Once the desired filtering is achieved, the Griffin-Lim (1983) algorithm reconstructs the signal in the time domain.

Although beamforming significantly improves target signal isolation, the output still contains residual low-frequency drone noise, particularly a 150Hz tone and a broad-band component between 200 and 600Hz. To address this, a single-stage spectral subtraction noise reduction was applied to each five-second segment of the beamformed signal. Mid-range drone noise is time-varying and partially overlaps speech, so spectral subtraction, which relies on an accurate noise estimate, can only attenuate it if the target speech is to be preserved; complete removal of the drone noise would leave residual harmonics or cause speech distortion artefacts.

The beamformed signal, denoted as x[n], is generated by steering the 16-microphone array towards the target. To provide an independent noise estimate, a feed-forward reference microphone (d_{ff}[n]) was positioned 40 cm vertically above the centre of the array. Although a feedback microphone (located 2 cm below the array) is available, it was not used in the current implementation but remains an option for future adaptive cancellation techniques.

To ensure that the noise captured by the reference microphone is phase-coherent with the beamformed mixture below 1kHz, the reference channel was aligned using an integer sample delay. The delay was calculated using the following expression:

$$\tau = \left\lfloor \left(\frac{\Delta z}{c}\right) f_s \right\rfloor = 18 \text{ samples @ } 48\text{kHz} \quad (6.1)$$

Where τ is the delay in samples, Δz is the vertical separation, c is the speed of sound, and f_s is the sampling rate. For the 0.4m offset at a 48kHz sampling rate, the required delay was determined to be 18 samples.

For each five-second audio clip (N = 240,000 samples at 48kHz), a single Fast Fourier Transform (FFT) was computed for both the beamformed signal and the aligned noise reference:

$$X(k) = \mathrm{F}\{x[n]\},\ D(k) = \mathrm{F}\left\{d_{ff}[n - \tau]\right\}. \quad (6.2)$$

Where *X(k)* is the FFT of the beamformed signal, and D(k) is the FFT of the delayed reference channel.

Previous work combining beamforming and post-filtering have been used for drone-mounted microphone arrays. Hioka *et al.* (2016) report on speech enhancement using a drone-mounted array with a beamforming stage coupled to post-filtering, targeting improved extraction of ground speech in the presence of strong rotor/ego-noise. More recently, Manamperi *et al.* (2024) proposed a drone-audition framework that applies multichannel Wiener filtering for ego-noise reduction and a Gaussian-mixture-model-based parametric Wiener post-filter to suppress residual noise, exploiting motor-current-specific noise statistics and demonstrating performance at very low signal-to-drone-noise ratios.

The methodology in this thesis is similar in that it uses covariance/array-based spatial filtering (MVDR) to suppress off-axis interference and applies a Wiener/mask-type post-filter to attenuate residual noise in the beamformed output. The differences are that the present work focuses on an audio-zooming pipeline evaluated within a physics-based built-environment simulation and validated through field measurements. And the noise reduction stage uses Wiener/TF masking and related post-filtering, which is designed to be applicable without requiring motor-current telemetry. The present work uses MVDR beamforming as the primary spatial filter for steering,

whereas Manamperi *et al*. centre the approach on multichannel Wiener filtering for drone-noise reduction.

6.2.11 Source Separation Techniques

Once the noise has been reduced, the simulator further disentangles the sound sources, starting with the highest spatial-separation condition and progressively decreasing the separation to the lowest.

In theory, ICA-based sound source separation takes the original sound source and reduces it to the individual components that are best separated from all the other sound sources affecting the beamformed audio signal. In practice, the process still struggles with highly reverberant and noisy conditions, producing many artefacts in such situations. Therefore, the MVDR, in combination with the spectral masking and Wiener filtering approach, may yield more favourable results without implementing ICA sound source separation in some cases. Even so, it remains a valuable tool for the end user to possess in specific high-noise conditions.

6.2.12 Visualisation and 3D Representation of Sound Propagation

The developed simulation's visual component produces a 3D representation of the environment, sound sources, microphone arrays, and their interactions. This component is integral in understanding the physics of how sound travels through space, how it interacts with various obstacles or surfaces, and how well (or poorly) a microphone array might capture the signals under different conditions. The calibration of these components can make the difference between a "highly directional" array in that it picks up sounds from directly in front and above it, or one that is just "good enough" to deal with sounds reflected around the environment.

6.3 Field Experiments at Exemplar Houses

The final phase of the experimental work involved real-world tests at exemplar houses to validate the system in typical residential environments.

6.3.1 Equipment Needed for Self-Contained Tests

A portable, self-contained test rig was assembled for the house experiments. The rig includes a compact, battery-powered audio recorder, portable amplifiers, and a simplified sensor array

designed for rapid setup in residential settings as shown in Figure 6.9. All equipment was selected for its ease of deployment and reliability in non-laboratory conditions.

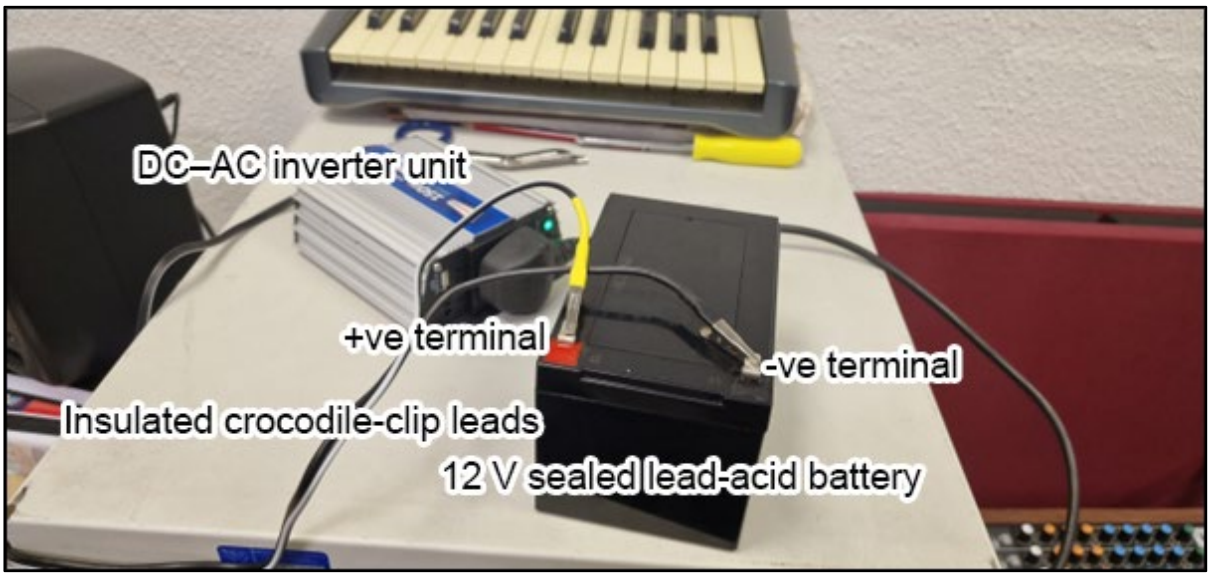

Figure 6.9. Inverter and 12V Battery for mobile recording session.

6.3.2 Initial Speaker and Array Test

This section describes the first outdoor test of the full circular microphone array, using a single loudspeaker as the source. The purpose of this initial experiment was to assess the basic functionality of the custom-built microphones and speaker system in a real-world context, in addition to allowing the user to verify that the array could capture and enhance a target sound source before more complex multi-speaker tests were attempted.

Figure 6.10 provides an overview of the experimental setup, showing the relative positions of the microphone array and loudspeaker during the single-speaker test. This arrangement was chosen to establish a controlled scenario for evaluating direct sound pickup and basic array performance outside of the laboratory environment.

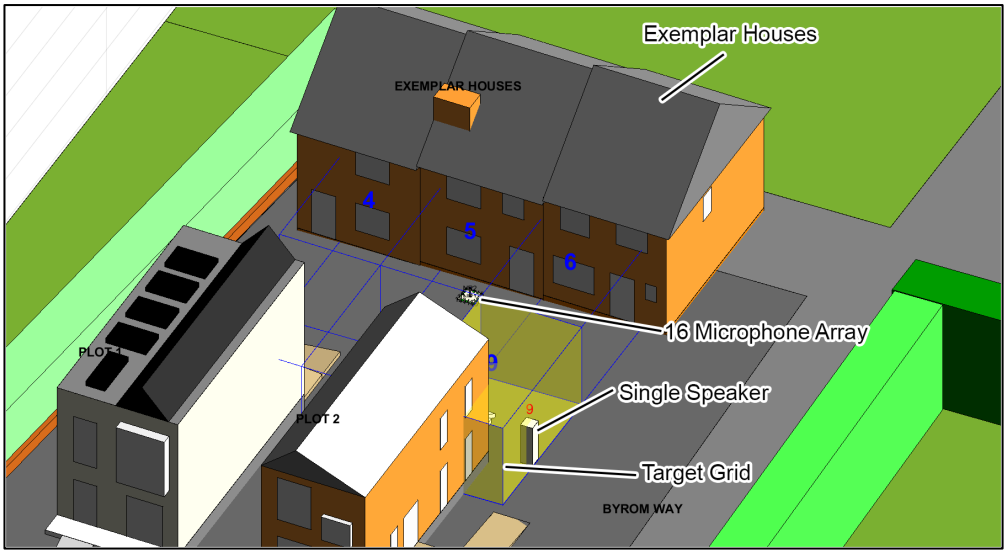

Figure 6.10. Simulated overview of the single-speaker test.

The initial test was performed with the self-contained real-world set-up to assess the performance of the custom-built mics and speakers in a real-world environment, record the performance of the sensor array and judge its ability to accurately capture audio sources and then apply the beamforming algorithm to the real-world outdoor signals. This would be compared to the simulation results and measured for signal clarity after audio zooming.

Although this experimental set-up does not reproduce all aspects of the Church Street incident, it provides a controlled outdoor environment with realistic source distances and building reflections against which the core beamforming behaviour can be evaluated.

In addition to direct sound capture, reflections from nearby surfaces are a key factor influencing array performance in the field. The main expected reflection paths for the single-speaker setup are illustrated in Figure 6.11.

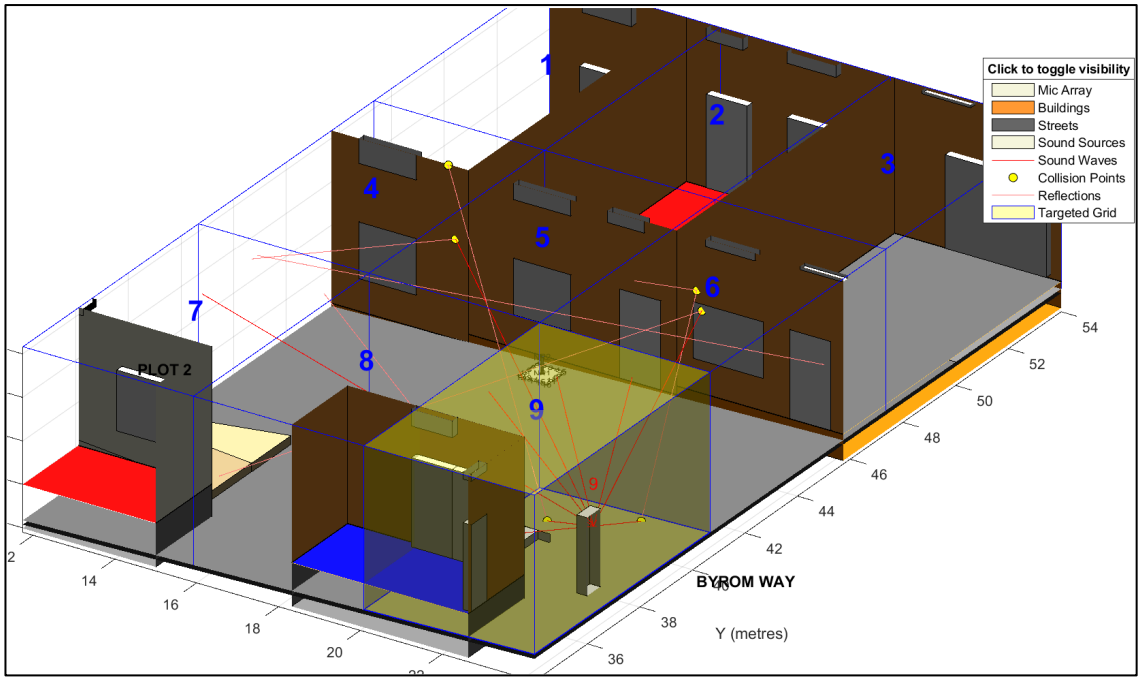

Figure 6.11. Expected reflection paths for the single-speaker test.

By documenting these reflection paths, it is possible to compare the real-world acoustic environment to simulated results and identify potential challenges for beamforming and source separation in practical applications.

6.3.3 Field Test with Nine Speakers

A more extensive test involved recreating the simulation of Exemplar houses with the configuration of nine speakers to record a field test. This test was again designed to test the algorithm against a complex multi-source environment to see if the audio zooming algorithm could isolate and enhance a target signal amidst real-world competing sources. The collection of detailed measurements of inter-speaker interference, spatial accuracy, and overall system performance data from this test would be used to inform further work needed to refine both the hardware configuration and the signal processing algorithms.

6.4 Summary

This chapter explained the elements of a MATLAB -based audio zooming algorithm built around an MVDR beamformer with additional noise reduction and source separation. This method could be an effective solution for audio zooming in the complex acoustic environments encountered in drone surveillance.

The MATLAB -based simulation and implementation of an audio zooming algorithm based on beamforming have been shown in this chapter, starting with the basics of the relevant physics. Beamforming is a signal processing technique used to control the direction of an array of sensors. When it comes to audio signals, beamforming can effectively isolate a sound source, even in a very noisy environment. One major use of audio beamforming is in drone surveillance applications. The chapter described setting up an acoustic simulation within a confined space that accurately mimics real-world conditions.

The beamforming process involves several basic steps. First, the time delays required to form a beam are calculated. This is essentially identifying how much earlier or later a signal will reach each microphone if it is going to travel along the direction in which the microphone array is aimed. The calculation of time delays is most commonly performed using an analytical method. This method assumes that all of the signals were generated in phase and that they will all reach the microphones travelling in the same direction, but with different starting points (the different sound sources) and different ending points (the different microphones).

Moreover, this chapter details the combination of various noise reduction approaches − subtractive noise reduction, spectral subtraction, and others − and applies these techniques to the beamformed signal to produce a clear and intelligible result. Finally, optional source separation techniques for the beamformed signal can be used to remove the remaining mixed sound sources and to achieve the largest possible improvement in the quality of the resulting sound signal.

The next chapter discusses the procedures used in the project's various experiments. These experiments were designed to help evaluate how viable the different methods of audio zooming could be in complex audio environments. The chapter deals with the evolution of the project, MATLAB-based simulation methods and real-life experiments, whose results would validate the simulation output.

Chapter 7: Experimental Results

7.1 Overview

Chapters 4, 5 and 6 presented the approaches to gathering data from the audio zooming experiments. This chapter presents the results from the controlled experiments, simulations, and real-world tests. Each section presents the experimental results and includes a discussion of the findings and their implications, before the chapter concludes with an overall synthesis and benchmarking against related work.

All crime scene and simulation experiments reported in this chapter use a fixed-duration speech sample of 5 **s** (48 kHz, 24-bit) as the target signal, mixed with the relevant interference sources for each environment. Unless explicitly stated otherwise, the same speech excerpt is used across test conditions to ensure that differences in measured SNR and STOI are attributable to the processing chain and environment rather than changes in content, tone or duration.

7.2 Initial Experiments in Sound Booth Environment

In the controlled environment of the isolated sound booth, an initial series of experiments was carried out to establish fundamental baselines of the sound-capturing system and how the system could perform in ideal conditions. There was a series of tests that tested sine wave capture, inverse square law behaviour, wind pressure tolerance, absorption coefficients of the booth and how they all combined to interfere with the microphone array.

7.2.1 Sensitivity and Frequency Sweep Tests

The results of the test carried out in section 4.2.5, where sine waves at predetermined frequencies were played through the nine-speaker setup and recorded by the prototype array are presented below. The resulting signals from the four-microphone array were saved and then analysed in MATLAB. The resulting time-domain waveform and frequency spectrum showed minimal distortion and a consistent amplitude envelope. The measured phase and amplitude data confirmed that the system's delay calculations and synchronisation across channels were accurate, serving as a reasonable baseline for further testing.

The recorded time-domain waveforms closely replicate the shape of the input sine signals, apart from the expected drop in level introduced by the recording chain. Figure 7.1 compares the input sine waves with the corresponding recorded outputs for each test frequency. Cosine similarity

was then calculated. The resulting cosine similarity values (shown in each subplot title) are all close to 1, indicating that the recorded waveforms are consistent with the corresponding input sine waves.

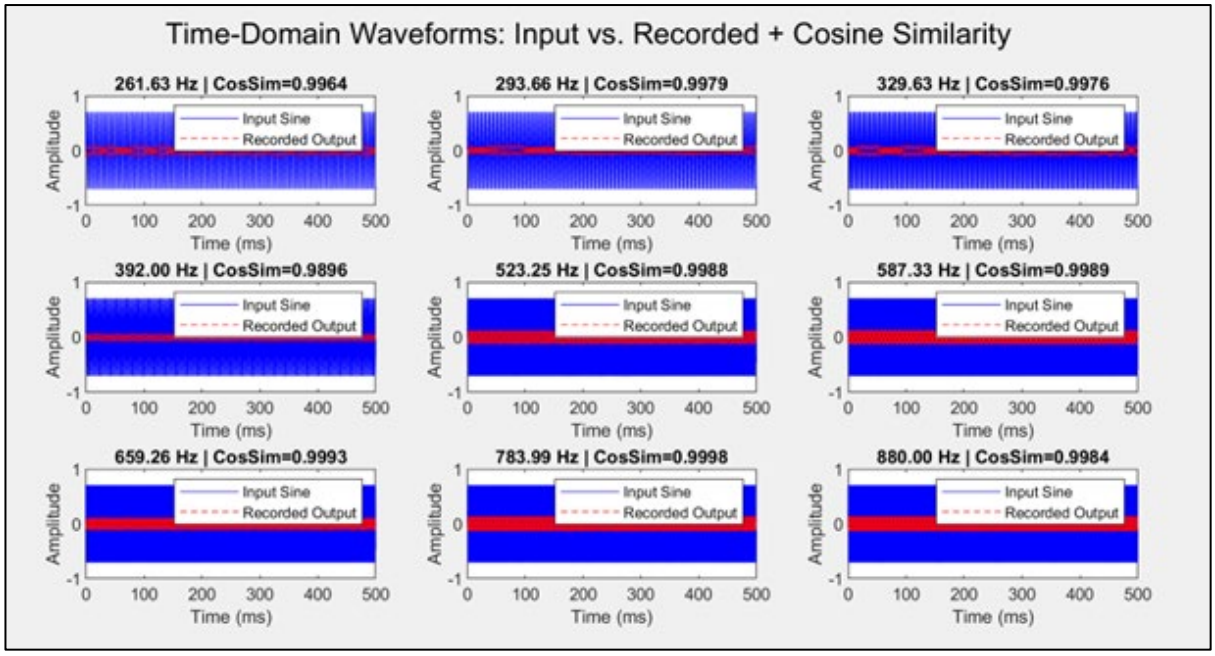

Figure 7.1. Four-microphone recordings (red) overlaid with input sines (blue), showing uniform chain attenuation.

Converting these initial recordings to the frequency domain produces spectra dominated by the intended fundamental frequency. Any harmonic components lie at least 17.7dB below the fundamental and fall beneath the system's noise and distortion floor. The plots in Figure 7.2 show FFT magnitude spectra for each test tone. The corresponding peak magnitudes of the digital input and the recorded microphone signal at the fundamental frequency are listed in Table 7.1, quantifying the attenuation introduced by the loudspeaker–room–microphone path. Magnitude spectra are used rather than power spectral density (PSD) because the stimuli are single-frequency sine waves; the aim was to compare the level of the fundamental and any distortion components, for which an FFT magnitude representation is sufficient.

The lower recorded amplitude reflects the expected attenuation and gain differences of the loudspeaker-room-microphone recording chain, including propagation loss and preamp gain, while the input trace is the original digital stimulus in arbitrary full-scale units; Table 7.1 quantifies this attenuation.

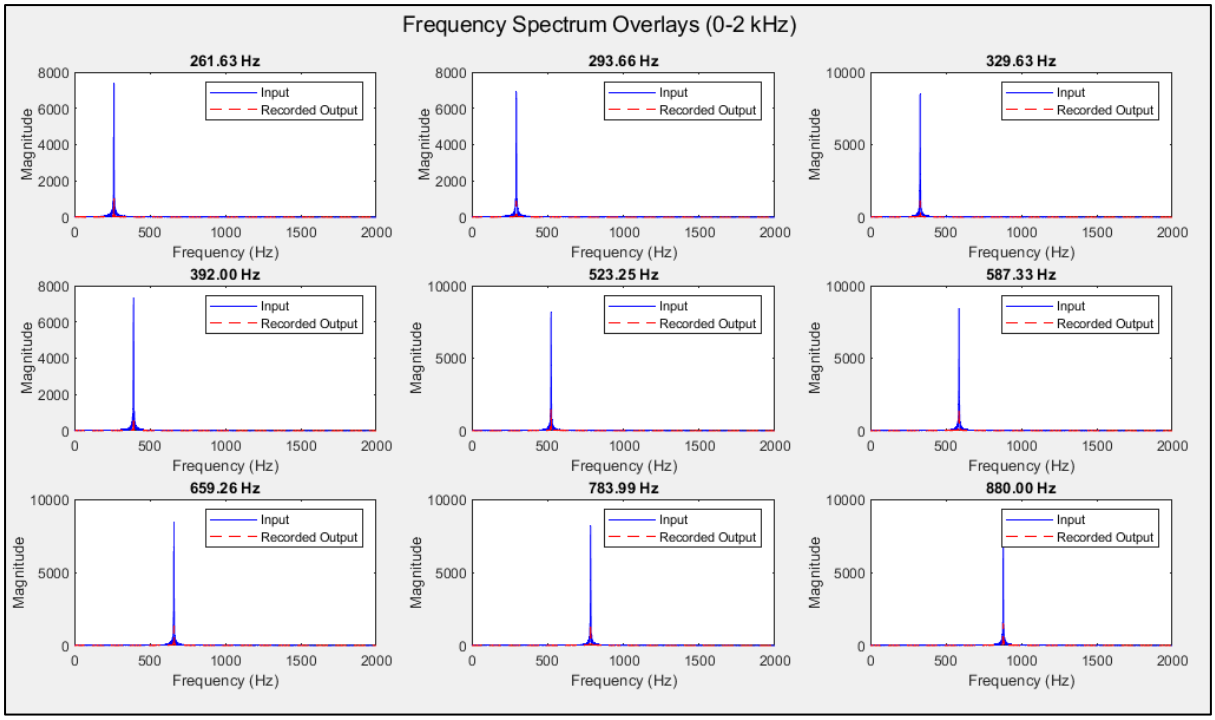

Figure 7.2. Frequency spectrum overlays of the recorded signals.

Table 7.1 summarises the total harmonic distortion (THD) measured at each test frequency; the THD values span from –52.1dB to –17.7dB, confirming that the recording chain behaves linearly over the sweep. Each test frequency was windowed, and MATLAB's 'thd' function was used to obtain the dB values you see in Table 7.1.

Table 7.1. Total harmonic distortion (THD) is measured by the prototype array.

Frequency (Hz)	**THD (dB)**
261.63	-23.308
293.66	-25.042
329.63	-24.721
392.00	-17.703
523.25	-30.861
587.33	-34.731
659.26	-32.669
783.99	-52.087
880.00	-46.329

To evaluate amplitude fidelity across input levels, the recorded peak amplitude was plotted against the known input level and fitted with a straight line. Figure 7.3 shows the harmonic content generated from the system running at room volume measured at 83dBA at the array.

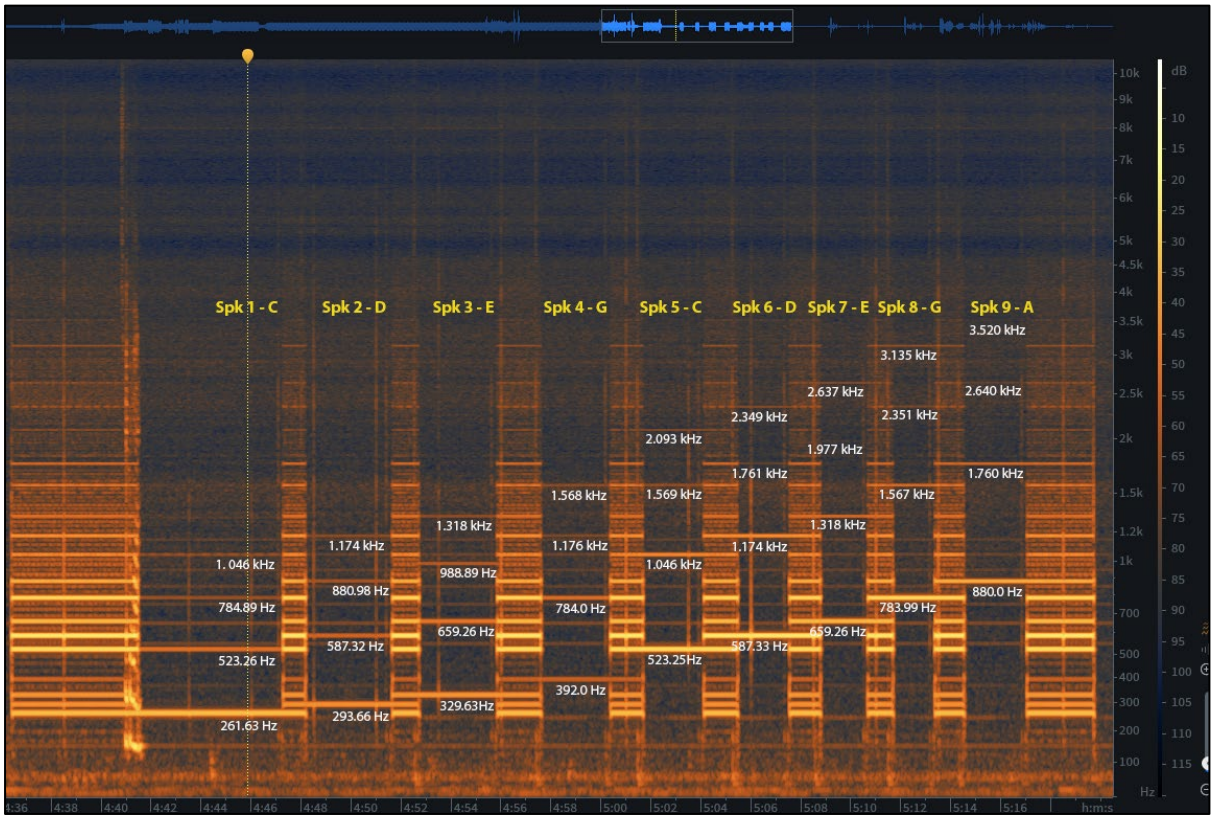

Figure 7.3. Harmonics from nine speakers in the sine wave test.

The recorded amplitude was plotted against the gain staged input levels and fitted a straight line ($R^2 = 0.89$), as shown in Figure 7.4. The less than perfect R^2 level (1 being a perfect calibration) reflects normal experimental scatter rather than a failure of the system and seems an appropriate value.

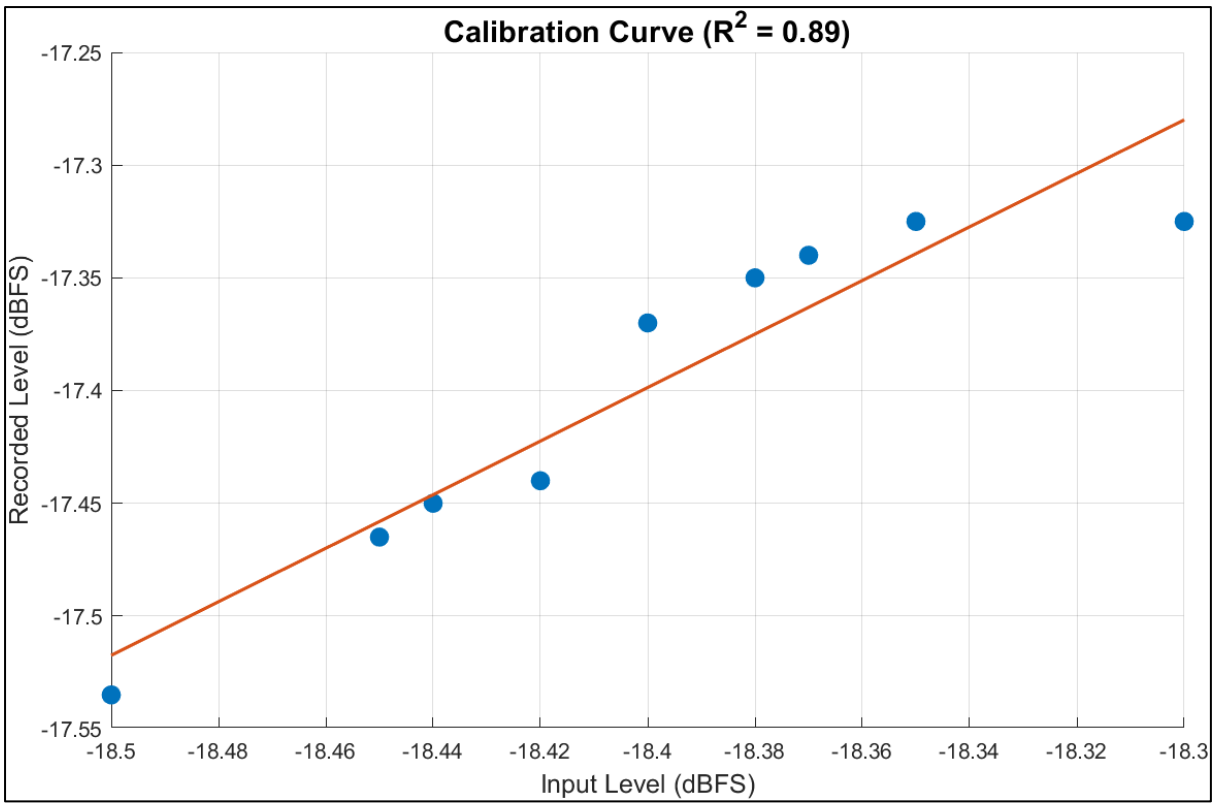

Figure 7.4. Recorded amplitude vs input with a linear fit of (R^2 = 0.89).

In this section, the prototype array is shown to reproduce the input sine signals with a controllable uniform gain loss attributable to the recording chain. The single-tone spectra confirm that any residual harmonics lie below the noise floor, while the nine-speaker tests reveal the expected harmonic display under full volume multi-frequency exposure. Total harmonic distortion remained low across the sweep (from –52.1dB to –17.7dB), and the calibration curve exhibited a predominantly linear amplitude response (R^2 = 0.89), with scatter consistent with normal experimental variability. These findings validate the array's fidelity and provide a solid platform for the beamforming experiments to follow.

7.2.2 Inverse Square Law Test

To verify that the array's measurements adhered to physical acoustic principles, the speaker was positioned at several distances from the array. The results of the test carried out in Chapter 4.2.6 were recorded and presented below.

The sound pressure level (SPL) was recorded at each position, and the data were plotted to confirm that the SPL decreased proportionally with the square of the distance, as predicted by the inverse square law. The experimental results reasonably matched the theoretical model, validating the distance estimation module which was a part of the algorithm.

The time domain waveforms in the DAW in Figure 7.5, show the attenuation of the signal over distance for all microphones. Fluctuations or deviations from the theoretical model can be explained by the environmental factors of testing in a university corridor and the resulting reflections coming from the nearby materials (glass, metal), resulting in a temporary shift in amplitude.

The slightly steeper decay for the small capsule (green waveform) is expected as it has a smaller diaphragm than the other microphones, and therefore its sensitivity is lower overall. This behaviour is also consistent with its higher self-noise and omnidirectional pattern: as the direct sound falls, the measurement includes proportionally more room noise and reverberant energy, so the measured peak level reduces more quickly than for the more directional microphones.

Figure 7.5. DAW results showing signal attenuation with distance.

As seen in Figure 7.6, Sound pressure level (SPL) was measured in dB against distance for four microphones (Audio Technica, Rode GT2, SM57, Small Capsule Electret), normalised to 0dB at 1m. The dashed line indicates the theoretical inverse square law ($-20 \cdot \log_{10}(d)$). All microphones track the theoretical slope at short to medium distances, with deviations at longer distances reflecting the environmental noise floor and any system limitations. To account for this, all subsequent calibration and analysis used only the distance range over which the microphones followed the theoretical 6dB per distance-doubling within approximately ±1dB. Data points at larger distances, where the SNR was low and the measurements were dominated by background noise and residual reflections, were excluded from the inverse-square fit and are shown only to indicate the practical operating limit of the measurement setup in this environment.

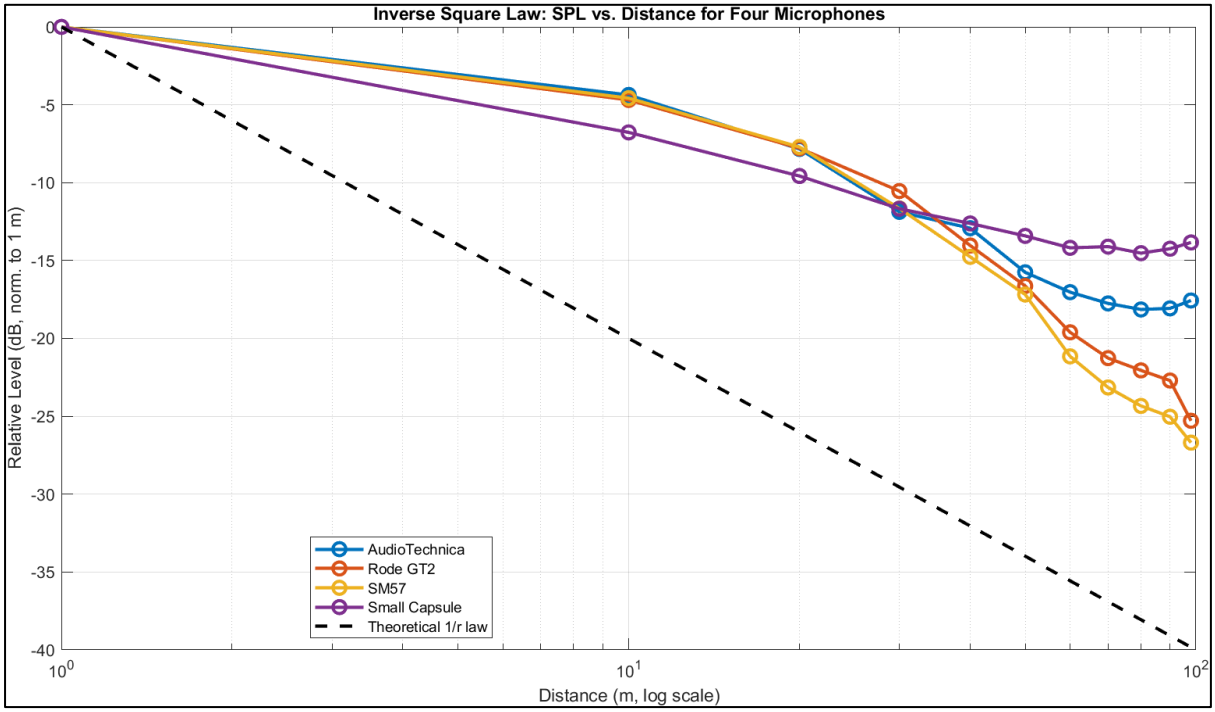

Figure 7.6. Measured SPL versus distance for four microphone types.

The inverse square law measurements validated the prototype microphone array's ability to accurately capture distance-related amplitude changes, approximating the theoretical prediction, but allowing for aspects introduced by the environment involved. Although environmental reflections and noise introduced deviations at greater distances, the overall adherence to the expected inverse square relationship suggests the system would be suitable for use with an audio zooming algorithm. The small capsule omni-directional electret

microphones performed well and picked up signals at further distances performing better than some of the more expensive microphones. This is likely due to the omni design not rejecting reflections from the rear, unlike the other microphones tested, but the lightness in weight, in addition to the performance, makes it a sensible choice for the array.

7.2.3 Wind Pressure Test

Simulated wind was introduced into the booth by operating a calibrated fan adjacent to the array. Microphone signals were monitored for fluctuations caused by wind pressure. This fan-based arrangement does not reproduce the spatially varying, turbulent wind conditions experienced by a drone in outdoor flight. Instead, it provides a controlled, repeatable airflow for comparing microphone types and array configurations under steady loading. The results should therefore be interpreted as relative indicators of wind susceptibility, not as a full model of in-flight performance.

Figure 7.7. displays the wind pressure test results using the SM57 microphone. The images show the experimental setup and microphone positioning relative to the fan, with and without the windscreen. Underneath are the time-domain waveforms illustrating significant amplitude increases in wind noise when the microphone is aimed at the outer rim of the blades. Underneath the waveforms are the corresponding spectrograms highlighting prominent low-frequency wind noise when the microphone is aimed at the blades, significantly reduced when the microphone is aimed at the centre of the fan or outside of the blades.

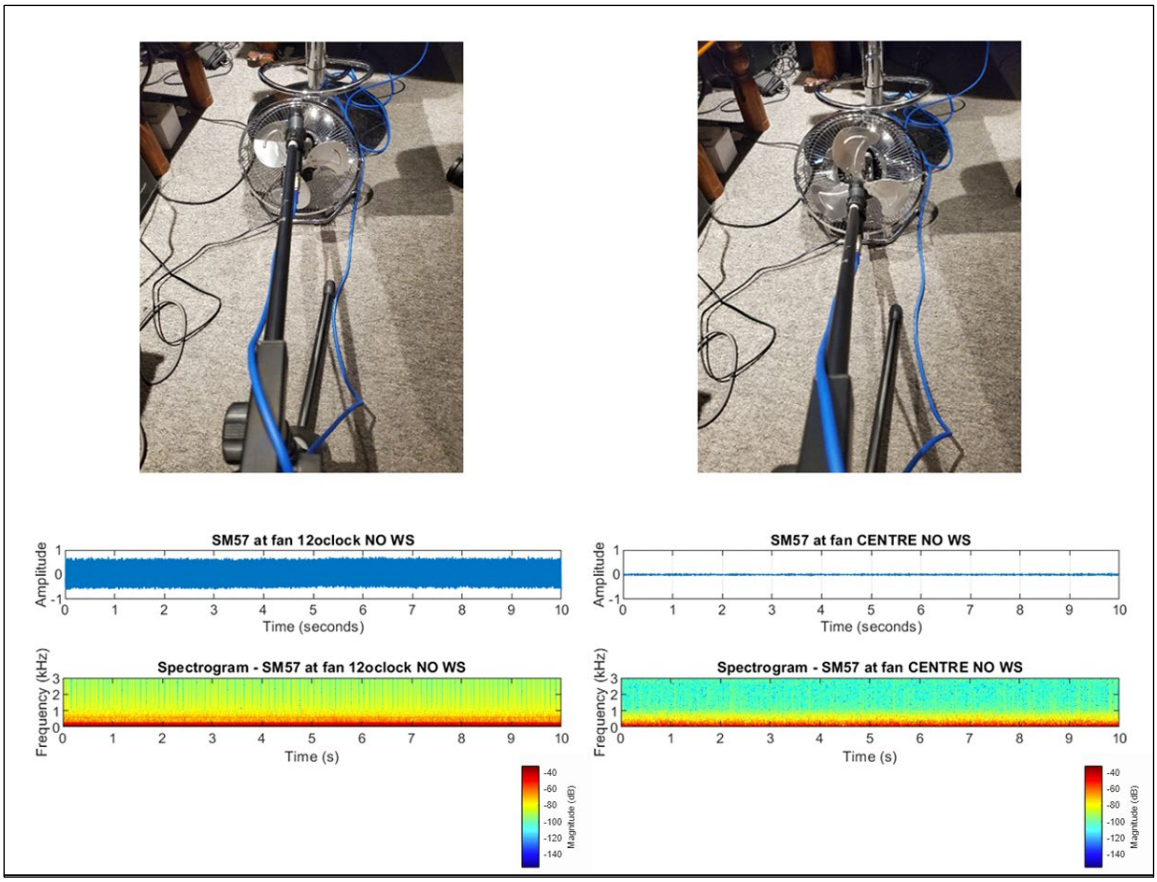

Figure 7.7. Wind pressure test results using the SM57 microphone. Colour scale shows spectral magnitude (blue = low energy, red = high energy, in dB relative to the maximum).

Figure 7.8. shows the sweep test results showing amplitude and spectral content of the SM57 output as the microphone is moved past the fan blades. A clear reduction in low-frequency noise is observed when the microphone is positioned in the low-pressure region near the fan centre. “NO WS” denotes that no additional windshield was fitted to the microphone; the capsule was exposed directly to the airflow. This condition provides a worst-case reference for wind-induced noise.

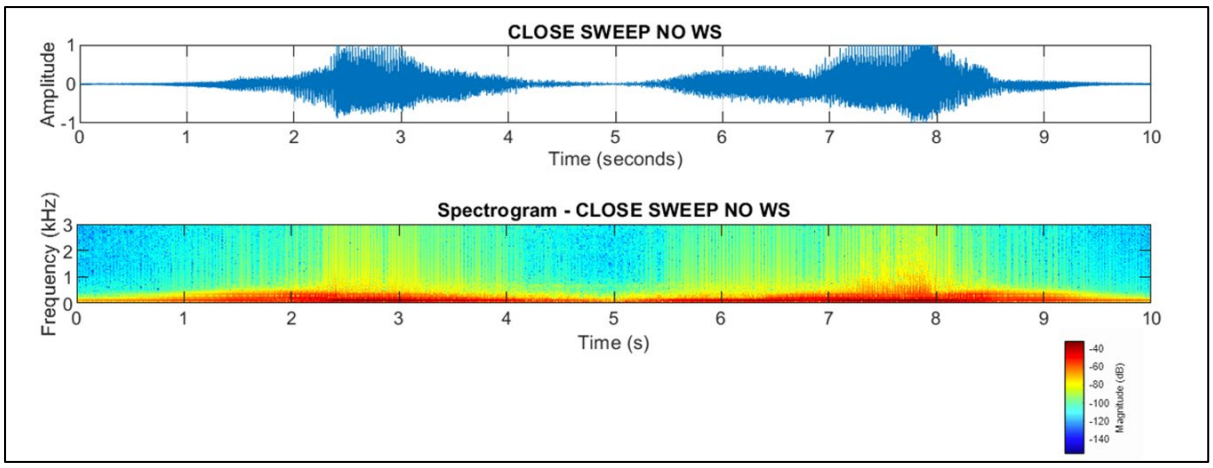

Figure 7.8. Sweep test to find the area of low pressure near the centre of the fan blades.

These wind pressure tests demonstrate that wind-induced noise can significantly degrade microphone performance, especially at low frequencies and in the absence of protective measures such as careful array placement or windscreens. Identifying optimal microphone positioning, such as specifically locating low-pressure regions near the fan centre, effectively reduces this noise before any filtering is needed. Locating these low-pressure zones, particularly near the fan centre, is important for practical deployment: placing capsules in these regions reduces wind-induced rumble and the risk of overloading the diaphragm before any electronic filtering or beamforming is applied. This provides a worst-case reference for exposed capsules and a baseline for the pressure variations likely to be encountered in real flight conditions. Additionally, these findings further confirm the efficacy of beamforming techniques in isolating target signals from wind interference, ensuring robust audio capture even under challenging environmental conditions.

7.2.4 Absorption Test

Absorptive materials (carpets and acoustic foam panels) were placed in the sound booth to help evaluate the system's response to absorption. These experiments also informed the "Sabine absorption" section of the MATLAB simulation, which uses the classical Sabine room-acoustics model to approximate how sound energy decays in a space as a function of the absorption of its surfaces. The recorded signals showed a reduction in reflected energy consistent with the known absorption behaviour of the materials. This confirmed that the system and simulation correctly account for material absorption in such environments and that the captured impulse responses are representative of the sound booth's acoustic characteristics.

Table 7.2 summarises the measured reduction in sound pressure levels provided by the carpet tile across various frequencies and stimulus types.

Table 7.2. Measured reduction in sound pressure level (dB) by the recycled carpet tile for each stimulus and frequency.

Frequency (Hz)	White Noise	Pink Noise	250Hz	500Hz	2kHz
250	–2.0	–1.0	–4.3	0.0	–0.1
500	–3.2	–2.8	–3.7	–3.2	–1.0
1,000	–7.1	–10.1	–5.7	–1.1	–1.0
2,000	–14.4	–14.5	–11.7	–10.5	–17.2
4,000	–25.7	–21.9	–16.7	–14.2	–20.9
8,000	–32.9	–26.6	–7.2	–0.7	–20.7
16,000	–26.1	–26.1	–0.4	–0.2	–13.3

Table 7.2 shows that the carpet tile provides substantial sound pressure attenuation at higher frequencies, particularly from 2kHz upwards, while offering limited absorption in lower-frequency ranges.

Figure 7.9 shows the SPL reduction provided by the carpet tile as a function of frequency, as measured with white noise, pink noise, and pure tones. The values correspond with Table 7.2, confirming that the greatest attenuation occurs at high frequencies (4–16kHz), while low-frequency absorption remains modest.

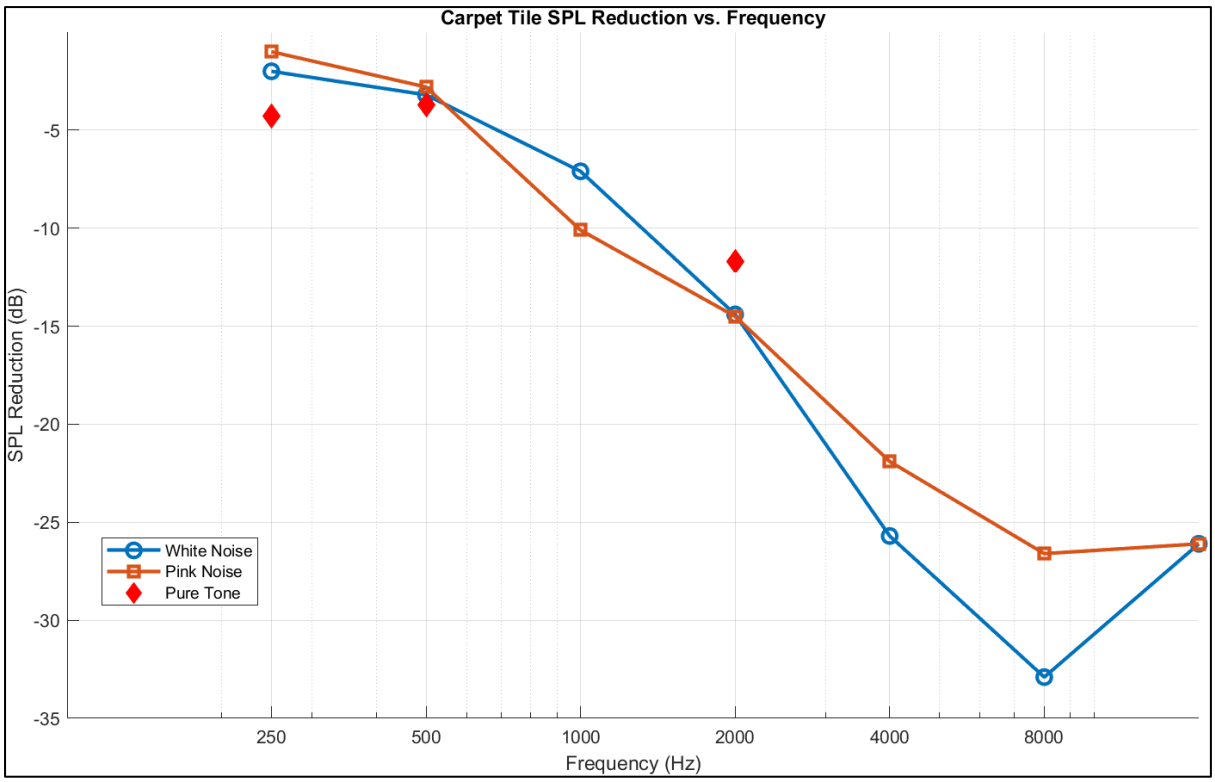

Figure 7.9. SPL reduction measured across frequencies for white noise, pink noise, and pure tones.

The absorption tests have shown the expected reduction of reflected acoustic energy, which is consistent with theoretical Sabine predictions. High-frequency attenuation provided by the carpet tile was significant, while low-frequency absorption was comparatively modest, matching known characteristics of such absorptive materials. These results confirm that the recording setup accurately represents acoustic treatment effects, ensuring that subsequent measurements and analysis reflect realistic environmental acoustics.

7.2.5 Recording Session

A comprehensive recording session was conducted in the sound booth environment to evaluate the performance of the array and the initial delay-and-sum beamforming and noise reduction algorithms. Nine speakers were used to generate a series of individual tones, summing to creating a musical chord ($C^{6/9}$), creating a realistic yet repeatable audio scenario. The aim of this session was to quantify any improvements in signal quality and intelligibility offered by beamforming techniques.

Figure 7.10 illustrates the results of the simulated SNR and beamforming test, in which the a four-microphone delay-and-sum (DAS) beamformer was steered towards a target signal in the presence of a single strong broadband interferer. In this scenario, the input SNR at the reference microphone was approximately -5.0dB. After beamforming, the output SNR improved to approximately -2.0dB, corresponding to a beamformer gain of approximately +3.0dB.

This result is representative of the SNR improvements achievable with a four-microphone delay-and-sum array under challenging noise conditions and demonstrates the ability of spatial filtering to enhance the target signal above the background interference. The zoomed time-domain panel and the PSD/spectrogram views show that the output remains noise-corrupted, but with a measurable reduction in broadband noise level consistent with the reported SNR gain.

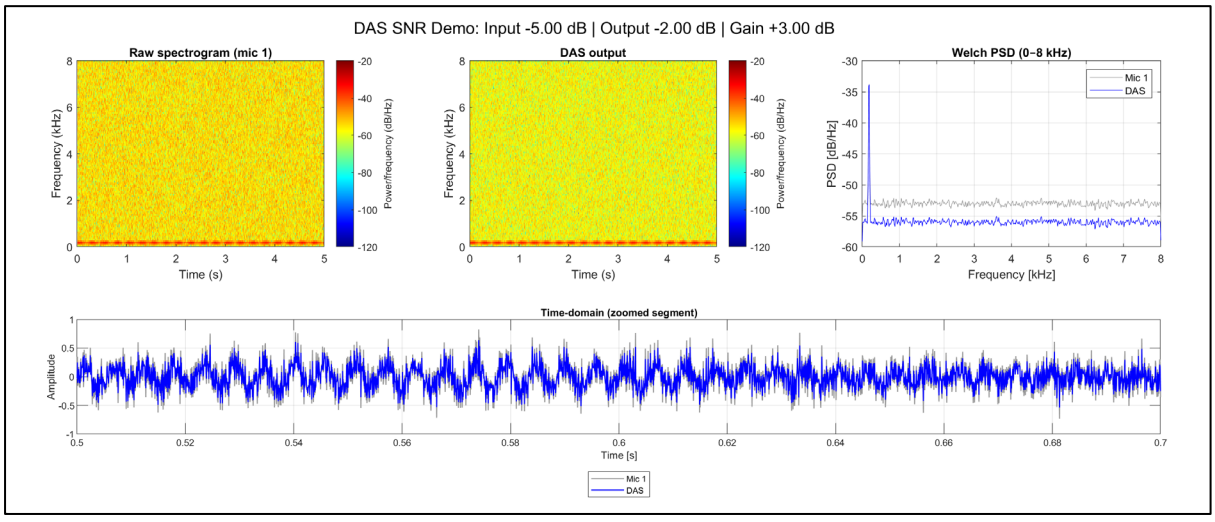

Figure 7.10. Spectrogram and waveform of raw mix vs beamformed output showing SNR gain.

The recording session results confirm the practical effectiveness of the beamforming approach within controlled yet realistic audio environments. The tests demonstrate that significant noise suppression and improved signal clarity can be achieved, highlighting the suitability of beamforming techniques for real-world acoustic scenarios.

7.3 Further Experiments within the Sound Booth Environment

Following baseline performance validation, further experiments were carried out to assess the system's stability and consistency over extended periods within the sound booth environment.

An additional recording session, using the newly established 16 microphone array with small capsule omnidirectional electret microphones and the multi-source configuration, aimed specifically at improving temporal performance and consistency under prolonged operating conditions.

7.3.1 Recording Session

A second, longer-duration recording session was performed with the same multi-source setup. This session focused on evaluating temporal stability and consistency. SNR measurements and spectral analyses across the duration confirmed that the system maintained robust performance over extended recordings, and subtle changes in environmental noise were successfully filtered. Figure 7.11 illustrates the outcomes of the simulated beamforming session, specifically targeting loudspeaker 1. The initial input SNR was approximately -4.3dB, while the output SNR after beamforming improved to just under 0.26dB. This corresponds to a beamformer gain of approximately +4.04dB, demonstrating a substantial and sustained reduction in background noise relative to the target signal.

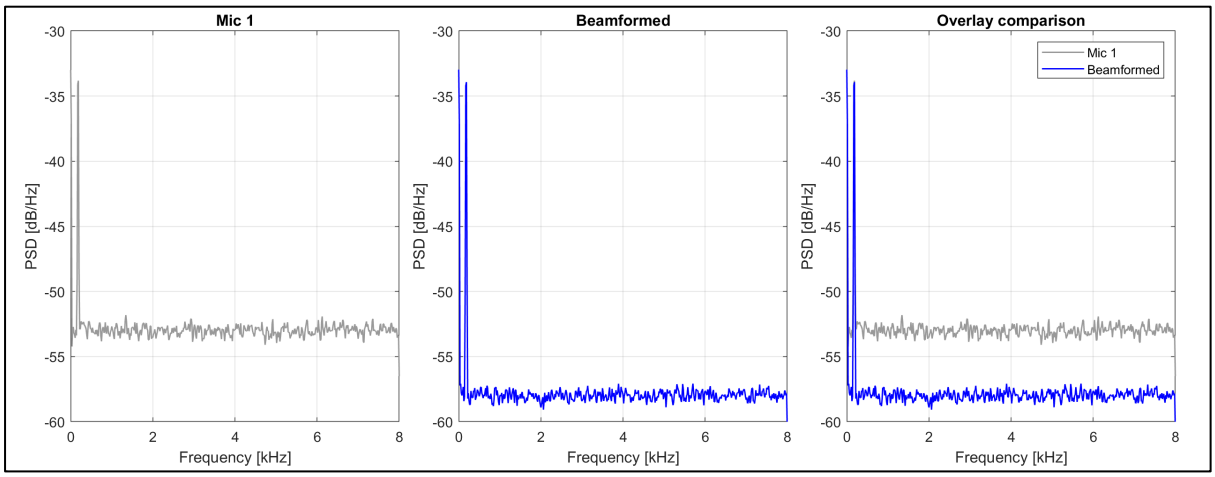

Figure 7.11. Results of beamforming towards loudspeaker 1 using the 16-microphone square-array configuration.

Table 7.3 summarises the measured time delays between the loudspeaker (Speaker 1) and each microphone in the 16-microphone square-array setup. These measurements confirm consistent microphone placements, an essential prerequisite for stable beamforming performance.

Table 7.3. Time Delays between Microphones and Speaker 1 in the Sound Booth.

Microphone	Delay in Samples	Delay in time (ms)	Microphone	Delay in Samples	Delay in time (ms)
1	400	8.33	9	353	7.35
2	381	7.94	10	433	9.02
3	363	7.56	11	367	7.64
4	346	7.21	12	445	9.27
5	328	6.83	13	428	8.92
6	410	8.54	14	412	8.58
7	340	7.08	15	397	8.27
8	421	8.77	16	382	7.96

Table 7.4 presents the sound profile data for each loudspeaker, detailing sound pressure levels (SPL), azimuth and elevation angles, and beam widths, demonstrating the carefully controlled directional characteristics of the audio sources used.

Table 7.4. Sound Profile for each Speaker

Speaker	dB_{SPL} at 1 meter	Azimuth Angle (°)	Elevation Angle (°)	Horizontal Beam Width (°)	Vertical Beam Width (°)
1	80	210	10	10	10
2	80	180	10	10	10
3	80	150	10	10	10
4	80	270	10	10	10
5	0	90	10	10	10
6	80	45	10	10	10
7	80	330	10	10	10
8	80	0	10	10	10
9	80	30	10	10	10

These further experiments undertaken in the sound booth confirm the temporal stability and consistency of the beamforming system over prolonged periods. Stable SNR improvements and minimal spectral variations underscore the reliability of the array and processing methods.

The accurate measurement of microphone delays and consistent loudspeaker profiles reinforce the reproducibility and robustness of the experimental setup, validating the array's effectiveness for realistic and sustained audio scenarios.

7.4 Discussion: Sound Booth Environment

The four-microphone booth trial established a baseline, raising SNR from -4.99dB to -1.99dB, a gain of +3.01dB after delay-and-sum beamforming. Although modest, the improvement verified delay estimation accuracy and microphone synchronisation. Subsequent tests with the sixteen-microphone square array yielded a beamformer gain of +4.04dB and preserved detail across extended recordings. These figures demonstrate that increasing the sonic aperture and number of microphones translates into increased directivity and improved noise rejection, matching expected results in theory (Qualcomm Technologies, 2021).

7.5 Simulation of Sound Booth Environment

In parallel with physical experiments, a virtual simulation of the sound booth environment was created using MATLAB. This allowed precise replication of the booth's acoustic characteristics, including dimensions, reflectivity, and absorption properties. The simulated environment provided virtual microphone signals for direct comparison with physical recordings, enabling validation and refinement of acoustic and beamforming models.

7.5.1 MATLAB Algorithm

A virtual sound booth was constructed in MATLAB to simulate the experimental conditions from the earlier physical tests. The aim was to test algorithm performance in a fully controlled environment, allowing simulation of fault conditions, directional beamforming scenarios, and array robustness under real-world signal mixtures. In this simulation, the audio scene consisted of a harmonically complex music bed and a voice source located at Grid 7, which acted as the target. The simulation also included a deliberate microphone failure to test the algorithm's resilience.

Figure 7.12 compares polar response patterns of the 16-microphone array under different beamforming conditions. The purple lobe shows the polar pattern when beamforming is applied to the 16-microphone array and directed toward the voice source (Grid 7). The grey lobe shows the effect of simulating a microphone failure to reduce the array to 13 microphones. While

some degradation is visible, the main lobe still points toward the target, indicating that the algorithm maintains directional integrity even with three failed sensors.

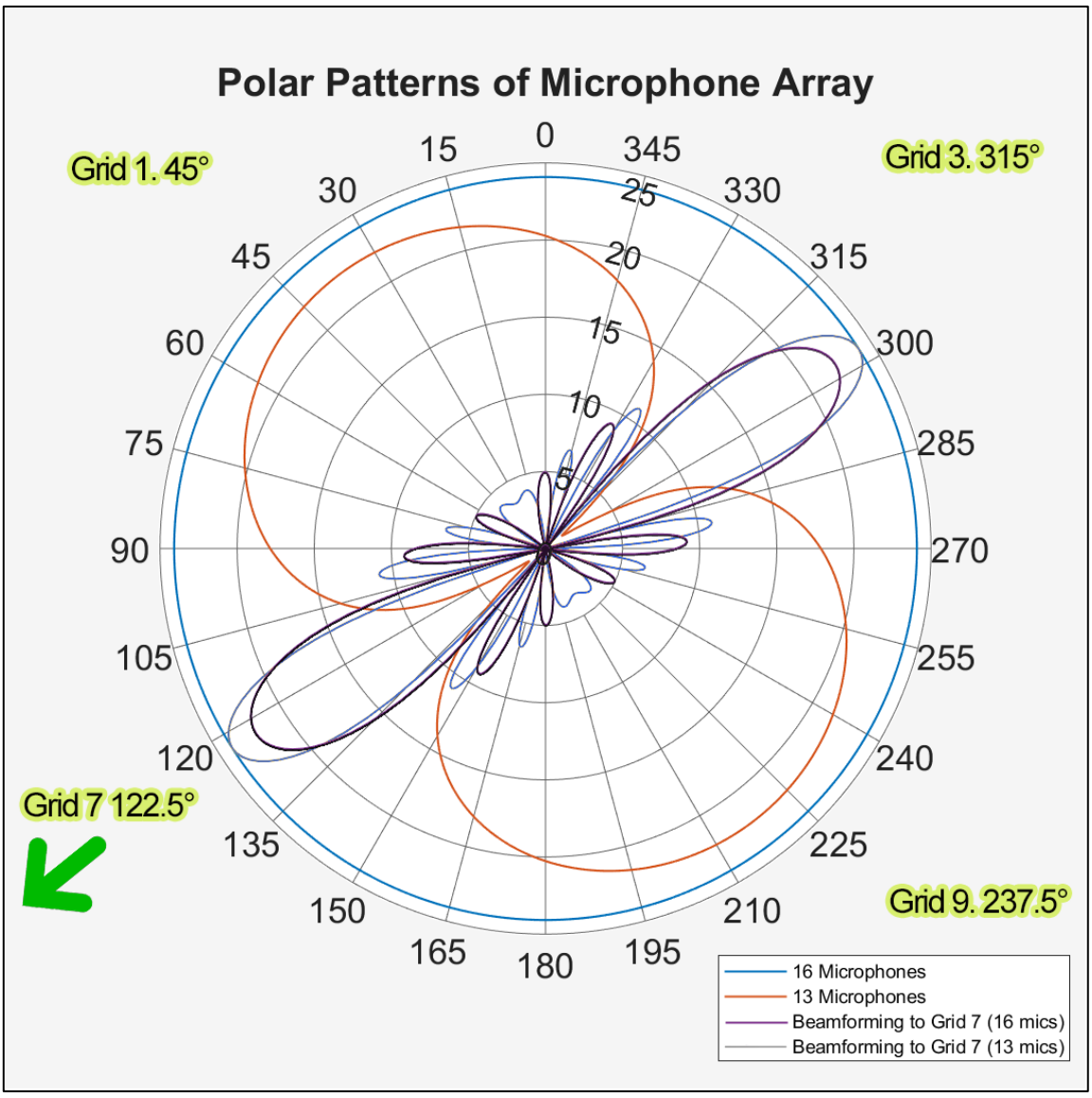

Figure 7.12. Comparison of polar patterns of the 16 microphone array with and without beamforming to a Grid and simulated failure of 3 sensors.

Figure 7.13 provides a direct comparison between beamformed outputs using 13 microphones versus the full 16-microphone configuration. Despite the reduced number of sensors, the waveform remains largely intact, with only minor spectral loss. This result demonstrates the robust nature of the beamforming algorithm and confirms that graceful degradation occurs when some elements of the array are lost.

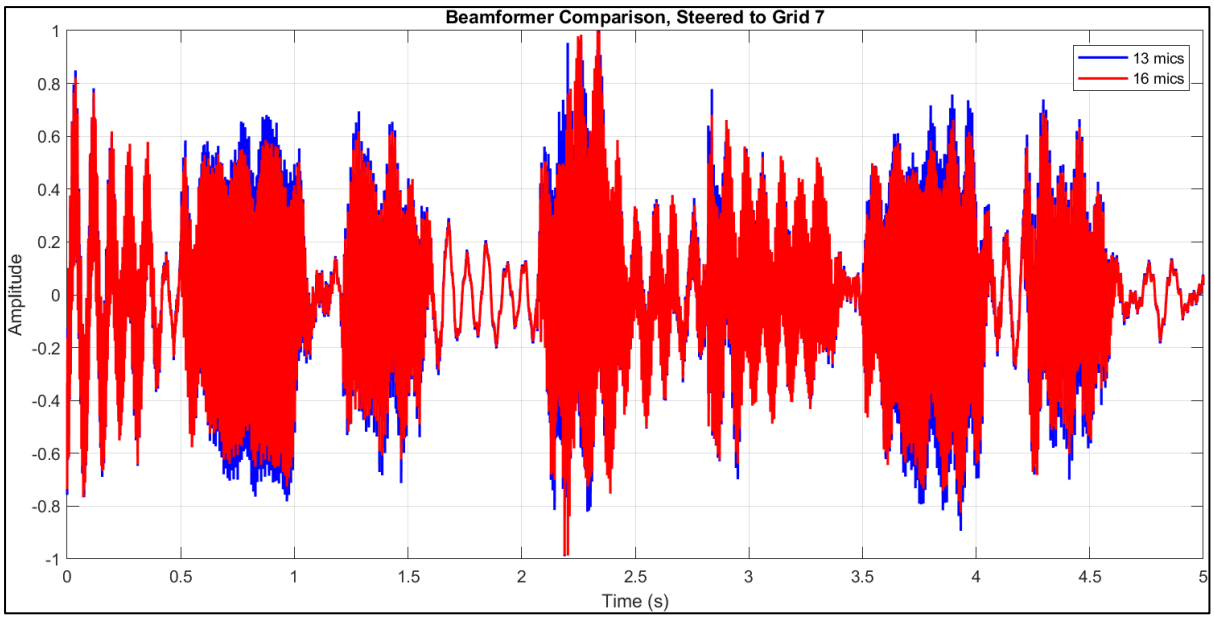

Figure 7.13. Waveform comparison between 13-microphone and 16-microphone beamformed outputs. Minimal degradation confirms the robustness of the algorithm.

Figure 7.14 displays the final output from the simulated beamforming test, including both waveform and spectrogram. The target voice remains prominent and intelligible despite the presence of background music and simulated faults. The spectrogram confirms that key speech bands are preserved and that musical interference is significantly reduced.

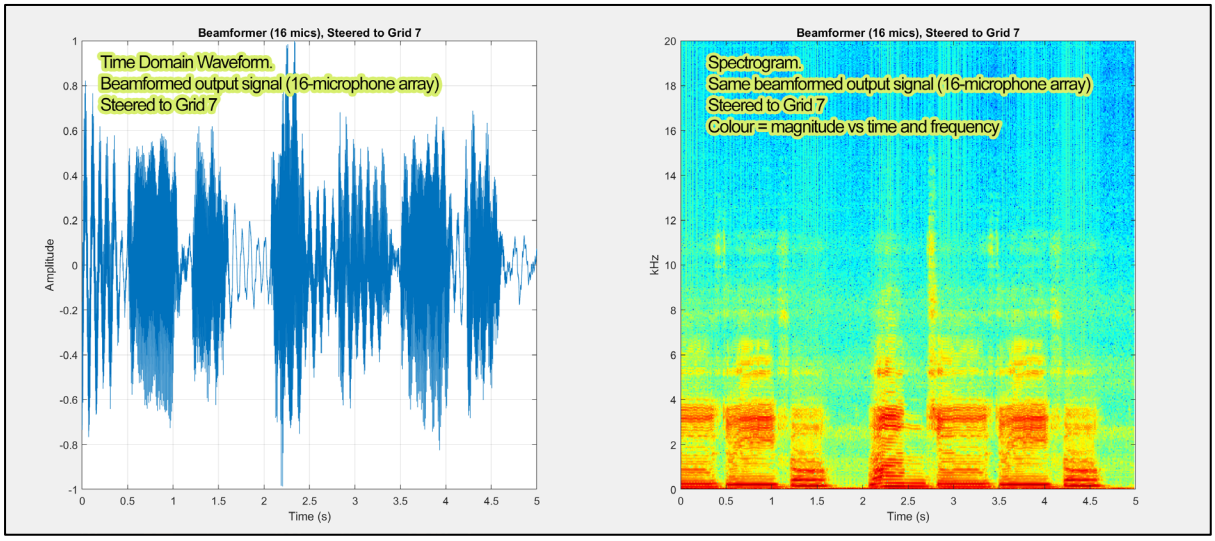

Figure 7.14. Final beamformed output in the simulated booth.

This simulation demonstrates the beamforming algorithm's ability to maintain performance under fault conditions and varying array configurations. The target voice signal remained intelligible, and directional focus was preserved, even with a reduced number of microphones and a simulated fault. These results confirm that the system is not only functional under ideal conditions but remains robust and reliable when challenged, making it suitable for practical deployment in real-world applications.

7.6 Simulation of Crime Scene

To assess the practical limitations and capabilities of the beamforming algorithm in a realistic environment, a simulated urban crime scene was developed in MATLAB. This model aimed to capture the acoustic complexity of an open-air city centre location, including variable reflective surfaces, different absorption properties, and the irregular geometry typical of urban settings. Although the full test program was later shifted to the exemplar houses at John Moores University for practical reasons, the results presented here reflect the initial phase of the city-centre simulation.

7.6.1 MATLAB Algorithm

The MATLAB model was extended to account for multiple overlapping reflections, frequency-dependent absorption, and a non-uniform arrangement of virtual microphone and source positions. This allowed for the generation of highly complex impulse responses and provided a meaningful test of the beamforming algorithm's ability to function in the presence of significant interference and reverberation.

Figure 7.15 illustrates the beamformer's magnitude response as a function of azimuth angle, steered to different target grids within the 3D model. The 3D plots capture the main lobe's behaviour and reveal how well the algorithm can localise a target signal in a challenging, multi-reflective acoustic environment. The results demonstrate that, even in a highly reverberant model of an urban space, the main lobe remains well-defined, and the beamformer is able to focus energy in the direction of the target sound.

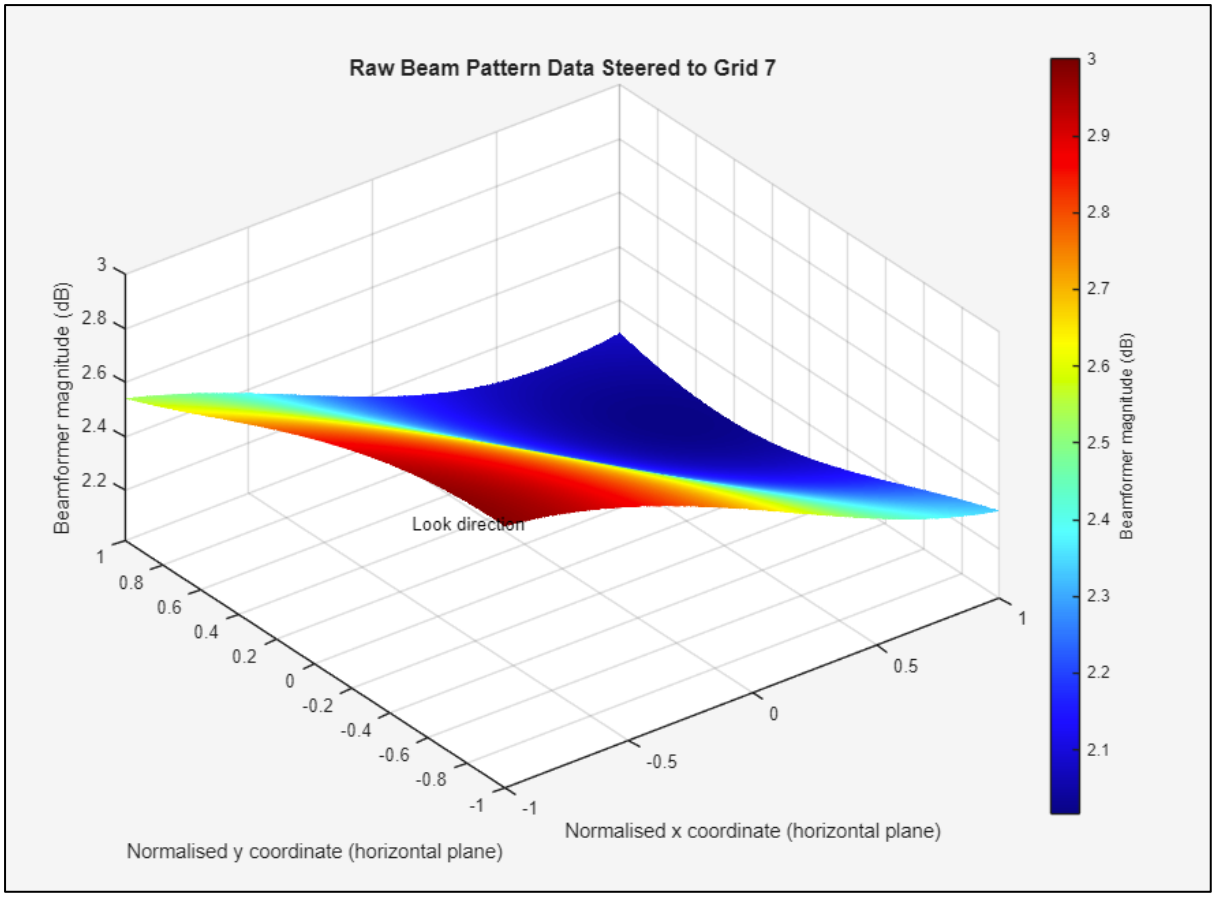

Figure 7.15. Magnitude response of the beamformer as a function of azimuth angle, shown in a 3D plot.

Figure 7.16 presents the polar response of the beamformer when steered to a specific grid. This view confirms that the expected directivity is preserved, with the main lobe aligned to the target and off-axis rejection still evident despite the complexity of the simulated scene.

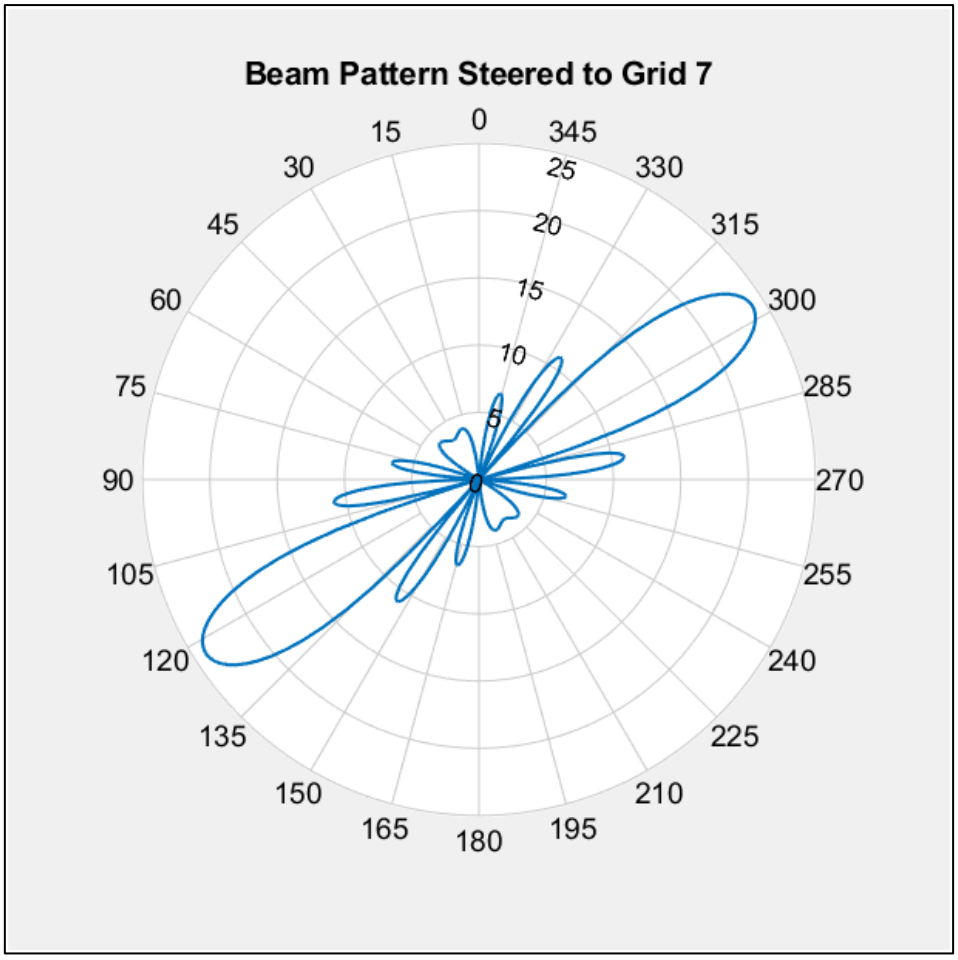

Figure 7.16. Polar plot of the steered beamformer's directivity pattern in the urban simulation. The main lobe remains focused on the target grid, with off-axis attenuation.

The MATLAB algorithm was extended to include multiple reflective surfaces, variable absorption coefficients, and irregular room geometry typical of an urban setting. This simulation generated complex impulse responses with multiple reflections and reverberation tails. The results confirmed that our beamforming algorithm could still isolate target signals even when the acoustic scene included significant interference and overlapping reflections. Although the scope of this urban crime scene simulation was necessarily limited, the results confirm that the developed beamforming algorithm is capable of isolating target signals even under complex acoustic conditions. The retention of clear main lobe focus and off-axis suppression, as shown in both magnitude and polar plots, supports the use of this approach for challenging real-world forensic audio applications. These preliminary tests produced a set of observations that served as a springboard for the more focused experiments later carried out at the exemplar house test sites.

7.7 Simulation of Exemplar Houses

Once the City centre, urban scenario had been explored, the study moved to a refined MATLAB model that reproduced experimental residential houses on the Liverpool John Moores University campus. These "Exemplar Houses" served as an accessible base within easy reach of the laboratory and presented acoustic conditions that mirror those found in real surveillance work.

7.7.1 MATLAB Algorithm

The upgraded simulation specified the exact floor plans, surface finishes, and construction materials used for each building, allowing the software to generate realistic spatial impulse responses. Figure 7.17 shows the circular microphone array used in these tests; a second, purpose-built array for noise reduction appears beside it. Examining both configurations provided a direct comparison between directional accuracy and noise-suppression effectiveness.

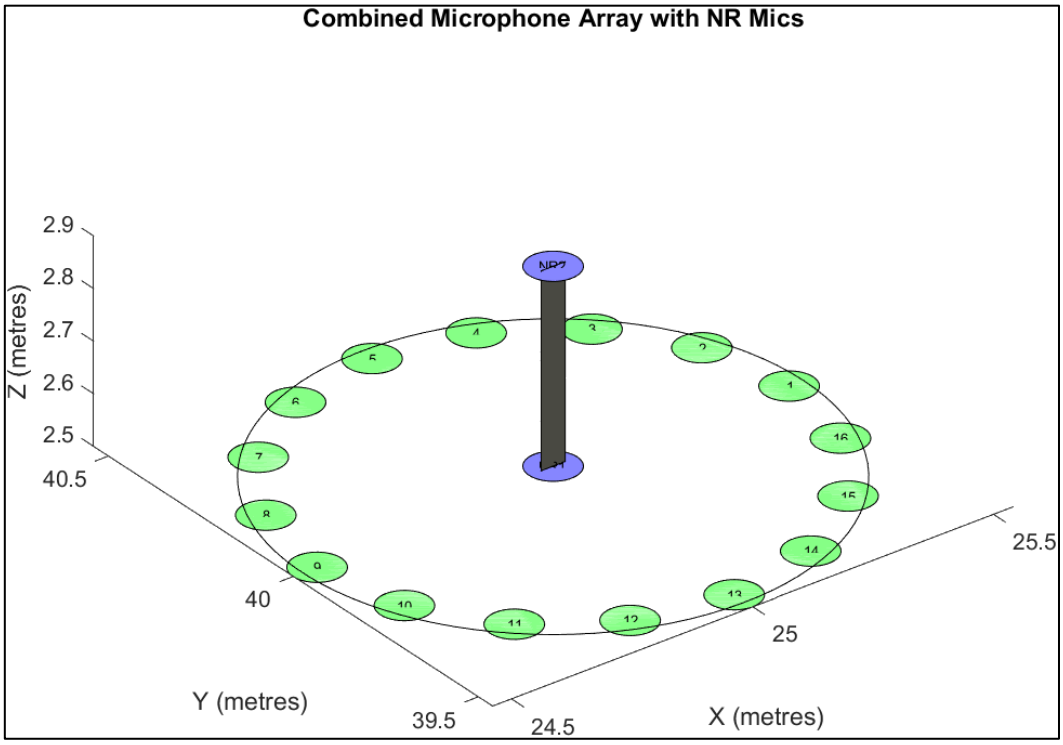

Figure 7.17. Circular Microphone Array with additional Noise Reduction Array.

Figure 7.18 illustrates example waveforms captured from the circular array within the simulated house environment. The recordings demonstrate how spatial arrangement influences the captured signal and highlight the presence of reverberation and noise in a typical domestic space.

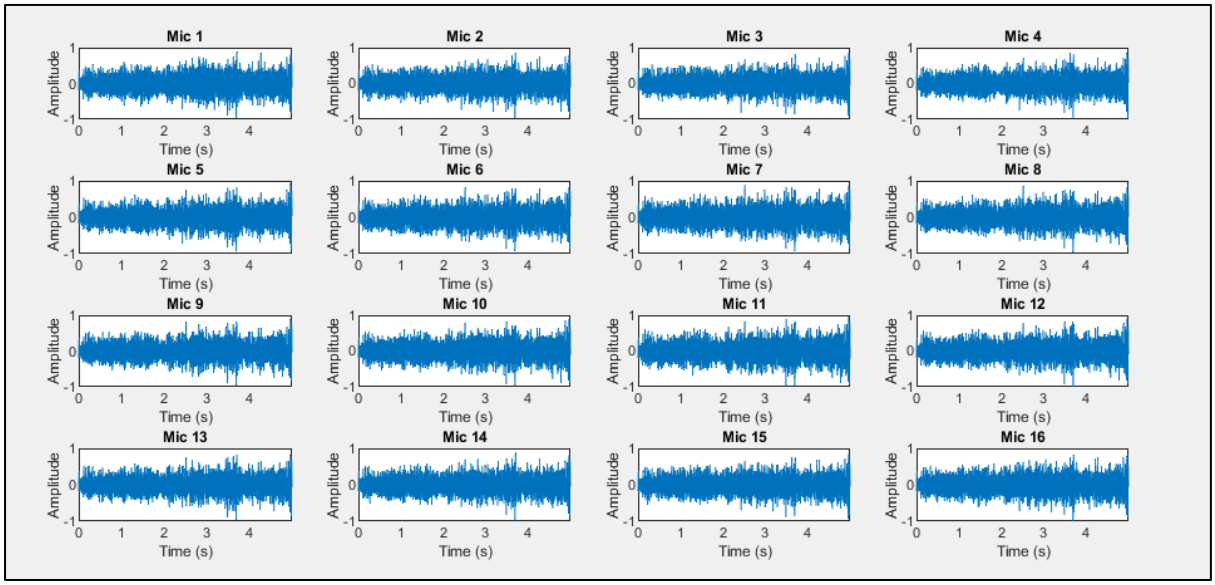

Figure 7.18. Waveforms of captured audio from the circular microphone array.

Figure 7.19 presents the beam pattern of the circular array when steered to Grid 1, illustrating a strong and well-focused main lobe. The result confirms that the array and algorithm can effectively target specific locations in a cluttered acoustic environment.

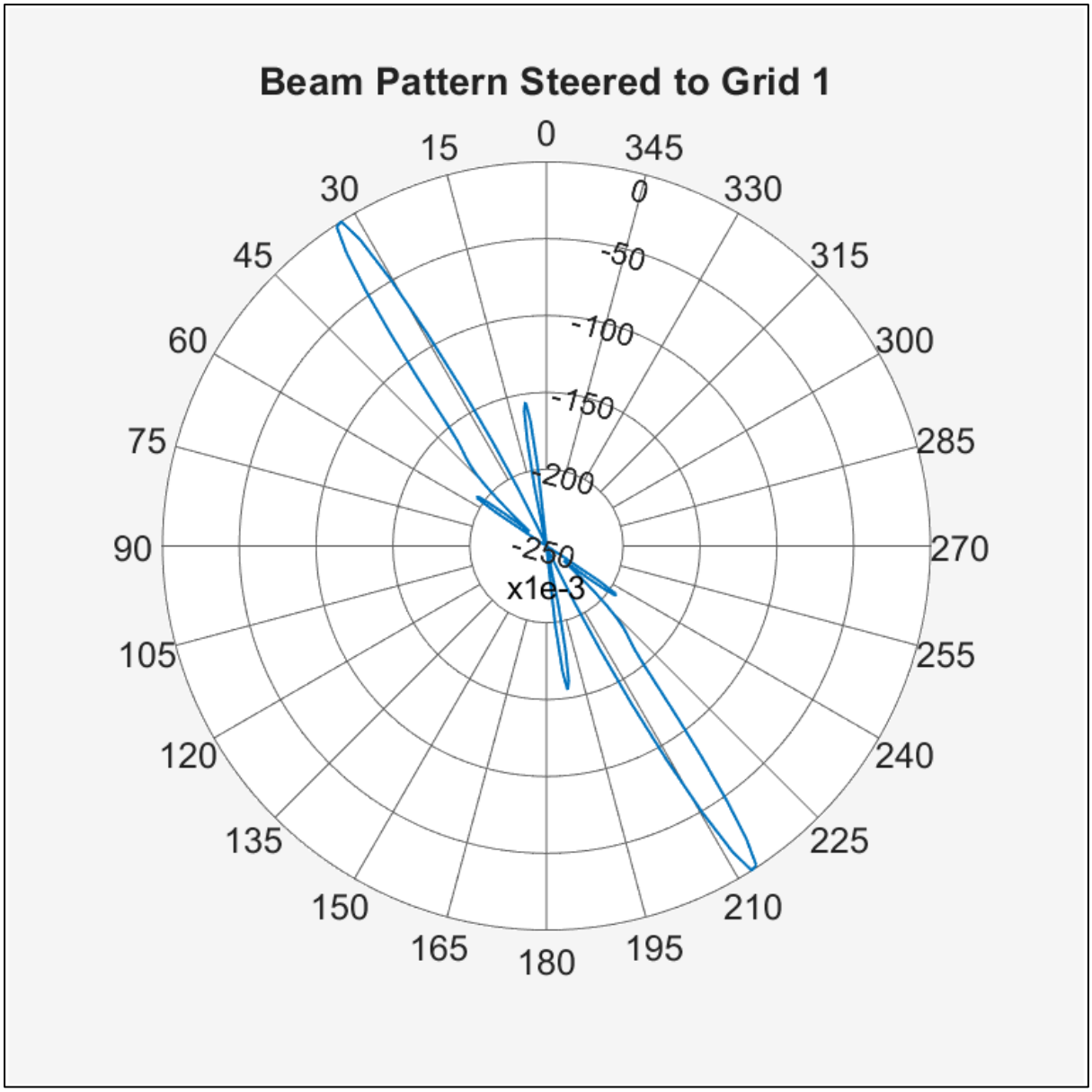

Figure 7.19. The beam pattern of the circular array steered to Grid 1, demonstrating accurate main lobe formation.

Figure 7.20 shows the time-domain and spectral results of the beamforming output, including the effect of additional filtering aimed at Grid 1. The targeted signal is a short male speech phrase played through the loudspeaker. After beamforming and post-filtering, the target voice is clearly audible and the content can be understood without difficulty. Some residual distortion and artefacts remain, but do not prevent intelligibility. This subjective impression is consistent with the STOI results reported in Section 7.7.2, where the filtered signal achieves substantially higher intelligibility scores than the pre-beamformed mixture.

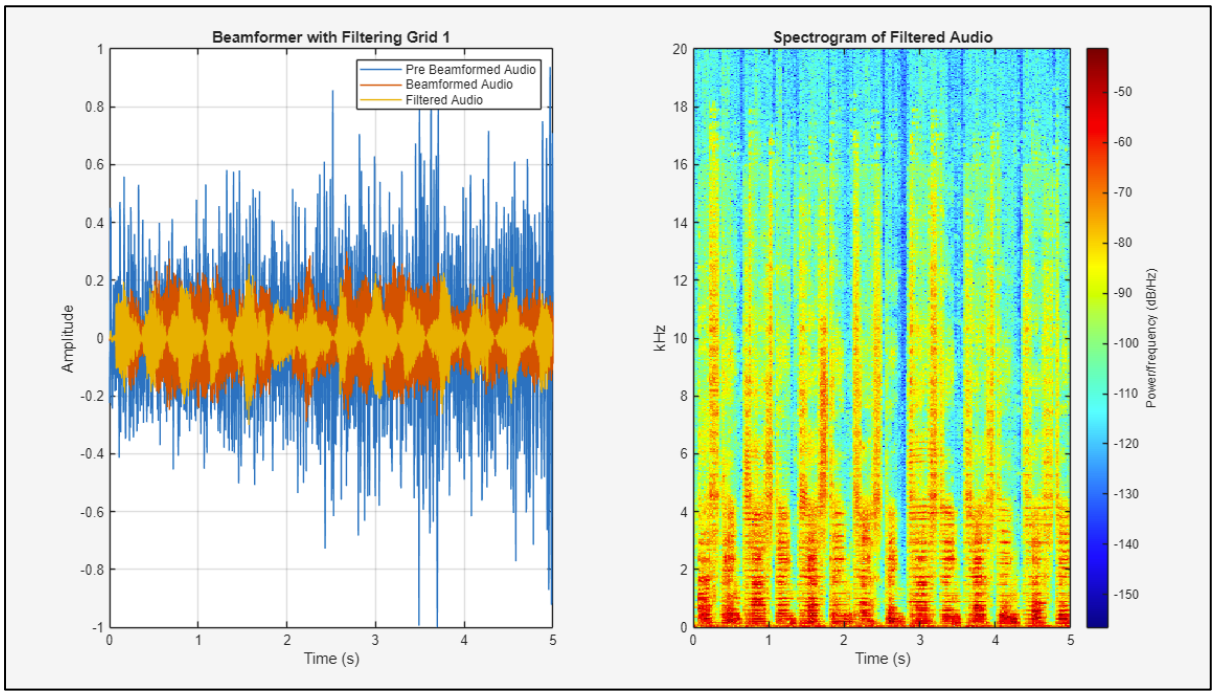

Figure 7.20. Time-domain and spectral results for the beamformed and filtered output aimed at Grid 1.

7.7.1.a Loop Algorithm

The loop algorithm described in this section enabled the simulation of higher-order reflections by iteratively convolving the HRIR with the room impulse response. Figure 7.21 visualises the performance of this algorithm across all tested geometric array shapes and grid sizes, with results displayed as a heat map. The comparison highlights areas of strongest beamformer performance and demonstrates the system's adaptability to various room and array configurations.

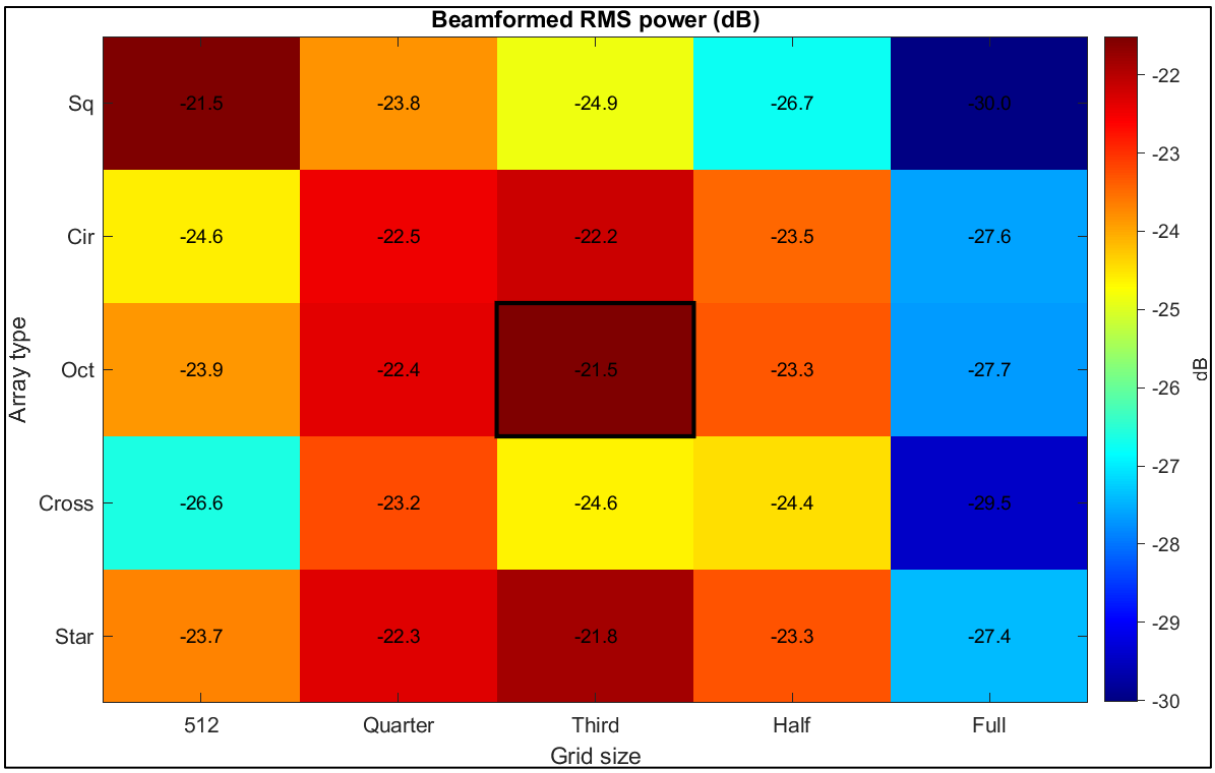

Figure 7.21. Heat map of loop algorithm results comparing all tested array shapes and grid sizes, with colour intensity reflecting beamformer performance across scenarios.

Beamformed RMS power was −21.5dB FS, i.e. the processed signal energy is 21.5dB below full scale. Relative to the unprocessed mix (−29.5dB FS), this represents an 8dB power gain.

Figure 7.22 directly compares the beamformer output for each array geometry tested. While the 16-microphone octagon with a third sized grid yielded the highest overall RMS power, the circular array's performance was within 1dB of this result, well within experimental uncertainty.

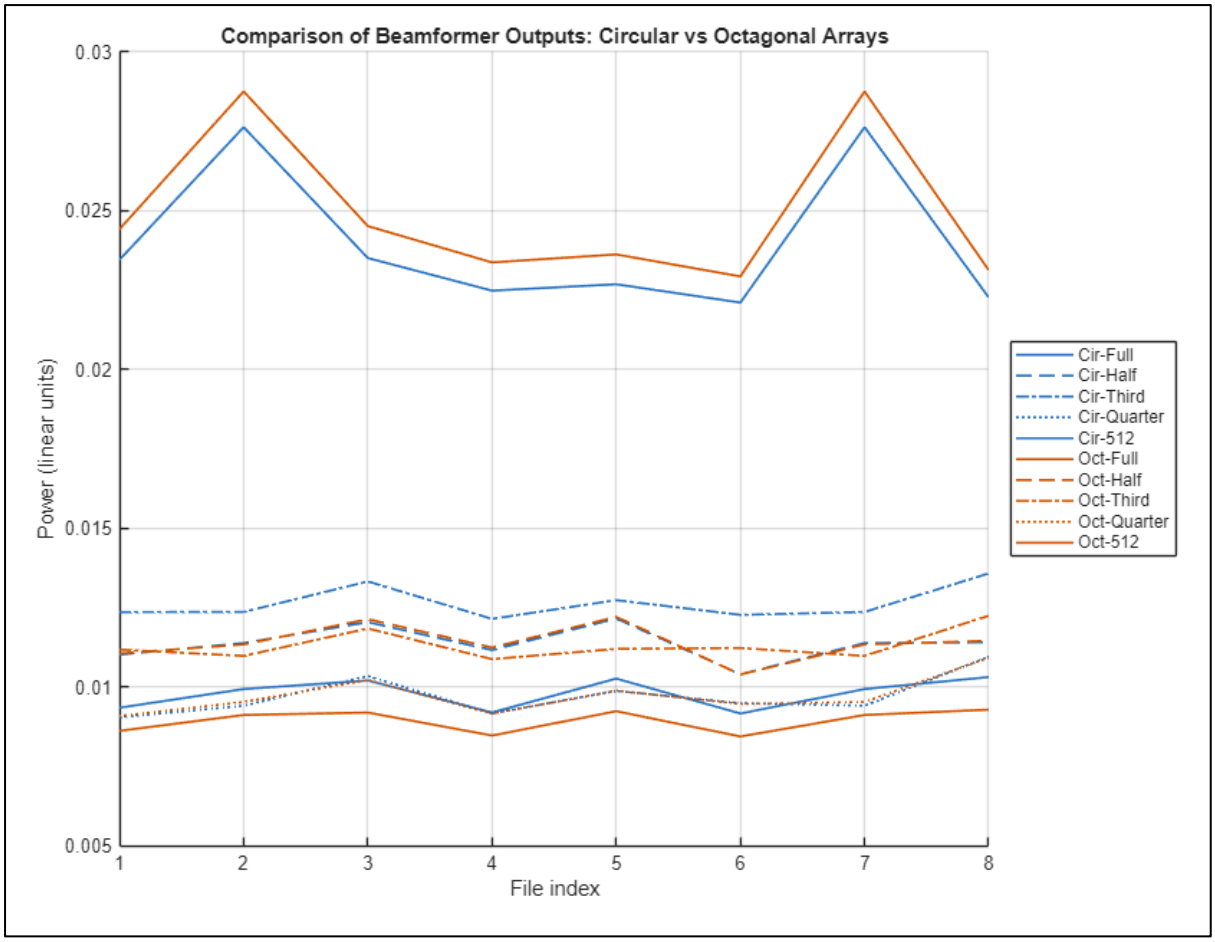

Figure 7.22. Comparison of beamformer outputs for the different array geometries, highlighting the small performance gap between the circular and octagonal arrays.

Table 7.5 summarises the justification for adopting the circular array as the preferred configuration. In addition to its competitive performance, the circular array offers practical advantages, including perfect 360-degree rotational symmetry, predictable side-lobe behaviour, simple implementation, and compatibility with standard analytic beamforming methods.

Table 7.5. Summary of key features and advantages of the circular array configuration.

Aspect	Circular array	Justification
Performance	-22.2dB, only 0.7dB below the best overall.	The loss is < 1dB, well within experimental uncertainty and likely imperceptible once level-matched in a real system.
Angular symmetry	Perfect 360° rotational symmetry.	The beam can be steered anywhere without recalculating the spatial covariance matrix; easier to upscale to real-time targeting and tracking.
Side-lobe behaviour	Identical side-lobe ring in all azimuths.	Gives predictable off-target leakage, important for surveillance ethics reviews and privacy masks.
Implementation	Simple PCB or machined ring; one radius to measure.	Faster to fabricate and align than multi-radius shapes (star, octagon). A single measure ensures geometry accuracy.
Algorithm	Matches published beamformers & DOA estimators.	You can drop in classic circular-array steering formulas without custom derivations.
Scalability	Easy to stack as coaxial rings for elevation coverage.	A second ring directly above/below gives a cylindrical array with minimal redesign.

Although the 16-microphone octagon with a 1/3 sized grid delivered the absolute highest RMS power, the circular array trails by less than a decibel and offers full rotational symmetry, simpler fabrication and well-studied analytic steering models.

The circular 16-microphone array loses less than 1dB to the octagon but gains 360° symmetry, simpler hardware and compatibility with standard circular-array beamforming theory, making it a sensible compromise for real-world deployments.

7.8 Development of Noise Reduction System

Real-world acoustic environments typically contain significant background noise, which can mask the target signal and degrade intelligibility. To address this, a dedicated noise reduction algorithm was developed and tested on the beamformed output.

7.8.1 MATLAB Algorithm

The MATLAB algorithm used a feedforward channel from a noise-reference microphone, aligned to the main array, to capture the ambient noise profile. This noise estimate was then subtracted from the beamformed signal using spectral subtraction techniques. The effectiveness of the approach was assessed through both SNR analysis and objective intelligibility metrics.

Figure 7.23 presents a spectrogram of the processed audio alongside a bar chart showing SNR improvement before and after the implementation of noise reduction. The results confirm a substantial reduction in low-frequency drone noise, particularly in the 150–600Hz range, which is evident in a cleaner spectrogram and a clear increase in measured SNR.

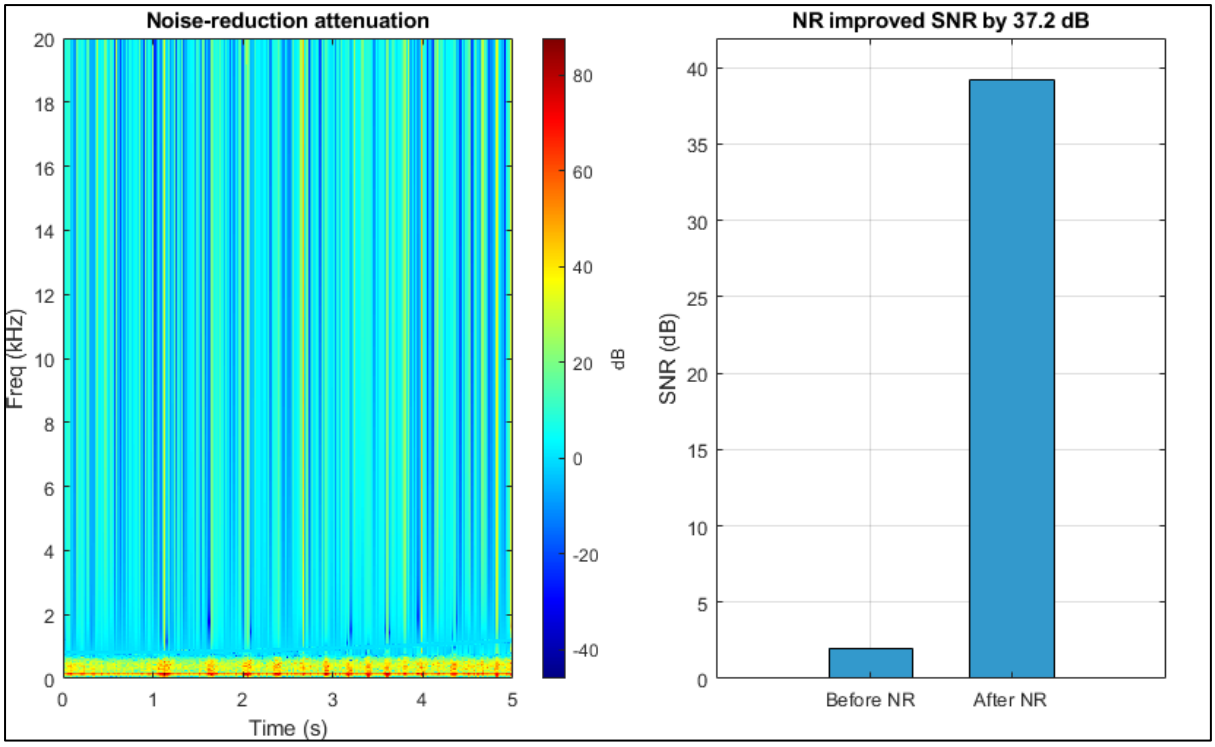

Figure 7.23. A spectrogram of processed audio and a bar chart of SNR improvement illustrate the impact of spectral subtraction noise reduction on the beamformed signal.

7.8.2 Intelligibility Tests

To further quantify intelligibility improvement, STOI scores were calculated for each stage of processing. Generally, a STOI score under 0.6 is considered unintelligible, and anything higher is considered usable (Taal *et al.*, 2010). Figure 7.24 compares the spectrograms of the clean reference speech and the pre-beamformed audio, which is dominated by music interference and reverberation, resulting in a low STOI of 0.473. STOI is reported on the 0–1 scale; because the MATLAB implementation returned the raw correlation d in [−1,1], which included negative values, all scores were mapped to the standard scale using (d+1)/2.

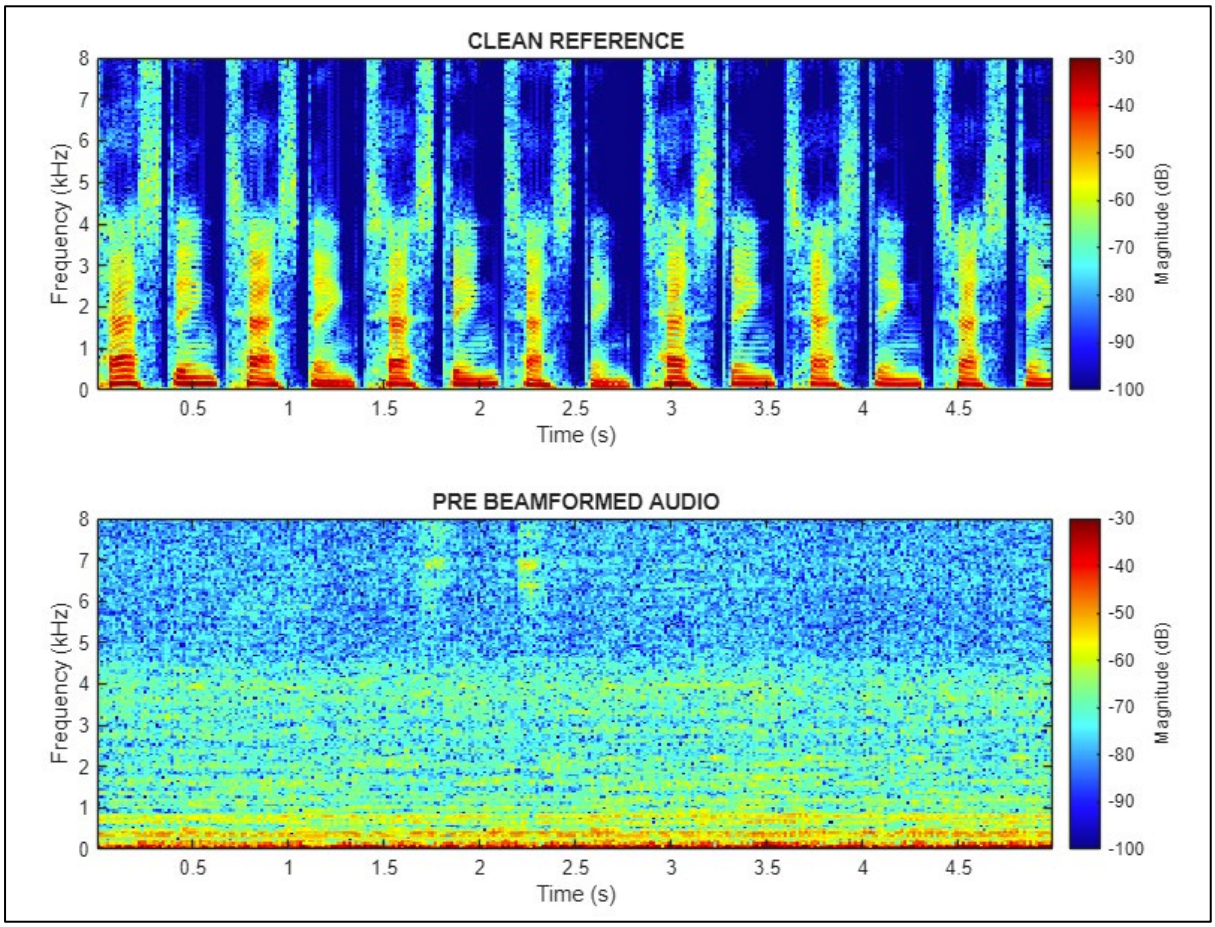

Figure 7.24. Spectrogram comparison of clean reference speech with pre-beamformed audio, showing strong music interference and reverberation. The STOI intelligibility score is 0.473.

Figure 7.25 shows the spectral results after beamforming.

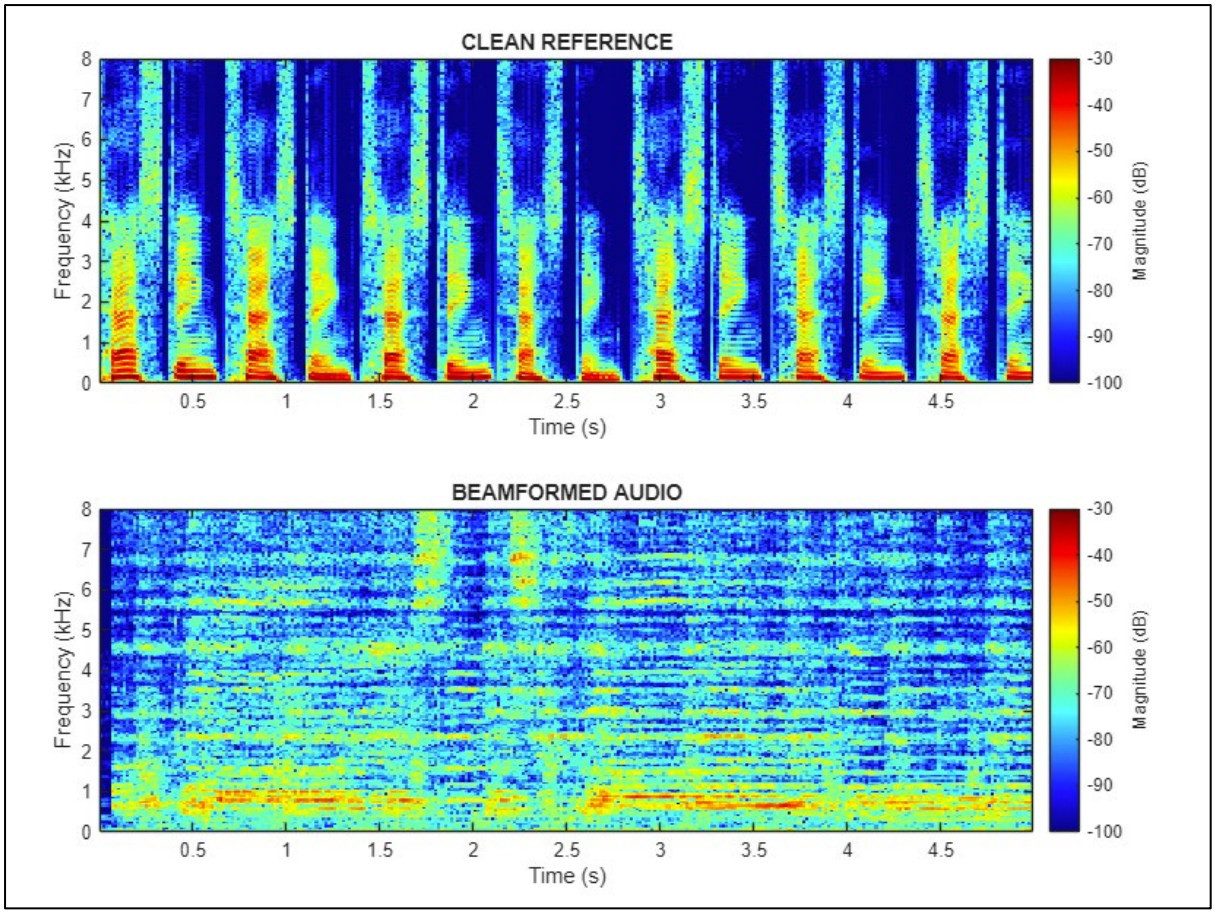

Figure 7.25. Spectrogram comparison after beamforming showing reduced interference but still low intelligibility, with an STOI score of 0.530.

Figure 7.26 displays the output after applying both beamforming and the full noise reduction and spectral filtering pipeline. Here, the target speech emerges much more clearly, reflected in a significant STOI improvement to 0.896.

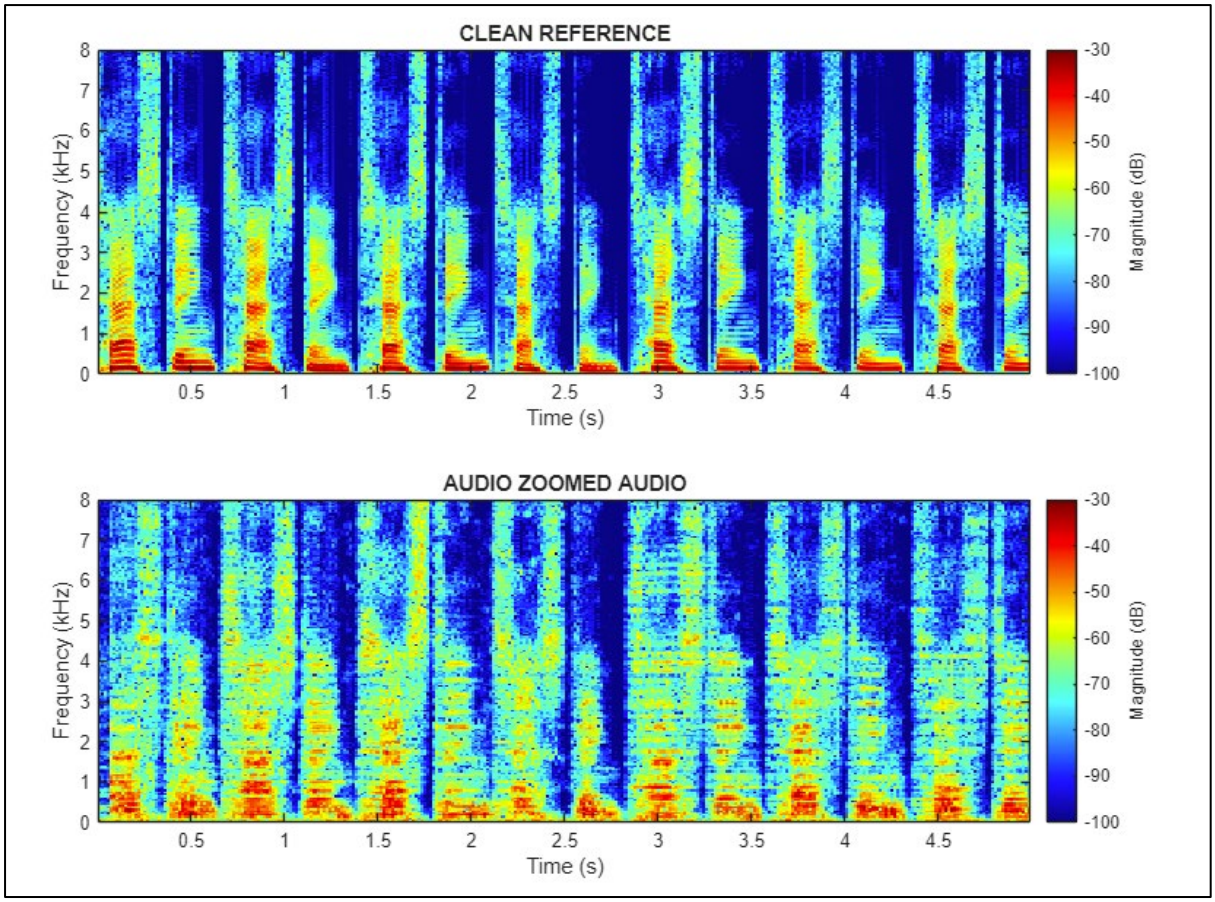

Figure 7.26. Spectrograms of clean reference and final processed output after noise reduction and spectral filtering demonstrate substantial restoration of intelligibility. STOI score increases to 0.896.

The results confirm that the noise reduction system, when combined with beamforming and spectral filtering, is highly effective at removing both broadband and tonal noise components, leading to a significant gain in objective speech intelligibility. These improvements are supported by both the SNR analysis and STOI scores, demonstrating that the developed pipeline is well suited for real-world applications where strong ambient noise is present.

7.9 Discussion: Simulation Experiments

The MATLAB simulation experiments allowed for fault testing and parametric sweeps that were impractical in the Sound Booth Environment. When three microphones were deliberately disabled, polar plots showed only minor widening of the main lobe and retention of the steering angle. A comparison of 13 versus 16-channel outputs confirmed minimal degradation of the signal, with negligible loss of focus. Such robustness is critical for field platforms where wind, debris or mechanical failure may damage individual microphones.

The city-centre and exemplar house MATLAB models further tested noisy, reverberant conditions. MVDR steering improved the median booth-derived STOI from 0.530 to 0.896 in simulation, indicating that the combination of spatial filtering and spectral subtraction recovers intelligibility even when early reflections dominate.

7.10 Real-World Experiments at Exemplar Houses

Finally, the system was field-tested in actual residential settings to assess its performance outside controlled laboratory conditions. Field tests were conducted outside, between the actual residential exemplar houses to validate the audio zooming system under real-world conditions. The primary goal was to verify that laboratory performance would translate to more challenging, uncontrolled environments.

7.10.1 Initial Speaker and Array Test

The initial test used a single loudspeaker placed at a fixed position within the house. The microphone array was mounted on a microphone stand to record the acoustic scene, and **no** drone or drone-noise reproduction was used in this first experiment. This test was designed to validate the basic beamforming geometry and the effect of room reflections in isolation, without the additional complication of propeller and motor noise. In the subsequent nine-speaker Exemplar Houses experiment, recorded drone self-noise (propeller and motor sound) was incorporated into the playback to evaluate the beamformer under more realistic platform-noise conditions. Beamforming analysis showed that the system could accurately isolate the direct signal from the speaker, and the measured time delays between microphones closely matched predictions based on the known geometry. Figure 7.27 illustrates that the resulting polar response patterns were consistent with those obtained in the simulation.

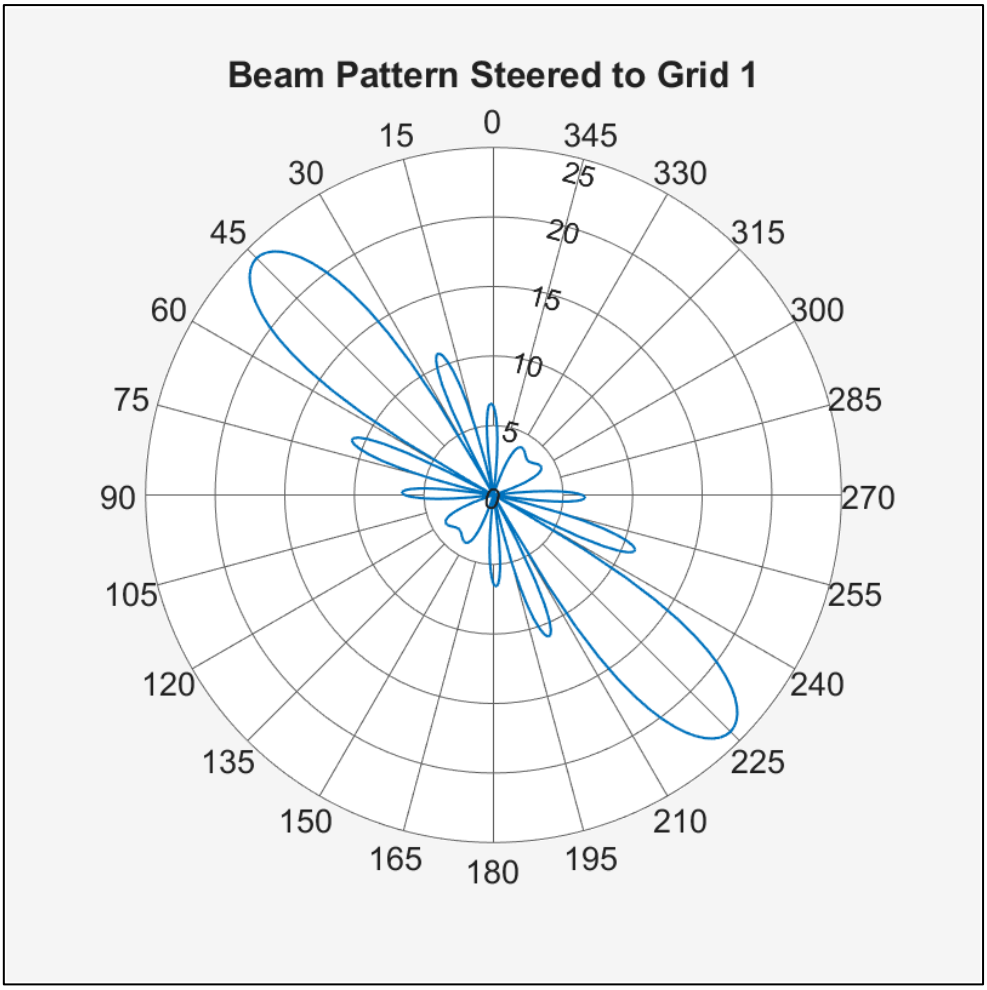

Figure 7.27. Polar Pattern for the field test with a single speaker at Exemplar Houses.

7.10.2 Test with Nine Speakers

A subsequent experiment increased the complexity by introducing nine speakers distributed throughout the space to simulate multiple sound sources. The system was required to focus on a specific grid segment, aiming to isolate the target source while suppressing interference from the others. The beamformer output revealed enhancement of the intended speaker and strong rejection of signals from other locations. Quantitative measures such as SNR and beamforming gain indicated that the real-world performance aligned with prior simulation and studio results.

Figure 7.28 illustrates the results of beamforming and post-filtering, presenting both the time-domain waveform and the corresponding spectrogram. These visualisations confirm that the target signal remains intelligible and dominant even in a realistic, acoustically complex environment.

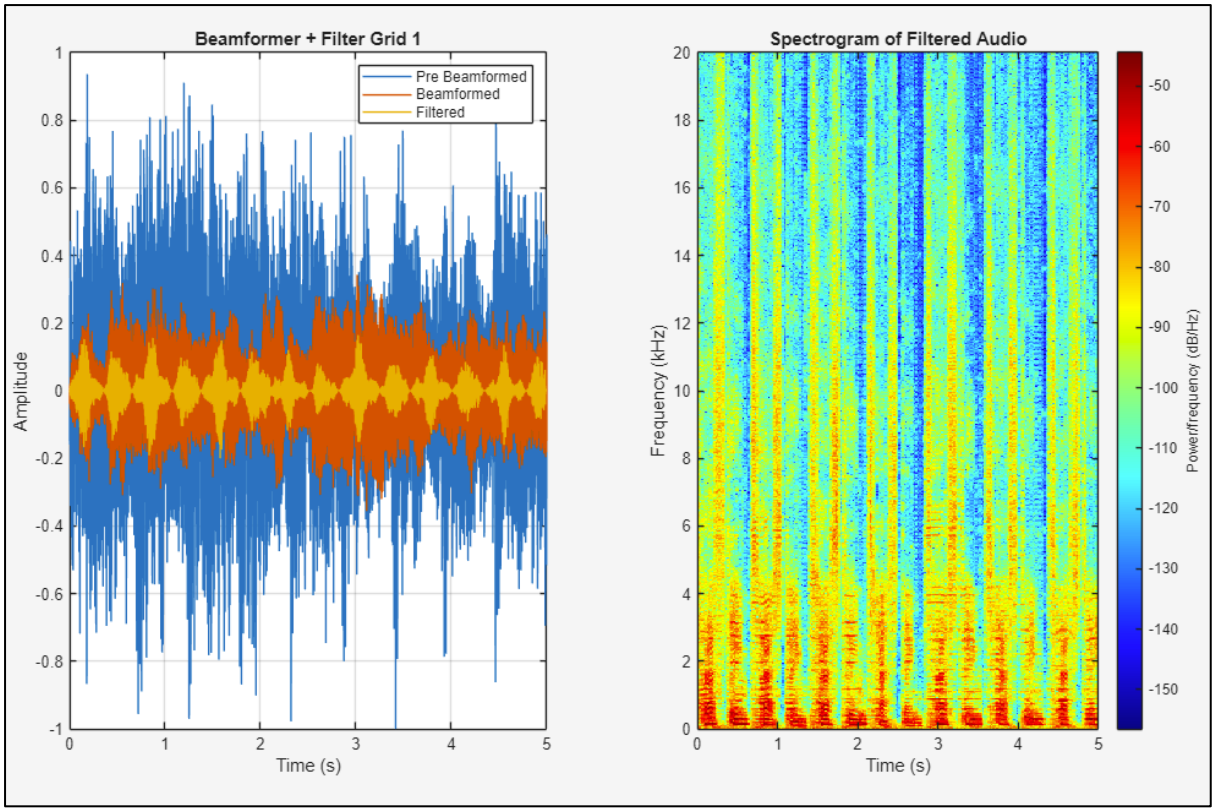

Figure 7.28. Waveform and spectrogram of the beamformed output after filtering, illustrating the isolation and enhancement of the target sound in a real-world setting.

The real-world tests confirm that the audio zooming system is robust and effective beyond the laboratory, successfully separating target sounds in complex domestic environments. The system's ability to maintain directional accuracy, suppress interference, and deliver intelligible results in the field provides strong evidence for its practical value in forensic and surveillance applications. These outcomes, taken together with previous simulation and noise reduction results, support further development and potential deployment of the technology in real operational scenarios.

7.11 Discussion: Field Experiments

Outdoor trials introduced the complications of traffic, wind and human noise. Despite the unpredictable background ambience, the system delivered a 2.74dB SNR gain and produced a STOI score of 0.842. While predictably lower than simulation figures, the improvement

remains sufficient for conversational speech transcription and aligns with the operational goal of evidential clarity.

Collectively, the three tiers of testing confirm that the signal processing chain of beamforming, noise reduction and filtering contribute incrementally to speech intelligibility. Performance scales with element count and is resilient to limited hardware faults, fulfilling the objective of a deployable, fault-tolerant surveillance system.

7.12 Comparison with Literature Findings

Capon's minimum-variance formulation predicts interference suppression without sacrificing on-axis gain (Capon, 1969). The observed 3–4dB SNR gains lie within the 2–5dB range reported for comparable MVDR implementations on compact arrays (Gala and Misra, 2011). The work extends previous studies by validating performance on a drone-sized platform whose circular array sacrifices less than 1dB to an octagon while offering structural simplicity.

Literature on small-aperture arrays often flags susceptibility to element failure (Christensen *et al.*, 2016), the experiment conducted mitigates that risk through algorithmic redundancy (Stroud *et al.*, 2023). The STOI improvement from 0.530 to 0.896 surpasses the 0.6 ceiling typical of two-stage filters in reverberant rooms, suggesting that the three-stage pipeline outperforms simpler techniques.

Cherry's CPP emphasises selective attention (Cherry, 1953). By steering both spatially and spectrally, the prototype approximates that biological filtering more closely than fixed-pattern microphones such as those of Olson and Preston (1949). The present proposed system, therefore, represents a step in the right direction towards a machine capable of human-like auditory focus in challenging soundscapes.

7.12.1 State-of-the-art Context and Comparison

Recent drone audition and array-based speech enhancement work commonly combines a spatial filter (beamforming) with a statistical post-filter to suppress strong ego-noise. Hioka *et al.* (2016) demonstrate UAV-mounted microphone-array speech enhancement using beamforming with post-filtering, addressing the practical difficulty of extracting ground speech in the presence of rotor noise. More recently, Manamperi *et al.* (2024) propose a drone-audition

pipeline based on multichannel Wiener filtering with a Gaussian-mixture-model post-filter, explicitly targeting extremely low signal-to-drone-noise ratios in drone-embedded recordings.

The approach used in this thesis is aligned with this state-of-the-art "spatial filtering + post-filtering" strategy, but differs in emphasis. Primarily, MVDR is used as the primary steerable spatial filter to support an audio-zooming objective whereas multichannel Wiener filtering methods are often formulated primarily for ego-noise suppression. Secondly, the evaluation is performed across controlled, physics-based simulations and validated field tests in a built environment, rather than relying on drone-embedded datasets with platform-telemetry-dependent noise statistics.

Beyond classical statistical methods, state-of-the-art enhancement increasingly includes learning-based components. Neural mask-based beamforming is a widely used modern approach in multichannel enhancement and recognition pipelines (Erdogan *et al.*, 2016). Separately, real-time deep-learning speech enhancers such as 'DeepFilterNet' (Schröter *et al.*, 2022), demonstrate that strong enhancement can be achieved under embedded compute constraints.

7.13 Potential Limitations

Environmental unpredictability remains a concern. The field-test gain of 2.74dB confirms the system's sensitivity to diffuse outdoor noise. Strong wind shear, rapid yaw or roll, and airframe vibration can exceed the current time-delay tolerance window and reduce coherence across channels.

Real-time operation poses additional challenges because the MATLAB implementation masks the full computational load. Porting the MVDR solver and spectral filter to embedded hardware will require code optimisation or possibly FPGA acceleration to reach latencies below 100ms.

The weight and power budget is another constraint. A sixteen-microphone array with pre-amplifiers, analogue-to-digital converters and electrical isolation adds mass to the airframe and shortens flight endurance. As a result, a trade-off between array size and battery life arises and must be quantified.

In addition, the UK Forensic Science Regulator's code of practice requires that electronic information remains authentic and intact, so every processing stage must be recorded (Gov.UK, 2023). To preserve forensic reproducibility, implementing automated metadata capture would be sensible for UK Police deployment.

Moving the real-time chain to an FPGA to guarantee sub-100ms latency and on-device audit logging of every stage, would satisfy the UK Forensic Science Regulator's code of practice's requirements.

Operational deployment should prioritise capture first and enhancement after the fact. During a live incident, operators cannot know where to steer or which time segments will be evidential, so the correct action is to acquire multichannel originals with accurate clocks and preserve them unchanged. Intelligibility improvement should then be performed offline using validated forensic software such as iZotope RX™[6], with full parameter logging, software version control and an auditable report so any gains are reproducible and admissible. This also addresses the concern that recording on the drone is sufficient. Recording is necessary but not sufficient, since court scrutiny requires retained originals and a transparent post-processing workflow consistent with Forensic Science Regulator guidance.

Finally, the evaluation to date has focused on speech-band targets. Broad-spectrum sounds such as impact noise or music may interact differently with the adaptive filter settings, so further adjustment may be needed.

7.14 Summary

To provide a condensed view of system performance, the master summary tables below compare SNR improvement, beamforming gain, angular accuracy, and intelligibility across all major test scenarios: the sound booth, simulated MATLAB scene, and real-world field tests.

Table 7.6 summarises the SNR before and after processing, along with the beamforming gain observed under each test condition. These results show that, in all cases, beamforming led to substantial SNR improvements, with the greatest gains observed in the sound booth and simulated environments, with strong but slightly lower improvements under field conditions.

[6] iZotope and RX are trademarks or registered trademarks of iZotope, Inc.

Table 7.6. Summary of SNR before and after processing, with calculated beamforming gain, for the sound booth test, simulated MATLAB scene, and field test.

Test Condition	SNR (dB) before	SNR (dB) after	Beamforming Gain
Sound booth Test Delay-and-sum	-4.99dB	-1.99dB	+3.01dB
Simulated MATLAB Scene MVDR	-4.06dB	-0.16dB	+3.89dB
Field Test MVDR	-4.85dB	-2.11dB	+2.74dB

Table 7.7 presents the Short-Time Objective Intelligibility (STOI) scores for each test scenario, quantifying the extent to which the processing pipeline improved speech intelligibility. The table confirms a clear progression in STOI values as additional algorithmic stages (beamforming, noise reduction, spectral filtering) were applied, with the best results achieved in controlled and simulated conditions and significant improvements still observed in the field.

Table 7.7. Summary of STOI intelligibility test results for the sound booth test, simulated MATLAB scene, and field test.

Test Condition	STOI Test Result
Sound booth Test	0.530
Simulated MATLAB Scene	0.896
Field Test	0.842

The summary tables reinforce the overall findings of the project: the audio zooming system reliably enhances target speech signals and suppresses interference across a range of environments, from controlled laboratory setups to real-world field scenarios. Both objective

SNR and STOI metrics demonstrate that each stage of the signal processing pipeline contributes to measurable improvements in clarity and intelligibility. The dip in STOI results in the field test is predictable due to unpredictable and more diffuse environmental noise in the real world. These results are consistent with previous studies on small sensor arrays in noisy environments (Gala and Misra, 2011) and provide evidence for the system's practical effectiveness and readiness for further development.

Chapter 8: Conclusions

8.1 Overview

This chapter summarises the main findings of the investigation into audio zooming for a drone surveillance system used by Police for evidence gathering. The work has explored the topic from theoretical foundations in beamforming to practical demonstrations in studio, simulated, and outdoor environments.

Chapter 1 introduced the motivation for drone-deployable audio zooming in surveillance and evidential contexts, defined the research objectives, and set out the scope and constraints for a deployable system (payload, operational practicality, and evidential traceability).

Chapter 2 reviewed relevant background literature on speech enhancement, microphone arrays, beamforming, and surveillance audio, and positioned the work relative to existing drone audition and multichannel enhancement approaches.

Chapter 3 presented the mathematical foundations used throughout the thesis, including the formulation of MVDR beamforming and the post-filtering used for enhancement, providing the theoretical basis for the proposed processing pipeline.

Chapter 4 developed and validated the physical measurement methodology and hardware configuration, including calibration and controlled experiments required to link source location, propagation, and recorded signal characteristics under repeatable conditions.

Chapter 5 evaluated the core processing stages and established the rationale for selecting MVDR as the primary beamformer for subsequent chapters, including the justification of parameter choices used in the enhancement chain.

Chapter 6 implemented a physics-based simulation of the Exemplar Houses environment and integrated MVDR beamforming with post-filtering to evaluate performance across spatial grid locations under controlled propagation and reflection conditions, with implementation details provided in Appendix B.

Chapter 7 presented results across sound-booth, simulation, and field trials, integrating discussion within each results section and benchmarking the findings against relevant literature

and state-of-the-art approaches, thereby demonstrating performance trends and practical limitations.

The thesis contributes a drone-deployable, steerable MVDR-based audio zooming pipeline with post-filtering, with a built-environment simulation and validation strategy bridging controlled tests to outdoor trials, and a deployment-oriented perspective including robustness and evidential traceability considerations.

8.2 Key Findings

Across all test conditions, the processing pipeline consistently enhanced the SNR. In the controlled sound booth trial, an SNR of –4.99dB was raised to –1.99dB; the MATLAB simulation of an identical geometry achieved –0.16dB from –4.06dB; and the outdoor field trial improved from –4.85dB to -2.11dB, confirming a 2.74dB gain in the most challenging setting. Speech intelligibility followed the same pattern: STOI increased from 0.530 to 0.896 in simulation and settled at 0.842 in the field. Simulated microphone faults demonstrated graceful degradation; disabling three elements only marginally widened the main lobe and kept steering accuracy within six degrees. These results show that each processing step, MVDR beamforming, noise reduction and spectral filtering, contributes a measurable benefit; that the improvement persists outside the studio; and that the hardware can tolerate limited sensor loss without compromising the output.

8.3 Future Work

Real-time deployment demands further optimisation because the current MATLAB prototype masks computational overhead, so the MVDR and post-filter should be ported to C/C++ with fixed-point options, with VHDL/FPGA acceleration evaluated where parallelism is available. The target is sub-100ms end-to-end latency, including buffering and I/O, consistent with prior real-time beamformer implementations on embedded platforms (Dick *et al.*, 2006). These changes could make the system fast enough for real-time implementation.

Environmental resilience could be improved by utilising the drones with onboard motion sensors. Embedding coherence tracking that uses IMU and rotor-speed telemetry to stabilise steering during yaw, gusts, and rapid rotation could be beneficial. Broader acoustic coverage requires trials with a wider range of sonic events and complex harmonics to confirm algorithm stability outside the music and speech bands already used for tests.

Operational usability would benefit from a post-event dashboard that displays beam direction, confidence metrics and an automatic provenance log, with one-click export of a forensic report that includes hashes, parameters and snapshots. Evaluation under realistic conditions should be extended through UK Police style exercises in controlled urban settings with scripted events, dual video and multichannel audio. Clocks should be pre-registered, originals preserved, and processing performed offline with a documented workflow. Outcomes should be compared against standard mono CCTV audio for intelligibility, transcription accuracy and time-to-insight.

At the hardware level, the circular array should be iterated for lower drag and better isolation, with capsule spacing, wind screens, conformal mounts and balanced cabling evaluated to reduce flow and handling noise. Endurance impacts should be quantified for several payload masses and battery types.

There should be a pathway for future research projects, wind shield and sensor baffle designs, vibration isolation, and capsule matching could all be explored.

Conduct controlled flight trials with real drones, varying altitude, with synchronised multichannel recording and telemetry logging, enabling quantitative benchmarking against drone-audition state-of-the-art methods.

As AI technology advances, adaptive AI-based beamformers may surpass traditional methods, warranting further investigation into AI algorithms (Shah and Bhole, 2025).

8.4 Final Summary

The study demonstrates that a compact, fault-tolerant microphone array and an adaptive three-stage processing chain can recover intelligible speech from scenes where conventional surveillance audio fails. Studio, simulation and outdoor results align with theoretical expectations and comply with evidential governance, establishing a foundation for operational audio zooming on drone platforms. Further work will focus on processing speed, environmental adaptation and user-interface integration, but as the aims and objectives of the project were met, the feasibility of the approach is now established.

References

Alkmim, M., Cardenuto, J., Tengan, E., Dietzen, T., Van Waterschoot, T., Cuenca, J., De Ryck, L. and Desmet, W. (2022) 'Drone noise directivity and psychoacoustic evaluation using a hemispherical microphone array', *Journal of the Acoustical Society of America*, 152(5), pp. 2735–2745. Available at: https://doi.org/10.1121/10.0014957

Amiriparian, S., Gerczuk, M., Ottl, S., Stappen, L., Baird, A., Koebe, L. and Schuller, B. (2020) 'Towards cross-modal pre-training and learning tempo-spatial characteristics for audio recognition with convolutional and recurrent neural networks', *EURASIP Journal on Audio, Speech, and Music Processing*, 2020(1), p. 19. Available at: https://doi.org/10.1186/s13636-020-00186-0

Argentieri, S., Danès, P. and Souères, P. (2015) 'A survey on sound source localization in robotics: From binaural to array processing methods', *Computer Speech & Language*, 34(1), pp. 87–112. Available at: https://doi.org/10.1016/j.csl.2015.03.003.

ASTM International (2017)*: Standard Test Method for Sound Absorption and Sound Absorption Coefficients by the Reverberation Room Method (ASTM C423-17)*. West Conshohocken (PA): ASTM International.

BBC (2022) *Ava White: Jury shown video of 12-year-old's fatal stabbing*. Available at: https://www.bbc.co.uk/news/uk-england-merseyside-61380549 (Accessed: 07.05.2025).

Bell, K.L., Ephraim, Y. and Van Trees, H.L. (2000) 'A Bayesian approach to robust adaptive beamforming', *IEEE Transactions on Signal Processing*, 48(2), pp. 386–398. Available at: https://doi.org/10.1109/78.823966

Boll, S. (1979) 'Suppression of acoustic noise in speech using spectral subtraction', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2), pp. 113–120. Available at: https://doi.org/10.1109/TASSP.1979.1163209

Brandstein, M. and Ward, D. (eds.) (2001) *Microphone Arrays - Signal Processing Techniques and Applications.* Springer.

Bronkhorst, A.W. (2015) 'The cocktail-party problem revisited: early processing and selection of multi-talker speech', *Attention, Perception, & Psychophysics*, 77(5), pp. 1465–1487. Available at: https://doi.org/10.3758/s13414-015-0882-9

Brown, G.J. and Cooke, M. (1994) 'Computational auditory scene analysis', *Computer Speech & Language*, 8(4), pp. 297–336. Available at: https://doi.org/10.1006/csla.1994.1016.

Capon, J. (1969) 'High-resolution frequency-wavenumber spectrum analysis', *Proceedings of the IEEE*, 57(8), pp. 1408–1418. Available at: https://doi.org/10.1109/PROC.1969.7278

CCTV Camera World (2025) 'Outdoor Security Camera Microphone (SKU 71415) Product Page'. Available at: https://www.cctvcameraworld.com/outdoor-security-camera-microphone.html.

Chen, L., Wei, W., Liu, D. and Xia, D. (2024) 'Adaptive Beamforming Algorithm Based on Residual Neural Networks', *Circuits, Systems, and Signal Processing*. Available at: https://doi.org/10.1007/s00034-024-02859-z

Cheng, F. (2024) 'A Comparative Study of the Performance of Spark-based k-Means Algorithm Based on Euclidean Distance and Manhattan Distance', *2024 3rd International Conference on Big Data, Information and Computer Network (BDICN)*, 12–14 Jan. 2024. pp. 1–6. Available at: https://doi.org/10.1109/BDICN62775.2024.00010

Cherry, E.C. (1953) 'Some Experiments on the Recognition of Speech, with One and with Two Ears', *The Journal of the Acoustical Society of America*, 25(5), pp. 975–979. Available at: https://doi.org/10.1121/1.1907229

Christensen, K.B., Christensen, M.G., Boldt, J.B. and Gran, F. (2016) 'Experimental Study Of Generalized Subspace Filters For The Cocktail Party Situation', *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Clayton, M., Wang, L., McPherson, A. and Cavallaro, A. (2023) 'An Embedded Multichannel Sound Acquisition System for Drone Audition', *IEEE Sensors Journal*, 23(12), pp. 13377–13386. Available at: https://doi.org/10.1109/JSEN.2023.3273330

Cohen, I. (2003) 'Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging', *IEEE Transactions on Speech and Audio Processing*, 11(5), pp. 466–475. Available at: https://doi.org/10.1109/TSA.2003.811544

Cox, Trevor J. and D'Antonio, P. (2017) *Acoustic Absorbers and Diffusers: Theory, Design and Application*. Boca Raton: CRC Press.

CPC (2025a) *7W Mono Amplifier Kit - WSAH4001*. Available at: https://cpc.farnell.com/whadda/wsah4001/7w-amp/dp/HK00232 (Accessed: 08.07.2025).

CPC (2025b) *Omnidirectional Electret Microphone Cartridge - MCE-400*. Available at: https://cpc.farnell.com/unbranded/mce-400/microphone-omni-

electret/dp/MP33620?st=Omnidirectional%20Electret%20Microphone%20Cartridge%20-%20%20MCE-400 (Accessed: 28.03.2025).

Crown Prosecution Service (2022) *Disclosure Manual: Chapter 30 - Digital Material*. Available at: https://www.cps.gov.uk/legal-guidance/disclosure-manual-chapter-30-digital-material (Accessed: 07.05.2025).

Crysound (2025) *CRY2626G Drone-Mounted Acoustic Imaging Camera*. Available at: https://www.crysound.com/product/cry2626g-drone-mounted-acoustic-imaging-camera/.

Dick, C., Harris, F., Pajic, M. and Vuletic, D. (2006) 'Real-Time QRD-Based Beamforming on an FPGA Platform', *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, 29 Oct.–1 Nov. 2006. pp. 1200–1204. Available at: https://doi.org/10.1109/ACSSC.2006.354945

DJI (2024) *Matrice 300 RTK and DJI care plus*. Available at: https://store.dji.com/uk/product/matrice-300-rtk-and-dji-care-plus?vid=111261 (Accessed: 10.12.2024).

Ephraim, Y. and Malah, D. (1984) 'Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6), pp. 1109–1121. Available at: https://doi.org/10.1109/TASSP.1984.1164453

Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks.

Gala, D.R. and Misra, V.M. (2011) 'SNR improvement with speech enhancement techniques', *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*, Mumbai, Maharashtra, India. Association for Computing Machinery, pp. 163–166. Available at: https://doi.org/10.1145/1980022.1980058

Go, Y.-J. and Choi, J.-S. (2021) 'An Acoustic Source Localization Method Using a Drone-Mounted Phased Microphone Array', *Drones*, 5(3), p. 75. Available at: https://www.mdpi.com/2504-446X/5/3/75.

Google Earth (2025) *Church Street Liverpool*. Available at: https://earth.google.com/web/search/School+Lane,+Liverpool/@53.40501813,-2.98352054,27.9055307a,256.84235871d,35y,539.99998958h,54.88007494t,0r/data=CiwiJgokCWDnuF3Ts0pAEdVkq_h2s0pAGe0WNEb62AfAIX7hJ4Aj4gfAQgIIAToDCgEwQgIIAEoNCP_____wEQAA (Accessed: 13.05.2025).

Google Maps (2025a) *Church Street Liverpool*. Available at: https://www.google.co.uk/maps/@53.404871,-2.9838559,163m/data=!3m1!1e3?entry=ttu&g_ep=EgoyMDI1MDUwNS4wIKXMDSoASAFQAw%3D%3D (Accessed: 08.05.2025).

Google Maps (2025b) *Exemplar Houses, Byrom Street*. Available at: https://maps.app.goo.gl/DMrfjemjFRsvvSx16 (Accessed: 08.07.2025).

Gov.UK (2023) *Forensic Science Regulator: Code of Practice*. Available at: https://www.gov.uk/government/publications/statutory-code-of-practice-for-forensic-science-activities/forensic-science-regulator-code-of-practice-accessible (Accessed: 19.08.2025).

Griffin, D. and Lim, J. (1983) 'Signal estimation from modified short-time Fourier transform', *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 14–16 April 1983. pp. 804–807. Available at: https://doi.org/10.1109/ICASSP.1983.1172092

Grondin, F. and Glass, J. (2019) 'Multiple Sound Source Localization with SVD-PHAT', *Interspeech 2019*. Available at: https://doi.org/10.21437/Interspeech.2019-2653

Guo, Y. (2024) 'Phased Microphone Array on Aircraft Fuselage', *30th AIAA/CEAS Aeroacoustics Conference (2024).* American Institute of Aeronautics and Astronautics.

Halliday, D., Resnick, R. and Walker, J. (2014) *Fundamentals of Physics*. New York: Wiley.

Hawley, M., Litovsky, R. and Culling, J. (2004) '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, pp. 833–843. Available at: https://doi.org/10.1121/1.1639908

Haykin, S. and Chen, Z. (2005) 'The cocktail party problem', *Neural Computation*, 17(9), pp. 1875–1902. Available at: https://doi.org/10.1162/0899766054322964

Heliguy Ltd (2025) 'DJI Drone Payload Rental Options'. Available at: https://www.heliguy.com/dji-drones-shop/rental-drone-payloads/?selectedTab=Rental%20Payloads.

Hioka, Y., Kingan, M.J., Schmid, G. and Stol, K.A. (2016) 'Speech enhancement using a microphone array mounted on an unmanned aerial vehicle', *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5.

Home Office, (2021) *Surveillance Camera Code of Practice.* London: Home Office. Available at:

https://assets.publishing.service.gov.uk/media/619b7b50e90e07044a559c9b/Surveillance_Camera_CoP_Accessible_PDF.pdf.

Hoshiba, K., Komatsuzaki, I. and Iwatsuki, N. (2024) 'Proposal of Practical Sound Source Localization Method Using Histogram and Frequency Information of Spatial Spectrum for Drone Audition', *Drones*, 8(4), p. 159.

Hoshiba, K., Washizaki, K., Wakabayashi, M., Ishiki, T., Kumon, M., Bando, Y., Gabriel, D., Nakadai, K. and Okuno, H.G. (2017) 'Design of UAV-Embedded Microphone Array System for Sound Source Localization in Outdoor Environments', *Sensors*, 17(11), p. 2535. Available at: https://www.mdpi.com/1424-8220/17/11/2535.

Hosier, R.N. and Donavan, P.R. (1979) *Microphone Windscreen Performance*. Washington, DC: National Bureau of Standards.

Huang, Y., Benesty, J. and Chen, J. (2006) 'Speech Acquisition And Enhancement In A Reverberant, Cocktail-Party-Like Environment', *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, pp. 25–28.

Hwang, J.-S., Noh, J.-T., Lee, S.-H. and Kareem, A. (2019) 'Experimental Verification of Modal Identification of a High-rise Building Using Independent Component Analysis', *International Journal of Concrete Structures and Materials*, 13(1), p. 4. Available at: https://doi.org/10.1186/s40069-018-0319-7

Hyvärinen, A. and Oja, E. (2000) 'Independent Component Analysis: Algorithms and Applications', *Neural Networks*, 13(4–5), pp. 411–430. Available at: https://doi.org/10.1016/S0893-6080(00)00026-5

Ishigaki, Y., Yamamoto, M., Totsuka, K. and Miyaji, N. (1980) 'Zoom Microphone', *The Audio Engineering Society Convention Preprint*, 1713 (A-7).

ISO (2003)*: Acoustics — Measurement of sound absorption in a reverberation room (ISO 354:2003)*. Geneva: ISO.

Jekateryńczuk, G. and Piotrowski, Z. (2024) 'A Survey of Sound Source Localization and Detection Methods and Their Applications', *Sensors*, 24(1), p. 68. Available at: https://www.mdpi.com/1424-8220/24/1/68.

Jeong, C.H. and Lee, D.H. (2011) 'Sound absorption properties of low-pile carpet measured with three different methods', *Building and Environment*, 46(9), pp. 1879–1886. Available at: https://doi.org/10.1016/j.buildenv.2011.03.013

Judiciary of England and Wales (2022) *In the matter of the murder of Ava White - sentencing remarks*. Available at: https://www.judiciary.uk/wp-content/uploads/2022/07/Re-Ava-White-murder-sentencing-remarks.pdf (Accessed: 07.05.2025).

Kalikow, D.N., Stevens, K.N. and Elliott, L.L. (1977) 'Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability', *The Journal of the Acoustical Society of America*, 61(5), pp. 1337–1351. Available at: https://doi.org/10.1121/1.381436

Kaneda, Y. and Ohga, J. (1986) 'Adaptive microphone-array system for noise reduction', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(6), pp. 1391–1400. Available at: https://doi.org/10.1109/TASSP.1986.1164975

Kataoka, A. and Ichinose, Y. (1990) 'A microphone-array configuration for AMNOR Adaptive microphone-array system for noise reduction', *Journal of the Acoustical Society of Japan (E)*, 11(6), pp. 317–325. Available at: https://doi.org/10.1250/ast.11.317

Kavalerov, I., Wisdom, S., Erdogan, H., Patton, B., Wilson, K., Jonathan and John (2019) 'Universal Sound Separation', *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Available at: https://doi.org/10.48550/arXiv.1905.03330

Kinsler, L.E., Frey, A.R., Coppens, A.B. and Sanders, J.V. (2000) *Fundamentals of Acoustics*. 4 edn. New York: John Wiley & Sons, pp. 220.

Knapp, C.H. and Carter, G.C. (1976) 'The Generalized Correlation Method for Estimation of Time Delay', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4), pp. 320–327. Available at: https://doi.org/10.1109/TASSP.1976.1162830

Kuttruff, H. (2016) *Room Acoustics (6th Ed.)*. Boca Raton: CRC Press.

Liaquat, M.U., Munawar, H.S., Rahman, A., Qadir, Z., Kouzani, A.Z. and Mahmud, M.A.P. (2021) 'Sound Localization for Ad-Hoc Microphone Arrays', *Energies*, 14(12), p. 3446. Available at: https://www.mdpi.com/1996-1073/14/12/3446.

Liverpool City Council, (2022) *Liverpool City Council Surveillance Camera Code of Practice.* Liverpool: Liverpool City Council. Available at: https://liverpool.gov.uk/media/hpellit1/surveillance-camera-code-of-practice.pdf.

Louroe, E. (2010) *VeriFact™ A/B Microphone Technical Data Sheet*. Available at: https://www.sdilink.com/Specsheet/VERIFACT%20A.pdf

Lyons, G.W., Hart, C.R. and Raspet, R. (2021) 'As the Wind Blows: Turbulent Noise on Outdoor Microphones', *Acoustics Today*, 17(4), pp. 29–38.

Makino, S., Sawada, H., Mukai, R. and Araki, S. (2005) 'Blind Source Separation of Convolutive Mixtures of Speech in Frequency Domain', *IEICE Transactions*, 88-A, pp. 1640–1655. Available at: https://doi.org/10.1093/ietfec/e88-a.7.1640

Manamperi, W., Abhayapala, T.D., Zhang, J. and Samarasinghe, P.N. (2022) 'Drone Audition: Sound Source Localization Using On-Board Microphones', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, pp. 508–519. Available at: https://doi.org/10.1109/TASLP.2022.3140550

Manamperi, W.N., Abhayapala, T.D., Samarasinghe, P.N. and Zhang, J. (2024) 'Drone audition: Audio signal enhancement from drone embedded microphones using multichannel Wiener filtering and Gaussian-mixture based post-filtering', *Applied Acoustics*, 216, p. 109818. Available at: https://doi.org/https://doi.org/10.1016/j.apacoust.2023.109818

MathWorks (2024) *wrapTo180 - MATLAB Documentation*. Available at: https://www.mathworks.com/help/map/ref/wrapto180.html (Accessed: 11.12.2024).

Matsumoto, M. and Naono, H. (1989) 'Stereo Zoom Microphone For Consumer Video Cameras', *IEEE Transactions on Consumer Electronics*, 35(4), pp. 759–766.

Merseyside Police (2022) *Boy, 15, sentenced for murder of Ava White*. Available at: https://www.merseyside.police.uk/news/merseyside/news/2022/july/boy-15-sentenced-for-murder-of-ava-white/ (Accessed: 07.05.2025).

Meyer, E., Neumann, E.-G. and Taylor, J.M. (1972) *Physical and applied acoustics : an introduction*. New York: Academic Press.

Neumann (2024) *What is Self-Noise (or Equivalent Noise Level)?* Available at: https://www.neumann.com/en-us/knowledge-base/neumann-im-homestudio/homestudio-academy/what-is-self-noise-or-equivalent-noise-level (Accessed: 13.05.2025).

Nodac, T. (2019) *OCB-CM40 CCTV Microphone Built-in AGC Circuit: Specifications*. Available at: https://www.nodactechnology.com/product/ocb-cm40-cctv-microphone-built-in-agc-circuit (Accessed: 13.05.2025).

Olson, H.F. and Preston, J. (1949) 'Single-Element Unidirectional Microphone', *Journal of the Society of Motion Picture Engineers*, 52(3), pp. 293–302. Available at: https://doi.org/10.5594/J12528

Owens, A. and Efros, A.A. (2018) 'Audio-Visual Scene Analysis with Self-Supervised Multisensory Features', *Computer Vision – ECCV 2018*. Available at: https://doi.org/10.1007/978-3-030-01231-1_39

Palacios, J. (1964) 'Inverse-square law in the theory of relativity', *Electronics and Power*, 10(10), pp. 362–363. Available at: https://doi.org/10.1049/ep.1964.0340

Parra, L.C. and Spence, C. (2000) 'Convolutive Blind Source Separation of Non-Stationary Sources', *IEEE Transactions on Speech and Audio Processing*, 8(3), pp. 320–327. Available at: https://doi.org/10.1109/89.841218

Paulraj, A., Roy, R. and Kailath, T. (1985) 'Estimation Of Signal Parameters Via Rotational Invariance Techniques- Esprit', *Nineteenth Asilomar Conference on Circuits, Systems and Computers*, 6–8 Nov. 1985. pp. 83–89. Available at: https://doi.org/10.1109/ACSSC.1985.671426

Pierce, A.D. and Beyer, R.T. (1989) 'Acoustics: An Introduction to Its Physical Principles and Applications. 1989 Edition', *The Journal of the Acoustical Society of America*, 87(4), pp. 1826–1827. Available at: https://doi.org/10.1121/1.399390

Prinz, M. and Ewert, S. (2020) 'Synthesis of real world drone signals based on lab recordings', *Acta Acustica*, 4, p. 45. Available at: https://doi.org/10.1051/aacus/2020056

Qualcomm Technologies, I. (2021) *Voice User Interface (Voice UI) Design Guide*. San Diego, CA: Qualcomm Technologies, Inc.

Ramirez, H. (2021) *DJI M300 Hover Accuracy Test*. Available at: https://www.youtube.com/shorts/IEQOqAijklk (Accessed: 02.04.2024).

Rode (2025a) *Lavalier Datasheet*. Available at: https://edge.rode.com/pdf/page/297/modules/986/lavalier_datasheet.pdf (Accessed: 28.03.2025).

Rode (2025b) *NTG-2 Datasheet*. Available at: https://edge.rode.com/pdf/page/342/modules/1186/ntg2_datasheet.pdf (Accessed: 28.03.2025).

Ruochen, W., Yuhong, Z. and Wei, Z. (2014) 'Acoustic Zooming Based on Real-Time Metadata Control', *Proceedings of IC-NIDC 2014*.

Beamforming-Based Acoustic Source Localization and Enhancement for Multirotor UAVs. 3–7 Sept. 2018. Available at: https://doi.org/10.23919/EUSIPCO.2018.8553514

Salvati, D., Drioli, C., Ferrin, G. and Foresti, G.L. (2020) 'Acoustic Source Localization From Multirotor UAVs', *IEEE Transactions on Industrial Electronics*, 67(10), pp. 8618–8628. Available at: https://doi.org/10.1109/TIE.2019.2949529

Schmidt, R. (1986) 'Multiple emitter location and signal parameter estimation', *IEEE Transactions on Antennas and Propagation*, 34(3), pp. 276–280. Available at: https://doi.org/10.1109/tap.1986.1143830

Schröter, H., Escalante, A.N., Rosenkranz, T. and Maier, A.K. (2022) 'Deepfilternet2: Towards Real-Time Speech Enhancement on Embedded Devices for Full-Band Audio', *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5.

Schulte-Fortkamp, B. and Jordan, P. (2023) 'Soundscape: The Holistic Understanding of Acoustic Environments', in Schulte-Fortkamp, B., Fiebig, A., Sisneros, J.A., Popper, A.N. and Fay, R.R. (eds.) *Soundscapes: Humans and Their Acoustic Environment.* Cham: Springer International Publishing, pp. 49–79.

Schultz-Amling, R., Kuech, F., Thiergart, O. and Kallinger, M. (2010) 'Acoustical Zooming Based on a Parametric Sound Field Representation', *Audio Engineering Society Convention Paper 8120*, pp. 1–9.

Shah, H. and Bhole, D. (2025) 'AI-Enhanced Beamforming: Advancements and Challenges in Spatial Audio and Voice Processing', *2025 IEEE International Conference on Electro Information Technology (eIT)*, 29–31 May 2025. pp. 289–294. Available at: https://doi.org/10.1109/eIT64391.2025.11103691

Shannon, C.E. (1949) 'Communication in the Presence of Noise', *Proceedings of the IRE*, 37(1), pp. 10–21. Available at: https://doi.org/10.1109/JRPROC.1949.232969

Shlens, J. (2014) 'A Tutorial on Independent Component Analysis', *ArXiv*. Available at: https://doi.org/10.48550/arXiv.1404.2986

Smita, S., Biswas, S. and Solanki, S. (2007) 'Audio Signal Separation and Classification: A Review Paper', 3297.

Southworth, G.C. (1946) 'Principles and Applications of Waveguide Transmission', *Bell System Technical Journal*, 25(1), pp. 74–100.

Strang, G. (2006) *Linear Algebra and Its Applications*. Brooks Cole.

Stroud, S., Jones, K.O., Edwards, G., Robinson, C., Ellis, D. and Chandler-Crnigoj, S. (2023) 'Robust Audio Zoom for Surveillance Systems: A Beamforming Approach with Reduced Microphone Array', *37th International Conference on Information Technologies (InfoTech-2023)*, Bulgaria, 20–21 Sept. 2023. IEEE Conference, Rec. #58 664, pp. 1–4. Available at: https://doi.org/10.1109/InfoTech58664.2023.10266894

Sylvestre-Williams, N. (2020) 'An update to ANSI/ASA S12.2-2019: Criteria For evaluating room noise—Where to go next?', *The Journal of the Acoustical Society of America*, 148(4_Supplement), pp. 2705–2705. Available at: https://doi.org/10.1121/1.5147496

Taal, C.H., Hendriks, R.C., Heusdens, R. and Jensen, J. (2010) 'A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech', *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, USA. pp. 4214–4217. Available at: https://doi.org/10.1109/ICASSP.2010.5495701

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

The Guardian (2022) *Boy, 14, guilty of murder of Liverpool schoolgirl Ava White*. Available at: https://www.theguardian.com/uk-news/2022/may/24/boy-14-guilty-of-murder-liverpool-schoolgirl-ava-white (Accessed: 07.05.2025).

Thiergart, O., Kowalczyk, K. and Habets, E.A.P. (2014) 'An acoustical zoom based on informed spatial filtering', *International Workshop on Acoustic Signal Enhancement (IWAENC 2014)*, Antibes, France, 2014. IEEE. Available at: https://doi.org/10.1109/iwaenc.2014.6953348

Thompson, R.C. (1997) *Principles of Vibration and Sound*. London: Taylor & Francis Ltd.

Tzinis, E., Venkataramani, S. and Smaragdis, P. (2019) 'Unsupervised Deep Clustering for Source Separation: Direct Learning from Mixtures Using Spatial Information', *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12–17 May 2019. pp. 81–85. Available at: https://doi.org/10.1109/ICASSP.2019.8683201

Tzinis, E., Wang, Z. and Smaragdis, P. (2020a) 'Sudo RM -RF: Efficient Networks for Universal Audio Source Separation', *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, 21–24 Sept. 2020. pp. 1–6. Available at: https://doi.org/10.1109/MLSP49062.2020.9231900

Tzinis, E., Wisdom, S., Hershey, J.R., Jansen, A. and Ellis, D.P.W. (2020b) 'Improving Universal Sound Separation Using Sound Classification', *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4–8 May 2020. pp. 96–100. Available at: https://doi.org/10.1109/ICASSP40776.2020.9053921

Tzinis, E., Wisdom, S., Jansen, A., Hershey, S., Remez, T., Ellis, D. and Hershey, J. (2021) 'Into the Wild with AudioScope: Unsupervised Audio-Visual Separation of On-Screen Sounds', *ICLR 2021*. pp. 1–27.

Van Waterschoot, T., Joos Tirry, W. and Moonen, M. (2013) 'Acoustic Zooming by Multimicrophone Sound Scene Manipulation', *Audio Engineering Society*, 61.

Veen, B.D.V. and Buckley, K.M. (1988) 'Beamforming: a versatile approach to spatial filtering', *IEEE ASSP Magazine*, 5(2), pp. 4–24. Available at: https://doi.org/10.1109/53.665

Visaton GmbH & Co. (2023) *FR 10 HM full-range loudspeaker — technical data sheet (Rev. 1.8)*. Available at: https://www.visaton.de/sites/default/files/dd_product/FR%2010%20HM_4898_4899_0.pdf (Accessed: 22.04.2024).

Wang, D.L. and Brown, G.J. (1999) 'Separation of speech from interfering sounds based on oscillatory correlation', *IEEE Transactions on Neural Networks*, 10(3), pp. 684–697. Available at: https://doi.org/10.1109/72.761727

Wang, L. and Cavallaro, A. (2020) 'A Blind Source Separation Framework for Ego-Noise Reduction on Multi-Rotor Drones', *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 28, pp. 2523–2537. Available at: https://doi.org/10.1109/taslp.2020.3015027

Wang, L. and Cavallaro, A. (2022) 'Deep-Learning-Assisted Sound Source Localization From a Flying Drone', *IEEE Sensors Journal*, 22(21), pp. 20828–20838. Available at: https://doi.org/10.1109/JSEN.2022.3207660

Wang, L., Clayton, M. and Rossberg, A.G. (2023) 'Drone audition for bioacoustic monitoring', *Methods in Ecology and Evolution*, 14(12), pp. 3068–3082. Available at: https://doi.org/https://doi.org/10.1111/2041-210X.14234

Webb, A.M., Reynolds, C., Chen, W., Reeve, H., Iliescu, D.-A., Lujan, M. and Brown, G. (2019) 'To Ensemble or Not Ensemble: When does End-To-End Training Fail?', p. arXiv:1902.04422. Available at: https://ui.adsabs.harvard.edu/abs/2019arXiv190204422W.

Weisstein, E.W. (2024) *Azimuth Angle - MathWorld*. Available at: https://mathworld.wolfram.com/Azimuth.html (Accessed: 11.12.2024).

Wiener, N. (1949) *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. The MIT Press.

Wilson, P.F. (2017) 'Multiple Sources in a Reverberant Environment: The "Cocktail Party Effect"', *2017 International Symposium on Electromagnetic Compatibility - EMC EUROPE, Angers, France*. Available at: https://doi.org/10.1109/EMCEurope.2017.8094822

Yen, B., Yamada, T., Itoyama, K. and Nakadai, K. (2024) 'A Performance Assessment on Rotor Noise-Informed Active Multidrone Sound Source Tracking Methods', *Drones*, 8(6), p. 266. Available at: https://www.mdpi.com/2504-446X/8/6/266.

Yost, W.A., Dye, R.H. and Sheft, S. (1996) 'A simulated "cocktail party" with up to three sound sources', *Perception & Psychophysics*, 58(7), pp. 1026–1036. Available at: https://doi.org/10.3758/BF03206830

Zhang, W., Chang, X., Boeddeker, C., Nakatani, T., Watanabe, S. and Qian, Y. (2022) 'End-to-End Dereverberation, Beamforming, and Speech Recognition in A Cocktail Party', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–16. Available at: https://doi.org/10.1109/TASLP.2022.3209942

Zoom Corporation (2022) *H3-VR_2 Operation Manual*. Available at: https://zoomcorp.com/media/documents/E_H3-VR_2.pdf (Accessed: 10.02.2024).

Appendix A: Publications Resulting from the Research

A.1 Journal Publications

Stroud, S., Jones, K.O., Edwards, G., Robinson, C., Ellis, D. and Chandler-Crnigoj, S. (2024) 'Advancing Audio Surveillance in Simulated Environments: Real-World Soundscapes and Targeted Noise Detection through Enhanced Beamforming Techniques', Journal of Advances in Engineering and Technology, 1(2). Presented at the 2024 SLIIT International Conference on

Engineering and Technology (SICET 2024), Sri Lanka. Available at: https://doi.org/10.54389/GCMF9805

A.2 Conference Publications

Stroud, S., Jones, K.O., Edwards, G., Robinson, C., Ellis, D. and Chandler-Crnigoj, S. (2023) 'Robust Audio Zoom for Surveillance Systems: A Beamforming Approach with Reduced Microphone Array', 37th International Conference on Information Technologies (InfoTech-2023), Bulgaria. Available at: https://doi.org/10.1109/InfoTech58664.2023.10266894

Stroud, S., Jones, K., Edwards, G., Robinson, C., Chandler-Crnigoj, S. and Ellis, D. (2024) 'Enhancing Environmental Sound Recognition in Digital Simulations: A Novel Approach to Beamforming and Signal Identification', International Conference on Computer Systems and Technologies 2024 (CompSysTech 2024), Ruse, Bulgaria. Available at: https://doi.org/10.1145/3674912.3674940

Stroud, S., Jones, K.O., Edwards, G., Robinson, C., Ellis, D. and Chandler-Crnigoj, S. (2024) 'Developing Audio Zoom in Virtual Environments: Real-World Soundscapes and Targeted Noise Detection', 38th International Conference on Information Technologies (InfoTech-2024), Bulgaria. Available at: https://doi.org/10.1109/InfoTech63258.2024.10701361

Appendix B: MATLAB Code

B.1 Overview

Scope and availability: This appendix provides selected MATLAB code excerpts illustrating the main simulation workflow, a text-based procedural listing of the pipeline, and a function index mapping each module to its role. Due to space constraints, the complete MATLAB project is provided in the associated GitHub repository with the link provided below.

B.2 GitHub Repository

https://github.com/stroud330/byromstreet-mvdr-audiozoom/releases/tag/v1.0-thesis

Software requirements: MATLAB R2024a. Toolboxes used by this codebase include the Phased Array System Toolbox (MVDR beamforming), Signal Processing Toolbox (spectrogram/STFT), Image Processing Toolbox (mask smoothing), and wrapTo180 support

Reproducibility note: The repository release above is the version corresponding to the thesis submission. To run the model, execute ByromStreetmodular.m from the project root with the Audio subfolder present.

Package contents.

Entry script : ByromStreetmodular.m

Expected folder structure (relative to the entry script):

- Audio/Music5secs.wav
- Audio/Speech5secs.wav
- Audio/Dronenoise5secs.wav
- *.m function files in the same folder

B.3 Examples of Code Excerpts

```
% Virtual Crime Scene - Byrom Street Campus 14.04.2025 | Steve Stroud
%% Clear all
clc; clear; close all;
%% DIMENSIONS
dimensions = define_dimensions();
%% CREATE 3D SCENE
[surfaces, world_objects] = create_3D_scene(dimensions);
%% PROMPT TO MOVE ARRAY
[array_location_choice,dimensions, input_line] = ...
set_microphone_array_location(dimensions);
% Check if the user wants to quit
    if strcmpi(array_location_choice, 'q') || strcmpi(input_line, 'q')
        disp('Exiting program.');
        return; % Exit the program
    end
%% CREATE MICROPHONE ARRAY + MICROPHONES
[array_shape_choice,mic_coordinates, dimensions, mic_array_group, micsOn] ...
= choose_microphone_array_type(dimensions);
% Check if the user wants to quit
    if strcmpi(array_shape_choice, 'q')
        disp('Exiting program.');
        return; % Exit the program
    end
%% CREATE A NOISE REDUCTION MICROPHONE ARRAY
[mic_coordinates_NR]=create_NR_microphone_array(dimensions);
%% CREATE GRID
[dimensions, grid_size_choice, input_line] = choose_grid_size(dimensions);
 % Check if the user wants to quit
    if strcmpi(grid_size_choice, 'q') || strcmpi(input_line, 'q')
        disp('Exiting program.');
        return; % Exit the program
    end
%% CREATE SOUND SOURCES
[spk_location_choice,speaker_coordinates,input_line] = set_speaker_coordinates();
[sound_source_group, speaker_coordinates]...
= create_sound_sources(dimensions, spk_location_choice, speaker_coordinates);
    if strcmpi(spk_location_choice, 'q') || strcmpi(input_line, 'q')
          disp('Exiting program.');
         return; % Exit the program
    end
%% CREATE THE SOUNDWAVES
[soundwave_choice, custom_soundProfile] = set_soundwaves();
if strcmpi(soundwave_choice, 'q')
      disp('Exiting program.');
      return; % Exit the program
end
[start_point, end_point, dimensions, soundwave_group, soundProfile, red_line_start, ...
red_line_end, ...
collision_array, reflection_group, collision_coords, collision_group, collision_array2,
collision_coords2, collision_array3, collision_coords3, collision_array4,
collision_coords4, reflection_data]...
= create_soundwaves(dimensions, speaker_coordinates, soundwave_choice, custom_soundProfile,
world_objects, surfaces);
%% CREATE ZOOMED 3D SCENE
[building_group, street_group] = create_zoomed_3D_scene(dimensions);
%% CHOOSE THEN FILL CHOSEN GRID
[beamform_grid_choice, Gridgroup] = choose_fill_chosen_grid(dimensions);
```

Figure B.1. Audio Zoom Matlab Modular Script (1 of 2).

```
% Check if the user wants to quit
    if strcmpi(beamform_grid_choice, 'q')
        disp('Exiting program.');
        return; % Exit the program
    end
%% CREATE INTERACTIVE LEGEND
createInteractiveLegend(mic_array_group, building_group, street_group, ... sound_source_group,
soundwave_group,collision_group, reflection_group, Gridgroup)
%% DISTANCE AND DELAY CALCULATIONS
distanceSpkMics = distanceSpkMics(speaker_coordinates, mic_coordinates);...
% Calculate the distance between the sound sources and the mics in meters
(distanceSpkMics 9 x 16)
timeDelay = timeDelay(speaker_coordinates, mic_coordinates);...
% Calculate the average time delays for each microphone, based on the distances from all 9
speakers (timeDelay 16 x 1)
timeDelayBetweenMics = timeDelayBetweenMics(mic_coordinates, timeDelay);...
% Calculate the time delay between each pair of microphones in seconds
(timeDelayBetweenMics 16 x 16)
timeDelaySpkMics = timeDelaySpkMics(speaker_coordinates, mic_coordinates);...
% Calculate the time delay between each speaker and each microphone (timeDelaySpkMics 9 x 16)
micDistances = micDistances(mic_coordinates);...
% Calculate difference in meters between each Mic. (0.16m or 16 cm between mic 1 and 2)
(micDistances 16 x 16)
%% CREATE VIRTUAL AUDIO SIGNALS
%[ALLSPKMics, fs, speedOfSound] = create_virtual_audio_signals(distanceSpkMics, ...
% mic_coordinates, speaker_coordinates, micsOn, soundProfile);
%% IMPORT REAL WORLD AUDIO SIGNALS
[ALLSPKMics, fs, speedOfSound, drone_noise] = import_audio_signals(distanceSpkMics,
mic_coordinates, ...
speaker_coordinates, micsOn, soundProfile, reflection_data);
%% MVDR BEAMFORMER
% Check if any mic is on if
any(micsOn)
    % Proceed with beamforming
if ~strcmp(beamform_grid_choice, 'q') % Only proceed if user_input is not 'q'
    [final_filtered_signal, beamformer_audio, beamformer, azimuth_grids]  =
beamformer_MVDR...
    (fs, dimensions, beamform_grid_choice, mic_coordinates, speaker_coordinates,
ALLSPKMics);
end else
    % Handle the situation when all mics are off
    warning('All microphones are turned off. Beamforming cannot proceed!');
end
%% DISPLAY SUMMARY
fprintf('\n\nSUMMARY:\n');
fprintf('------------------------------------------------\n');
fprintf(['Figure 1 = Full-size 3D Scene\nFigure 2 = 3D Scene zoomed into chosen grid (with
visibility toggle)\nFigure 3 = Microphone array 3D plot\nFigure 4 = Virtual audio signals from
the 16 microphones on the array\nFigure 5 = Polar plot of the steered beamformer pattern\nFigure
6 = Polar plot of the chosen array shape Pre Beamforming\nFigure 7 = Beamformer and Filter
plots\n']);
fprintf(['Upon successful beamforming, 3 audio files in .wav format will be stored in the
active directory\n1: pre beamforming, ' '2: beamformed array, 3: enchanced beamformed
array.\n']);
```

Figure B.2. Audio Zoom MATLAB Modular Script (2 of 2).

B.4 Key Code Tables

Table B.1 provides the dimensions used by the MATLAB function that creates the Exemplar Houses simulation.

Table B.1. MATLAB 'dimensions' initialisation code

```
function dimensions = define_dimensions()
%% Define the dimensions of the virtual crime scene
dimensions.heightgrid = 15;                          % height of the grid
dimensions.roomLength = 50;                          % length of scene
dimensions.roomWidth = 80;                           % width of scene
dimensions.roomHeight = 50;                          % height of scene
dimensions.arrayLength = 0.7;                        % length of the mic array
dimensions.arrayWidth = 0.7;                         % width of the mic array
dimensions.zMin_x = -0.01; %-0.01;                   % Lowest point of (x)
dimensions.zMax_x = 1.3;                             % Highest point (x)
dimensions.zMin_y = -0.01; %-0.01;                   % Lowest point of (y)
dimensions.zMax_y = -0.3;                            % Highest point of (y)
dimensions.speakerLength = 0.5;                      % length of the speaker
dimensions.speakerWidth = 0.5;                       % width of the speaker
dimensions.speakerHeight = 1.8;                      % Height of the speaker
dimensions.centreX = dimensions.roomLength /2;       % x centre of the scene
dimensions.centreY = dimensions.roomWidth /2;        % y centre of the scene
dimensions.centreZ = dimensions.roomHeight /2;       % z centre of the scene
dimensions.corners(1).x = 0;                         % x top left corner
dimensions.corners(1).y = dimensions.roomWidth;      % y top left corner
dimensions.corners(2).x = dimensions.roomLength;     % x top right corner
dimensions.corners(2).y = dimensions.roomWidth;      % y top right corner
dimensions.corners(3).x = 0;                         % x bottom left corner
dimensions.corners(3).y = 0;                         % y bottom left corner
dimensions.corners(4).x = dimensions.roomLength;     % x bottom right corner
dimensions.corners(4).y = 0;                         % y bottom right corner
dimensions.arraycentreX = dimensions.roomLength/2;   % default X array
dimensions.arraycentreY = dimensions.roomWidth/2;    % default Y array
dimensions.arraycentreZ = 1;                         % default Z array
dimensions.arrayRadius = 0.52;                       % Radius of circular array
```

The MATLAB code prompts the user to choose the starting location for the microphone array. The code is provided in Table B.2.

Table B.2. MATLAB code to prompt an array move.

```
%% PROMPT TO MOVE ARRAY
[array_location_choice,dimensions, input_line] = set_microphone_array_location
      (dimensions);
% Check if the user wants to quit
    if strcmpi(array_location_choice, 'q') || strcmpi(input_line, 'q')
        disp('Exiting program.');
        return; % Exit the program
    end
```

Table B.3 provides the MATLAB code to prompt the user to select a different geometric microphone array shape. The user has a choice of Square, Circular, Octagonal, Cross and Six-pointed Star.

Table B.3. MATLAB code to select an array type.

```
%% CREATE MICROPHONE ARRAY + MICROPHONES
[array_shape_choice,mic_coordinates, dimensions, mic_array_group, micsOn] =
choose_microphone_array_type(dimensions);
% Check if the user wants to quit
    if strcmpi(array_shape_choice, 'q')
        disp('Exiting program.');
        return; % Exit the program
    end
```

Table B.4 provides the MATLAB code to prompt the user to select the number of sound sources for the simulation and their location. The default setting is nine speakers, each in their own grid. Eight 'speakers' or sound sources are equilaterally spaced around the edge of the sound pickup target grid with a centre speaker underneath the array. To simulate custom scenes, the user can choose custom Cartesian coordinates for each of the nine sound sources or turn some off altogether.

Table B.4. MATLAB code to prompt the creation of sound sources.

```
%% CREATE SOUND SOURCES
[spk_location_choice,speaker_coordinates,input_line] = set_speaker_coordinates();
[sound_source_group, speaker_coordinates] = create_sound_sources(dimensions,
spk_location_choice, speaker_coordinates);
 if strcmpi(spk_location_choice, 'q') || strcmpi(input_line, 'q')
    disp('Exiting program.');
    return; % Exit the program
 end
```

Table B.5 displays the MATLAB code to prompt the user to create custom sound waves and choose the Sound Pressure Level (SPL) at 1 meter from the sound source, the azimuth angle, elevation angle, horizontal and vertical beam width for each sound coming from each source. Alternatively, the user can choose a default set of predetermined sound waves to set up repeatable experimental conditions quickly.

Table B.5. MATLAB code to prompt the creation of sound waves and reflections.

```
%% CREATE THE SOUNDWAVES
[soundwave_choice, custom_soundProfile] = set_soundwaves();
 if strcmpi(soundwave_choice, 'q')
        disp('Exiting program.');
        return; % Exit the program
 end
[start_point, end_point, dimensions, soundwave_group, soundProfile,
red_line_start, red_line_end, collision_array, reflection_group, collision_coords,
collision_group, collision_array2, collision_coords2, collision_array3, colli-
sion_coords3, collision_array4, collision_coords4, reflection_data] =
create_soundwaves(dimensions, speaker_coordinates, soundwave_choice, custom_sound-
Profile, world_objects, surfaces);
```

Table B.6 illustrates the code that triggers the various functions that perform distance and time delay calculations. These calculate the distances between each sound source and each microphone on the array, as well as the distances between each microphone on the array. The time delays between the microphones and the individual sound sources, as well as the delay between the microphones, are also calculated and stored in a MATLAB matrix array.

Table B.6. MATLAB code to run distance and delay calculation functions.

```
%% DISTANCE AND DELAY CALCULATIONS
distanceSpkMics = distanceSpkMics(speaker_coordinates, mic_coordinates);
% Calculate the distance between the sound sources and the mics in meters
(distanceSpkMics 9 x 16)
timeDelay = timeDelay(speaker_coordinates, mic_coordinates);
% Calculate the average time delays for each microphone, based on the distances
from all 9 speakers (timeDelay 16 x 1)
timeDelayBetweenMics = timeDelayBetweenMics(mic_coordinates, timeDelay);
% Calculate the time delay between each pair of microphones in seconds
(timeDelayBetweenMics 16 x 16)
timeDelaySpkMics = timeDelaySpkMics(speaker_coordinates, mic_coordinates);
% Calculate the time delay between each speaker and each microphone
(timeDelaySpkMics 9 x 16)
micDistances = micDistances(mic_coordinates);
% Calculate difference in meters between each Mic.
(0.16m or 16 cm between mic 1 and 2) (micDistances 16 x 16)
```

The code that runs the MVDR Beamformer function is shown in Table B.7. This code focuses the microphone arrays' response towards the grid of the user's choice.

Table B.7. MATLAB code to run the beamformer and filter function.

```
%% MVDR BEAMFORMER
% Check if any mic is on
if any(micsOn)
    % Proceed with beamforming
if ~strcmp(beamform_grid_choice, 'q') % Only proceed if user_input is not 'q'
    [final_filtered_signal, beamformer_audio, beamformer, azimuth_grids]  =
beamformer_MVDR(fs, dimensions, beamform_grid_choice, mic_coordinates, speaker_co-
ordinates, ALLSPKMics);
end
else
    % Handle the situation when all mics are off
    warning('All microphones are turned off. Beamforming cannot proceed!');
end
```

Table B.8 illustrates the MATLAB code, which triggers the functions that perform additional filtering and plotting of the beamformed signal.

Table B.8. MATLAB code to run simple noise reduction and filter functions.

```
function noise_estimation = estimate_noise(audio)
    % Simple estimation of noise using non-speech segments (placeholder)
    noise_estimation = mean(real(audio(audio < 0.05)));
    % Threshold-based noise estimation on real part
end

function enhanced_audio = wiener_filter(audio, noise_estimation)
    % Wiener filter implementation for noise reduction
    noise_power = noise_estimation^2;
    signal_power = var(real(audio)); % Use real part for variance calculation
    gain = max(0, signal_power / (signal_power + noise_power));
    enhanced_audio = gain * real(audio); % Apply gain to real part of audio
end
```

Table B.9 illustrates how spectral masking and Wiener filtering employ an STFT to achieve noise reduction, employ interpolation to synchronise the reference and beamformed audio, and after utilising the spectral mask to smooth the signal, it is reconstructed using the Griffin & Lim (1983) method.

Table B.9. MATLAB code for the Wiener Filter and Spectral Mask in the Filter function.

```
% Step 1: Compute the STFT of the beamformed signal and reference signal
[S_mixture, ~, ~] = spectrogram(beamformed_audio, 1024, 512, 1024, fs);
[S_voice, ~, ~] = spectrogram(original_voice, 1024, 512, 1024, fs);

% Step 2: Interpolate reference spectrogram to match the mixture dimensions
voice_resized = interp2(f_voice, t_voice, abs(S_voice)', f, t, 'linear', 0)';

% Step 3: Compute the Wiener filter mask
power_mixture = abs(S_mixture).^2;
power_voice = magnitude_voice_resized.^2;
spectral_mask = power_voice ./ (power_voice + power_mixture + eps);

% Step 4: Smooth and soften the mask to reduce artifacts
spectral_mask = imgaussfilt(spectral_mask, 1); % Gaussian smoothing
spectral_mask = spectral_mask .^ 0.8;         % Softening factor

% Step 5: Apply the mask to the mixture's magnitude and reconstruct
magnitude_separated = spectral_mask .* abs(S_mixture);
separated_spectrogram = magnitude_separated .* exp(1j * angle(S_mixture));

% Step 6: Use the Griffin-Lim algorithm to reconstruct the time-domain signal
final_signal = griffin_lim(separated_spectrogram, 1024, 512, 1024, fs, 50);
```

Table B.10 displays the MATLAB code that undertakes the optional ICA sound separation functions that can be applied to the beamformed and filtered signal.

Table B.10. MATLAB code to run signal separation functions.

```
% Separation
final_separated_signal1 =
Separation1(beamformed_audio, beamform_grid_choice, fs)
final_separated_signal2 =
Separation2(final_separated_signal1, beamform_grid_choice, fs)
final_separated_signal3 =
Separation3(final_separated_signal2, beamform_grid_choice, fs)
```

Table B.11 provides some of the code within the ‘create 3D scene’ function. This example shows a target within the scene visualised as a red circle.

Table B.11. MATLAB code within the create 3D scene function.

```
% Draw a red circle symbolising crime scene
x = 16.7 + cos(linspace(0, 2*pi, 100)) * 1.3;
y = 32 + sin(linspace(0, 2*pi, 100)) * 1.3;
z = calculate_z_positions(x, y, dimensions);
% Create the patch
patch('XData', x, 'YData', y, 'ZData', z+0.01, 'FaceColor', 'r', 'EdgeColor',
'none', 'FaceAlpha', 0.4);

% Labeling, lighting, view adjustments, etc.
xlabel('X (metres)');
ylabel('Y (metres)');
zlabel('Z (metres)');
%title('BYROM WAY LJMU');
view(3); % Set 3D (3) or 2D (2) view
axis equal
light('Position', [20, 20, 30]);
light('Position', [50, 10, 30]);
hold off
```

B.5 Selected Pseudocode

Algorithm B.1. Top-level procedure (ByromStreetmodular.m)

```
PROCEDURE ByromStreetmodular()

    Clear workspace, console and figures

    1) Define global scene dimensions and defaults
       dimensions := define_dimensions()

    2) Build and render the full 3D scene
       (surfaces, world_objects) := create_3D_scene(dimensions)

    3) Select microphone array location
       (array_location_choice, dimensions, input_line) :=
set_microphone_array_location(dimensions)
       IF user quits THEN EXIT

    4) Select microphone array geometry and create mic coordinates
       (array_shape_choice, mic_coordinates, dimensions, mic_array_group, micsOn)
:= choose_microphone_array_type(dimensions)
       IF user quits THEN EXIT

    5) Create the noise-reduction reference array geometry (separate array)
       mic_coordinates_NR := create_NR_microphone_array(dimensions)

    6) Select the grid size around the array
       (dimensions, grid_size_choice, input_line) := choose_grid_size(dimensions)
       IF user quits THEN EXIT

    7) Select / set speaker locations, then render speaker objects
       (spk_location_choice, speaker_coordinates, input_line) :=
set_speaker_coordinates()
       (sound_source_group, speaker_coordinates) :=
create_sound_sources(dimensions, spk_location_choice, speaker_coordinates)
       IF user quits THEN EXIT

    8) Select soundwave profile (default or custom)
       (soundwave_choice, custom_soundProfile) := set_soundwaves()
       IF user quits THEN EXIT

    9) Create soundwave line visualisation and compute reflections/collisions
       (..., reflection_data) := create_soundwaves(dimensions, speaker_coordinates,
soundwave_choice,
                                                   custom_soundProfile,
world_objects, surfaces)

    10) Create a zoomed 3D scene view for the chosen grid region
        (building_group, street_group) := create_zoomed_3D_scene(dimensions)

    11) Select the beamforming target grid (1-9) and render it
        (beamform_grid_choice, Gridgroup) := choose_fill_chosen_grid(dimensions)
        IF user quits THEN EXIT

    12) Create interactive legend to toggle scene elements
        createInteractiveLegend(mic_array_group, building_group, street_group,
sound_source_group,
                                soundwave_group, collision_group, reflection_group,
Gridgroup)

    13) Compute geometry matrices: distances and delays
        distanceSpkMics      := distanceSpkMics(speaker_coordinates,
mic_coordinates)
        timeDelay            := timeDelay(speaker_coordinates, mic_coordinates)
```

```
        timeDelayBetweenMics := timeDelayBetweenMics(mic_coordinates, timeDelay)
        timeDelaySpkMics     := timeDelaySpkMics(speaker_coordinates,
mic_coordinates)
        micDistances         := micDistances(mic_coordinates)

    14) Create multichannel microphone signals by importing audio and applying
distance effects
        (ALLSPKMics, fs, speedOfSound, drone_noise) :=
import_audio_signals(distanceSpkMics, mic_coordinates,

speaker_coordinates, micsOn,

soundProfile, reflection_data)

    15) Apply MVDR beamforming to the selected grid direction and post-filter the
output
        IF any microphones enabled (micsOn) THEN
            (final_filtered_signal, beamformer_audio, beamformer, azimuth_grids) :=
                beamformer_MVDR(fs, dimensions, beamform_grid_choice,
mic_coordinates,
                                speaker_coordinates, ALLSPKMics)
        ELSE
            Print warning and terminate beamforming stage
        ENDIF

END PROCEDURE
```

Algorithm B.2. Microphone signal construction (import_audio_signals.m)

```
PROCEDURE import_audio_signals(distanceSpkMics, mic_coordinates,
speaker_coordinates, micsOn, soundProfile, reflection_data)

    fs := 48000 Hz
    c  := 343 m/s
    N  := 5 seconds * fs samples

    Load WAV files:
        music  := Audio/Music5secs.wav
        speech := Audio/Speech5secs.wav
        drone  := Audio/Dronenoise5secs.wav

    mic_signals[nMic, N] := 0

    FOR each speaker spk
        Choose source signal for this speaker (e.g., speech for target speaker,
music for interferers)
        FOR each microphone mic
            IF micsOn(mic) = 1 THEN
                distance := distanceSpkMics(spk, mic)
                delay_samples := round((distance / c) * fs)

                attenuation := 1 / distance^2           // inverse square
                spl_gain := function_of(soundProfile)    // SPL scaling for this
source

                delayed := apply_integer_sample_delay(source_signal, delay_samples)
                mic_signals(mic,:) += delayed * attenuation * spl_gain
            ENDIF
        END FOR
    END FOR

    Add drone noise to each microphone channel (scaled)
    Normalise to avoid clipping
    Save a reference mic-channel WAV (pre-beamforming)

    Return microphone signals as cell array ALLSPKMics and sampling frequency fs
END PROCEDURE
```

Algorithm B.3. MVDR beamforming + post-filtering (beamformer_MVDR.m)

```
PROCEDURE beamformer_MVDR(fs, dimensions, beamform_grid_choice, mic_coordinates,
speaker_coordinates, ALLSPKMics)

    1) Compute azimuth/elevation angles for microphones and grid directions
       azimuth_mics, elevation_mics := angles_from_geometry(mic_coordinates,
dimensions.centre)
       azimuth_grids, elevation_grids := angles_from_geometry(speaker_coordinates,
dimensions.centre)

    2) Create conformal array and MVDR beamformer object
       mic_positions := [3 x nMic] from mic_coordinates
       mic_array := phased.ConformalArray(ElementPosition = mic_positions)

       beamformer := phased.MVDRBeamformer(
           SensorArray = mic_array,
           PropagationSpeed = 343,
           Direction = [azimuth_grids(target_grid); elevation_grids(target_grid)],
           DiagonalLoadingFactor = (model default),
           WeightsOutputPort = true)

    3) Apply beamformer to multichannel audio matrix
       audio_matrix := stack(ALLSPKMics)
       (beamformed_audio, weights) := beamformer(audio_matrix)

    4) Compute and save a pre-beamformed reference signal (sum across microphones)
       pre_array_audio := sum(audio_matrix across microphones)
       Save Pre_Beamformed_Audio.wav

    5) Compute beam pattern (steered) and plot polar response
       pattern_mvdr := compute_mvdr_beam_pattern(beamformer, azimuth sweep,
elevation = 0)
       Plot polarpattern(...)

    6) Apply post-filter
       final_filtered_signal := Filter(beamformed_audio, beamform_grid_choice, fs)

    7) Plot time-domain signals and spectrograms; save output audio
       Save Beamformed_Audio_Grid_<grid>.wav
       Save Filtered_Speech_Grid_<grid>.wav

    Return (final_filtered_signal, beamformed_audio, beamformer, azimuth_grids)
END PROCEDURE
```

Algorithm B.4. Post-filter (Filter.m)

```
PROCEDURE Filter(beamformed_audio, beamform_grid_choice, fs)

    Load original voice reference: Audio/Speech5secs.wav
    Resample / convert to mono and equal-length alignment

    STFT analysis of mixture and reference
    Compute magnitude spectra
    Resize / align reference spectrogram to mixture spectrogram size

    Construct Wiener-style soft mask from power spectra
    Smooth mask (Gaussian) and apply soft exponent

    Apply mask to mixture magnitude; use mixture phase for initial reconstruction
    Refine phase using Griffin–Lim iterations

    Return reconstructed time-domain signal
    Save Filtered_Speech_Grid_<grid>.wav
END PROCEDURE
```

B.6 Function Index

Table B.12 lists all MATLAB .m files in the code package and summarises their role in the workflow.

Table B.12. Function Index for MATLAB package.

File	Purpose
ByromStreetmodular.m	Entry-point script that orchestrates the full simulation (scene setup, user prompts, signal generation, MVDR beamforming, plotting, and audio export).
Filter.m	Post-filter stage: spectral masking/Wiener-style filtering with Griffin–Lim reconstruction; writes filtered speech output.
Separation1.m	Auxiliary separation stage (utility function in the package).
Separation2.m	Auxiliary separation stage (utility function in the package).
Separation3.m	Auxiliary separation stage (utility function in the package).
beamformer_MVDR.m	Creates and applies the MVDR beamformer toward the chosen grid direction; plots beampatterns; calls post-filter; saves output audio.

File	Purpose
calculate_z_positions.m	Computes Z coordinates for ground slope or surface elevations from X/Y positions.
check_collision.m	Collision/reflection helper (first stage) for ray interaction with scene surfaces/objects.
check_collision2.m	Collision/reflection helper (second stage).
check_collision3.m	Collision/reflection helper (third stage).
check_collision4.m	Collision/reflection helper (fourth stage).
choose_fill_chosen_grid.m	Prompts for the target grid (1–9) and renders/marks the selected grid region.
choose_grid_size.m	Interactive grid-size selection; sets the coordinate grid region around the array.
choose_microphone_array_type.m	Interactive selection and construction of the microphone array geometry; outputs mic coordinates and enable/disable flags.
compute_beam_pattern.m	Beam pattern helper function used for plotting/analysis.
compute_mvdr_beam_pattern.m	Computes MVDR beam pattern across an azimuth sweep for plotting.

File	Purpose
createInteractiveLegend.m	Creates an interactive legend to toggle visibility of plotted groups/lines.
create_3D_scene.m	Builds the 3D scene geometry and surface/material definitions used for plotting and reflections.
create_512_size_grid.m	Creates the grid at the 5.12 m scale option used by the script.
create_NR_microphone_array.m	Creates the separate noise-reduction reference array geometry.
create_circular_microphone_array.m	Creates circular array geometry and returns microphone coordinates and plot handles/groups.
create_cross_microphone_array.m	Creates cross array geometry and returns microphone coordinates and plot handles/groups.
create_grid.m	Creates the default 3x3 grid coordinate set around the array.
create_half_size_grid.m	Creates a half-size version of the default grid.
create_octagonal_microphone_array.m	Creates octagonal array geometry and returns microphone coordinates and plot handles/groups.

File	Purpose
create_quarter_size_grid.m	Creates a quarter-size version of the default grid.
create_sound_sources.m	Creates the speaker/source objects/groups for 3D plotting.
create_soundwaves.m	Creates soundwave/ray visualisation and computes reflections/collision data based on scene geometry.
create_square_microphone_array.m	Creates square-perimeter array geometry and returns microphone coordinates and plot handles/groups.
create_star_microphone_array.m	Creates star array geometry and returns microphone coordinates and plot handles/groups.
create_third_size_grid.m	Creates a third-size version of the default grid.
create_zoomed_3D_scene.m	Creates a zoomed-in 3D view of the selected region for clearer visualisation.
define_dimensions.m	Defines scene dimensions, centre point and global configuration parameters.
distanceSpkMics.m	Computes distance matrix between speakers and microphones.

File	Purpose
estimate_noise.m	Noise estimation helper for spectral noise reduction methods.
fastICA.m	Independent Component Analysis implementation used by some separation utilities.
import_audio_signals.m	Constructs multichannel microphone signals by importing WAVs and applying distance-based attenuation and delays; adds drone noise; saves reference audio.
istft.m	Inverse STFT utility used by reconstruction/separation stages.
micDistances.m	Computes pairwise distance matrix between microphones.
name_save_audio_file.m	Utility to standardise filenames and save audio outputs.
noise_reduction.m	Alternative noise reduction utility (not necessarily used in the main ByromStreetmodular run path).
pointInPolygon.m	Geometry utility used for containment/intersection checks.
set_microphone_array_location.m	Interactive selection of where the microphone array is positioned in the scene.

File	Purpose
set_soundwaves.m	Selects default vs custom soundwave profile parameters (e.g., SPL/beamwidth).
set_speaker_coordinates.m	Interactive selection or assignment of speaker/source coordinates.
timeDelay.m	Computes propagation time delays (from geometry).
timeDelayBetweenMics.m	Computes pairwise time-delay matrix between microphones.
timeDelaySpkMics.m	Computes time delays from each speaker to each microphone.
toggleVisibility.m	Helper used by interactive legend callbacks.

Appendix C: Supplementary Figures

C.1 Additional Technical Specifications

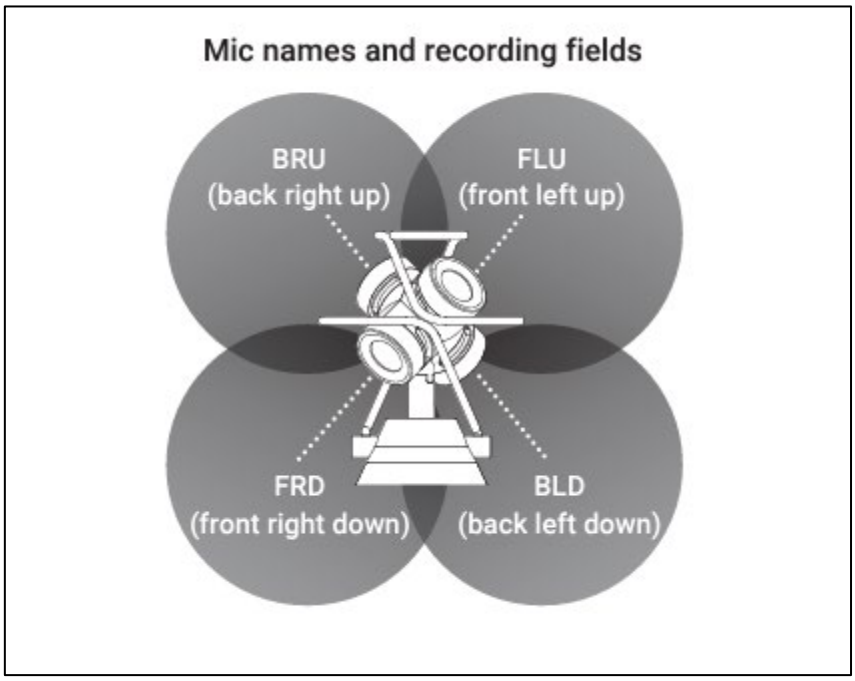

Figure C.1. A-Format ambisonics channel names
(Taken from Zoom Corporation (2022)).

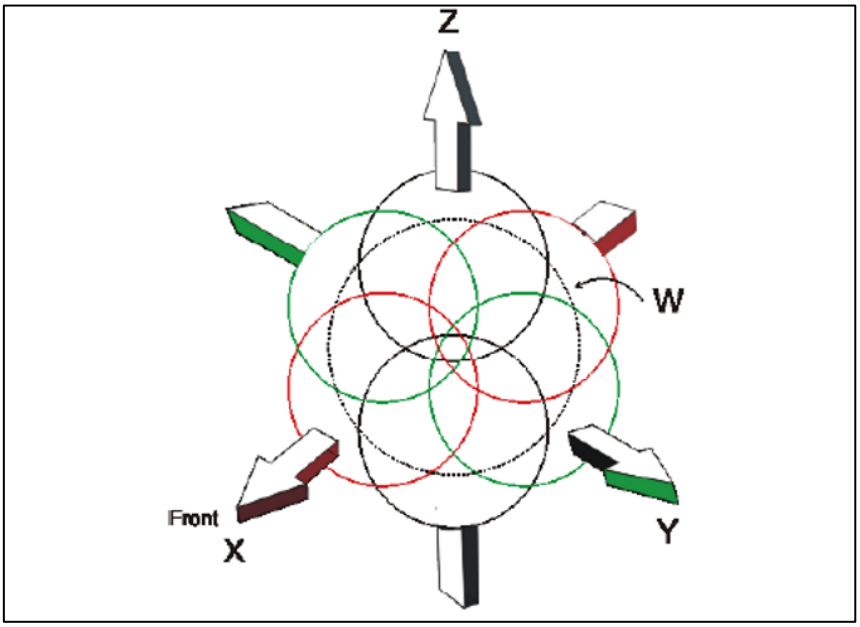

Figure C.2. B-format ambisonics capsule layout: three figure-of-eight plus one omni (Taken from Zoom Corporation (2022)).

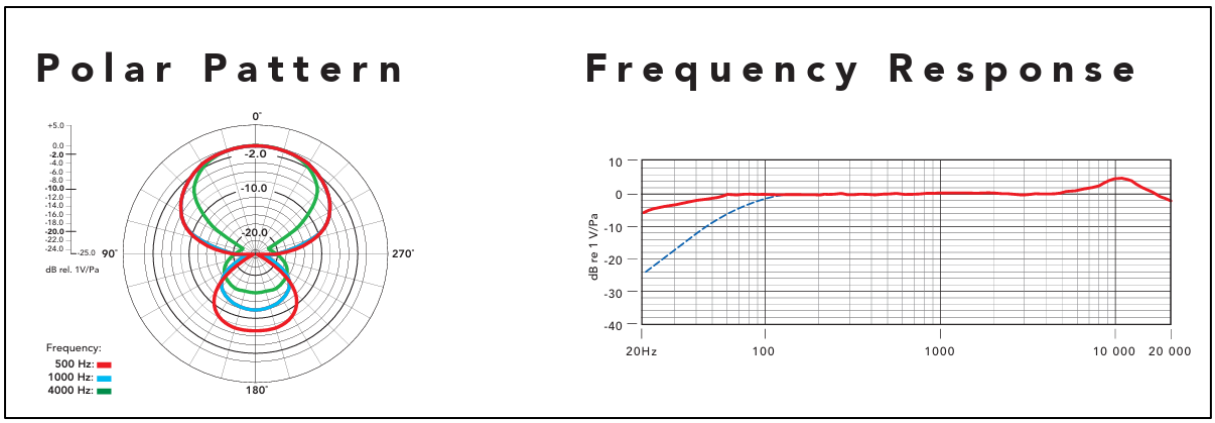

Figure C.3. Polar Pattern and Frequency Response of the Rode NTG-2 Directional Condenser Microphone (Rode, 2025b).

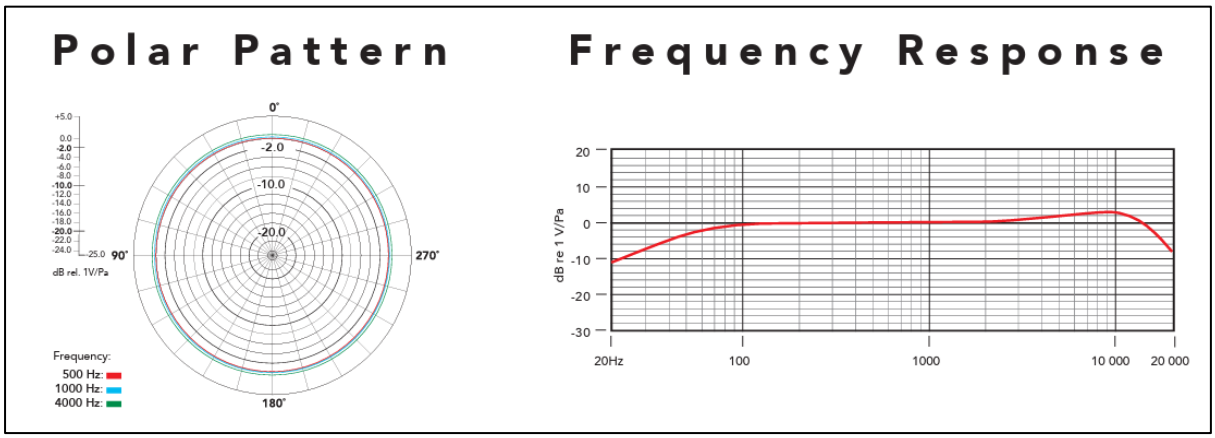

Figure C.4. Polar Pattern and Frequency Response of the Rode Lavalier Condenser Microphone (Rode, 2025a).

Appendix D: Awards Resulting from the Research

D.1 Awards Resulting from the Research

Figure D.1. Best Paper Award at SICET 2024 in SLIIT University, Malabe, Sri Lanka.

Figure D.2. First Place for Best Pitch Presentation at Liverpool John Moores University Post Graduate Research Day 2024.