EEE4022S Research Project
Final Project Report:
Laser Based Listening Device

Prepared by:
Molise Mokhakala

Prepared for:

Dr Stephen Paine

Department of Electrical and Electronics Engineering

University of Cape Town

Submitted to the Department of Electrical Engineering at the University of Cape Town
in partial fulfilment of the academic requirements for a
Bachelor of Science degree in Mechatronics Engineering

October 27, 2025

Keywords: laser microphone, optical vibrometry, speech enhancement, transimpedance amplifier, spectral subtraction, Wiener filtering, automatic speech recognition

Plagiarism Declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

2. I have used the **IEEE** convention for citation and referencing. Each contribution to, and quotation in, this report from the work(s) of other people has been attributed and properly cited. Any section taken from an internet source has been referenced to that source.

3. This report is my own work, and is in my own words (except where I have attributed it to others).

4. **I have not paid a third party to complete my work on my behalf. My use of artificial intelligence software has been limited to:** *coding help in VS Code, grammar correction, technical language refinement, and LaTeX formatting assistance.*

5. I have not allowed and will not allow anyone to copy my work with the intention of passing it off as his or her own work.

6. I acknowledge that copying someone else's assignment or essay, or part of it, is wrong, and declare that this is my own work.

Signed:

Molise K Mokhakala

October 27, 2025

______________________________ ________________________

Molise Mokhakala Date

Word count: 21793 _______________

Acknowledgements

> No man is an island entire of itself; every man is a piece of the continent, a part of the main.
>
> —*John Donne*

I would like to express my sincere gratitude to Dr. Paine for his continuous guidance, support, and valuable insights throughout this project. His expertise and encouragement were instrumental in its successful completion. I would also like to extend my appreciation to Tsepang, the Power Lab technician student, for his exceptional assistance in developing the test box—an essential component without which none of the experimental results would have been possible

Abstract

This project presents the design, implementation, and evaluation of a low-cost laser microphone system capable of remotely reconstructing intelligible speech from vibrating surfaces through optical sensing. Sound-induced surface vibrations modulate a 650 nm laser beam reflected from a target window, which is detected by a BPW34 photodiode and converted to electrical signals via a dual transimpedance amplifier (TIA) configuration. The dual-TIA architecture provides enhanced common-mode rejection (CMRR) to suppress ambient light interference, a critical improvement over single-channel designs. The analog front-end incorporates precision filtering and differential amplification stages, implemented on a custom PCB and powered via a standard 5 V USB interface, achieving a compact and portable form factor suitable for laboratory deployment.

Signal recovery is accomplished through a multi-stage digital signal processing pipeline combining FIR band-pass filtering (100–3400 Hz), Wiener filtering, spectral subtraction, and adaptive voice amplification. This hybrid approach balances noise suppression with speech preservation, achieving a Short-Time Objective Intelligibility (STOI) score of 0.157—a 15.7-fold improvement over unprocessed recordings. Material characterization revealed that a 3 mm glass window provided 0.53 dB higher SNR than Perspex due to superior optical reflectivity (8.0% vs. 7.8%) and mechanical rigidity. Automatic speech recognition testing using Whisper Medium demonstrated a word error rate (WER) of 15.0% on processed audio at 1 m distance, representing an 8.9% improvement over raw recordings and validating the system's practical utility for forensic and surveillance applications.

Range sensitivity analysis at distances of 1–3 m revealed non-linear performance trade-offs: while energy-based SNR increased with distance, perceptual intelligibility (STOI) degraded due to optical diffraction and beam coherence loss. The system successfully met all design specifications within a budget constraint of R2000, demonstrating that cost-effective laser microphone can achieve sufficient fidelity for automated transcription tasks. The project uses reproducible, open-source implementation integrating modern speech enhancement algorithms with classical optical sensing techniques. All hardware designs, PCB schematics, signal processing code, and experimental datasets are publicly available via GitHub, enabling reproduction and extension by future researchers.It advances the field by providing a reproducible, open-source implementation integrating modern speech enhancement algorithms with classical optical sensing techniques, and establishes quantitative benchmarks for low-cost laser microphone performance across multiple objective metrics.

Contents

List of Figures

ADC
Analog-to-Digital Converter – A device that converts an analog signal into its corresponding digital representation.

ASR
Automatic Speech Recognition – A technology that converts spoken language into text using machine learning or signal processing algorithms.

BOM
Bill of Materials – A comprehensive list of all components required for hardware assembly.

CMRR
Common-Mode Rejection Ratio – A measure of an amplifier's ability to reject common-mode signals, i.e., noise appearing simultaneously on both inputs.

DAC
Digital-to-Analog Converter – A circuit that converts digital values into continuous analog signals.

DSP
Digital Signal Processing – The numerical manipulation of signals, typically using algorithms to enhance, analyze, or transform data.

FIR
Finite Impulse Response – A type of digital filter with a finite duration impulse response, known for linear phase properties.

He–Ne Laser
Helium–Neon Laser – A gas laser operating typically at 632.8 nm, used for precise optical measurements due to its narrow linewidth and coherence.

IIR Infinite Impulse Response – A recursive digital filter type that uses feedback to achieve a desired frequency response.

LD Laser Diode – A semiconductor device that emits coherent light when forward biased.

LDV
Laser Doppler Vibrometer – A non-contact device that measures surface vibration by detecting the Doppler shift of reflected laser light.

PCB
Printed Circuit Board – A board that mechanically supports and electrically connects electronic components using conductive tracks and pads.

PIN Photodiode

Positive–Intrinsic–Negative Photodiode – A semiconductor light sensor with fast response and low noise characteristics.

SNR

Signal-to-Noise Ratio – The ratio between the desired signal power and the background noise power, expressed in decibels (dB).

STOI

Short-Time Objective Intelligibility – A quantitative metric (0–1 scale) assessing the intelligibility of processed or degraded speech signals.

TIA

Transimpedance Amplifier – An operational amplifier circuit that converts input current (from a photodiode) into a proportional output voltage.

VCSEL

Vertical-Cavity Surface-Emitting Laser – A semiconductor laser that emits light perpendicular to its surface, often used for compact optical systems.

WER

Word Error Rate – A standard metric for evaluating automatic speech recognition accuracy by comparing predicted text to a reference transcript.

Wiener Filter

– An adaptive filter that minimizes the mean-square error between estimated and desired signals, commonly used for noise reduction.

Whisper

– An open-source speech-to-text model by OpenAI designed for multilingual transcription and robust performance in noisy audio.

Window Reflectivity

– The proportion of light reflected from a window's surface, influencing the strength of optical feedback in laser microphones.

Photodiode

– A semiconductor device that converts light into an electrical current proportional to the incident optical power.

Chapter 1

Introduction

This report details the research, design, and development of a laser-based listening device referred to as a laser microphone in the paper, focusing on the hardware, optical, and signal processing components required to reconstruct audio from distant vibrating surfaces. The study aims to implement a functional prototype capable of capturing and processing acoustic information through non-contact optical sensing. The project was undertaken as part of the final year BSc (Eng) Mechanical Engineering degree in the Department of Electrical and Electronic Engineering at the University of Cape Town.

1.1 Background to the Study

Sound detection using light-based methods has become an emerging field of interest due to its ability to capture acoustic information without direct physical contact. Conventional microphones, although effective, are limited by distance, environmental noise, and the need to be close to the sound source.

Laser-based acoustic sensing provides a non-contact alternative by exploiting the minute vibrations induced on a reflective surface by sound waves. A coherent laser beam directed at the surface captures these vibrations through optical interference, which can then be processed into an audible signal. Early research into laser microphones has demonstrated their potential in surveillance, industrial monitoring, and structural diagnostics, but challenges remain to achieve reliable, noise-resistant, and cost-effective implementations.

This project builds on existing research by designing and testing a laser microphone system tailored for remote acoustic signal acquisition. The focus is on creating a practical laboratory-scale prototype capable of detecting and reconstructing speech or sound signals reflected off common surfaces, using photodiode-based sensing and digital signal processing algorithms for reconstruction.

1.2 Problem Statement and Objectives of this Study

Problem Statement: Despite decades of development in laser listening technology for covert surveillance and remote acoustic sensing, practical implementations remain constrained by high costs[1], low signal-to-noise ratios due to weak optical modulation and ambient interference, and limited accessibility for research applications. Existing systems either sacrifice real-world applicability for laboratory sensitivity or require specialised optical equipment beyond typical engineering budgets. This project addresses the fundamental challenge of whether integrated optical-electronic-algorithmic design can achieve transcription-grade speech intelligibility (WER <20%, STOI >0.15) from a low-cost (<R2000) laser microphone system operating at practical distances (1–3 m) on common reflective surfaces, while maintaining reproducibility through open-source implementation suitable for forensic, surveillance, and academic applications.

1.2.1 Problems to be Investigated

The primary issues to be investigated in this study include:

- Determining the effectiveness of a laser-based microphone system in reconstructing acoustic signals from a vibrating surface.
- Evaluating how different surface materials, distances, and environmental conditions affect signal accuracy and sensitivity.
- Designing an optical receiver system capable of translating laser reflections into usable electrical signals.
- Implementing digital signal processing algorithms for noise reduction, filtering, and intelligible sound reconstruction.
- Implement a transcription model as part of the device.
- Assessing system limitations, including range and surface profile.

1.2.2 Purpose of the Study

The purpose of this study is to develop and experimentally validate a laser-based listening device capable of converting surface vibrations into audible signals. The motivation arises from the potential applications of optical acoustic sensing in security surveillance, industrial inspection, and scientific experimentation. By investigating this approach, the study seeks to bridge theoretical understanding and practical implementation of laser vibrometry for acoustic reconstruction.

The significance of this research lies in demonstrating a proof-of-concept that utilises electronic principles, optical sensing, signal processing, and embedded system design to explore an innovative method of sound detection. The outcomes will provide valuable insights into signal quality optimisation, cost-effective design strategies, and environmental robustness for future implementations.

1.3 Scope, Limitations and Assumptions

1.3.1 Scope

The scope of this project includes:

- Theoretical and experimental investigation of optical vibration detection for sound reconstruction.
- Design and construction of a laser-based sensing system using commercially available components.
- Implementation of analogue front-end circuitry for photodiode signal conditioning.
- Development of digital signal processing algorithms in MATLAB to reconstruct and evaluate the detected signals.
- Performance evaluation under controlled laboratory conditions and transcription intelligibility.

1.3.2 Limitations

The limitations of this study include:

- The experimental setup is limited to short-range indoor measurements (up to a few meters).
- Environmental disturbances such as air turbulence, background light, and mechanical vibrations may affect measurement accuracy.

- The project is constrained by the available budget and laboratory resources, with the budget capped at R2000.
- Real-time embedded processing was not implemented in this phase; signal analysis was conducted offline using MATLAB.
- The study assumes a coherent and stable laser source with negligible beam divergence over the test distance.

1.4 Plan of Development

The report begins by introducing the underlying principles of optical sound detection and prior research on laser vibrometry and acoustic signal reconstruction. Chapter 2 presents a comprehensive literature review on laser-based sensing systems and relevant digital signal processing techniques.

Chapters 3 and Chapter 4 outlines the methodology adopted for the hardware and optical design, detailing component selection, circuit implementation, and system integration. Chapter 5 discusses the digital signal processing workflow, including noise reduction, filtering, and time-domain signal reconstruction. This chapter also shows implementation of a translation model and its assessment against other models.

Chapter6 then quantifies the system performance across the dectection, signal processing and transcription domains

The report concludes with a discussion of the system's performance, limitations, and recommendations for further development, including real-time processing and enhanced environmental robustness.

Chapter 2

Literature Review

This chapter situates the laser microphone project within a broader research and technological context by surveying the relevant literature. Although the basic concept of recovering sound from optical vibrations has existed for decades, its applications in both security and scientific settings present unique challenges and opportunities. Rather than simply duplicating past designs, this project aims to critically assess existing methods and technologies to inform the development of a system tailored for modern use cases, including remote surveillance, acoustic sensing, and machine learning–based audio reconstruction.

The objective of this review of the literature is therefore not only to outline technical precedents but also to identify the motivations, limitations, and innovations that have shaped this field. Particular emphasis is placed on the challenges associated with optical signal acquisition and noise suppression, especially in uncontrolled environments where ambient disturbances can dominate the true audio signal.

The chapter begins by reviewing existing implementations of laser microphone systems, ranging from intelligence-grade surveillance tools to academic and hobbyist prototypes. Then it delves into the photonic components that form the heart of such systems, namely lasers, photodetectors, and transimpedance amplifiers, highlighting performance trade-offs and noise sources. Following this, a detailed discussion of signal acquisition and processing challenges is presented, with a focus on the nature of optical vibrometry signals and the types of interference commonly encountered in this context.

To address these challenges, Section 5 reviews the traditional filtering and enhancement techniques used in optical audio processing pipelines. This is followed by a review of emerging machine learning–based methods, which promise more adaptive and robust denoising in complex environments. Finally, relevant research sources are cited to provide a roadmap for further exploration and benchmarking, setting the stage for the technical implementation that follows.

In doing so, this chapter aims to provide both a theoretical foundation and practical motivation for the system design choices made in subsequent stages of the project.

2.1 Overview and History of Laser Microphones

Laser microphones, also referred to as *laser interferometers*, *laser vibrometers*, or *laser accelerometers*, are non-contact devices that detect vibrations on a surface caused by sound waves. By directing a coherent light beam, typically a laser, onto a reflective surface such as a glass window, minute deflections caused by acoustic pressure modulate the reflected beam. This modulated light is then converted into an electrical signal using a photodetector, effectively reconstructing the original audio waveform [2]. The human voice typically occupies the 300 Hz–3400 Hz range, which falls comfortably within the operational bandwidth of modern laser microphone systems [3].

The concept of transmitting or detecting sound via light predates the invention of the laser. In the 1880s, Alexander Graham Bell developed the *photophone*, which transmitted speech using a modulated beam of sunlight reflected from a vibrating diaphragm onto a selenium cell receiver whose electrical resistance varied with light intensity. Bell considered this device even more significant than his telephone, describing it as capable of revealing “discoveries undreamed of just now” [4]. His work laid the foundation for photoacoustic and photothermal detection methods that underpin modern optical communication and remote acoustic sensing.

The first practical use of light for covert listening emerged decades later in 1947, when Léon Theremin developed the *Buran* system in the Soviet Union. Instead of a laser, it employed a low-power infrared beam to detect sound-induced vibrations in window glass at distances of up to 500 m. The system, reportedly used by the KGB to surveil U.S., British, and French embassies in Moscow, is widely regarded as the precursor to the modern laser microphone [5]. By the 1990s, advancements in laser diodes, optical detectors, and signal processing revived interest in laser-based eavesdropping systems, though their popularity later declined with the advent of digital and network-based surveillance technologies.

Operation In operation, laser eavesdropping relies on the fact that acoustic pressure waves cause solid surfaces—such as windows, cups, or walls—to vibrate. When a laser beam is reflected off such a surface, these vibrations induce phase and amplitude modulations in the reflected light. A remote receiver converts these modulations back into an audio signal. According to FCDO Services [1], this allows stand-off listening from distances exceeding 500 m without physical intrusion, making it a preferred technique for technical surveillance due to its low detectability.

Moses and Trout [4], describe a simplified academic demonstration of this principle, where a low-cost laser pointer and solar cell were used to reproduce the sound reflected from a vibrating window. In this setup, fluctuations in the curvature of the glass surface altered the divergence of the reflected beam, creating measurable intensity variations at the photodetector. Even minimal window vibrations produced audible signals after amplification, effectively mimicking the operation of professional-grade surveillance systems.

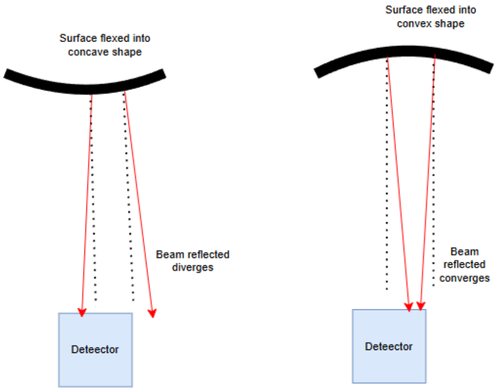

Surface flexed into concave shape
Beam reflected diverges
Deteector
Surface flexed into convex shape
Beam reflected converges
Deteector

Figure 2.1: Laser_beam divergence and convergence due to window vibration

Modern laser microphone technology has evolved significantly beyond its espionage origins. Contemporary implementations employ advanced receivers, low-noise amplifiers, and digital demodulation techniques, including double-heterodyne detection, to enhance sensitivity and suppress noise. High-end interferometric systems can now detect sub-nanometer vibrations, with documented sensitivity down to angstrom-scale displacements, and operate under challenging conditions such as fog or poor visibility [6]. Commercial laser vibrometers, originally designed for precision vibration analysis in engineering and materials testing, are conceptually similar to laser microphones and can be repurposed for acoustic monitoring, structural health assessment, or surveillance applications. .

2.2 Review of Existing Laser Microphone Systems

Research into laser microphones has a range of system designs that differ in their underlying optical principles, component choices, and intended applications. Early efforts were closely linked to the development of *laser Doppler vibrometers* (LDVs), which exploit the Doppler effect to measure vibration velocity from reflected light. Cai [7] demonstrates that the frequency shift between reference and reflected beams could be directly related to surface vibration velocity, enabling non-contact audio capture. Interferometric techniques such as scanning Laser Doppler Vibrometry (LDV) have also been employed to visualise sound fields non-intrusively by detecting refractive index variations in air [8][9] [10]. Such systems demonstrate the same optical phase modulation principle underlying modern laser microphones. While LDVs are highly sensitive, their reliance on coherent, mirror-like reflective surfaces limits deployment in practical settings where surfaces are rough or non-ideal. This highlights a recurring trade-off in laser microphone research: systems with high sensitivity often lack robustness in uncontrolled environments.

Alternative approaches leverage *pulsed laser vibrometers* (PLVs). Wang *et al.* [11] proposes a pulsed laser system coupled with photo-electromotive force (photo-EMF) sensors, incorporating multiple reflections of the probe beam to amplify the signal before detection. This design achieved higher sensitivity than conventional vibrometers, but at the cost of increased optical complexity and alignment challenges. The need for stable beam redirection across multiple bounces raises questions about system scalability and robustness outside laboratory conditions.

A third class of designs is based on the *self-coupling effect of laser diodes* (LDs), where variations in refractive index caused by acoustic pressure modulate the laser's own wavelength. Daisuke [12] reports that this approach, using a distributed feedback LD with integrated photodiode detection, provided improved sensitivity compared to traditional vibrometry. Importantly, the self-coupling method reduces the reliance on external interferometers, making it potentially more compact and cost-effective. However, performance still heavily depends on surface reflectivity and environmental stability, raising concerns about reproducibility under different operating conditions.

Commercial systems have also emerged, often marketed as *invasive or semi-invasive monitoring devices.* For example, Argo's multi-use infrared stethoscope system incorporates a laser microphone option for room surveillance, offering wide field of view monitoring and AM audio transmission at 840 nm **argo**. While such designs demonstrate practical integration, their technical specifications—such as wideband coverage and eye safety compliance—are rarely reported in detail, making independent evaluation difficult. Moreover, their use of infrared wavelengths prioritises surveillance robustness over high-fidelity sound reproduction, reflecting a divergence between commercial and academic design priorities.

A further commercial example is the *PKI 3000 Laser Microphone* by PKI Electronic Intelligence [13]. It employs an eye-safe 1550 nm Class 1 laser (< 10 mW) and an integrated interferometric receiver to extract speech information from Doppler shifts in the backscattered beam. Unlike earlier systems limited by reflection angle, the PKI 3000 achieves greater angular independence and can operate through closed glass surfaces at distances of 5 – 150 m. The device includes a built-in low-light CCD camera for alignment and selectable bandwidths between 150 Hz – 7 kHz, with analog and digital outputs for recording or monitoring. Though PKI notes higher-end variants (PKI 3100/3200) yield better audio quality, the PKI 3000 demonstrates a practical, field-ready implementation that integrates optical, electronic, and acoustic processing within a single portable unit, highlighting how commercial designs trade fidelity for deployability.

Across these systems, thematic contrasts emerge. LDV- and PLV-based systems prioritise sensitivity but require ideal reflective surfaces and precise alignment, whereas self-coupling LD systems sacrifice robustness for compactness and lower cost. Commercial implementations emphasize operational feasibility over academic metrics such as signal-to-noise ratio (SNR) or distortion.The optical source and receiver subsystem largely determines system performance. Laser wavelength, coherence length, power output, photodetector sensitivity, and bandwidth directly influence SNR, dynamic range, and environmental robustness. While early systems relied on gas lasers for coherence stability, advances in semiconductor laser diodes and integrated photonics now enable more compact, energy-efficient, and eye-safe field-deployable designs

2.2.1 Laser Source Selection

Traditional laboratory-grade systems typically employ Helium–Neon (HeNe) lasers at 632.8 nm due to their narrow linewidth and high coherence length, ideal for interferometric detection and phase-sensitive vibrometry [9]. However, such lasers are bulky and power-hungry, limiting their portability. Modern alternatives favour semiconductor-based sources, including distributed feedback (DFB) and vertical-cavity surface-emitting lasers (VCSELs), which provide tunable wavelengths, compact form factors, and stable output suitable for self-coupling and heterodyne detection schemes [7], [12].

2.2.2 Wavelength Considerations

Several studies emphasise that the type of laser diode employed can significantly affect signal capture fidelity. Distributed-feedback (DFB) laser diodes are commonly used due to their narrow linewidth and wavelength stability, which enhance sensitivity to minute surface vibrations [12]. However, their relatively high cost and power consumption may limit applicability in compact or low-budget systems. Conversely, standard edge-emitting laser diodes are more cost-effective but suffer from broader linewidths and reduced coherence, limiting their performance in interferometric setups.

The selection of operating wavelength also introduces critical trade-offs. Infrared wavelengths around 850 nm are commonly deployed in commercial systems for surveillance, partly because they balance eye safety, availability of components, and adequate reflectivity across common window glass materials **argo**. However, studies have demonstrated that longer wavelengths, such as 1550 nm, offer improved eye safety margins for higher power operation, albeit at the cost of reduced efficiency in some photodetectors [14]. Moreover, shorter wavelengths (e.g., 633 nm He-Ne lasers) may achieve higher sensitivity due to reduced speckle size on target surfaces, but their increased scattering makes them less robust in outdoor or turbulent conditions [15]. These wavelength-dependent trade-offs highlight the absence of a universally optimal solution, with system design often reflecting specific application constraints.

2.2.3 Photodetectors and Transimpedance Amplifier (TIA) Design

High-speed photodiodes combined with low-noise transimpedance amplifiers (TIAs) form the core of most laser microphone receiver architectures. The photodiode must provide sufficient bandwidth to cover the acoustic range (20 Hz–20 kHz) while maintaining a low noise-equivalent power (NEP) for adequate signal-to-noise ratio (SNR). Avalanche photodiodes (APDs) offer internal gain and enhanced sensitivity compared to PIN types but introduce excess multiplication noise and require higher bias voltages [16], [17], [18].

The TIA stage plays a central role in determining overall system noise and stability. As Carter notes, “Noise is a purely random signal that cannot be completely eliminated, but its effects can be minimized through careful circuit design” [19]. In a TIA, dominant sources include op-amp voltage and current noise, as well as the thermal (Johnson) noise of the feedback resistor. According to Carter, “the total output noise is the root-sum-square of the voltage and current noise contributions,” and the total input capacitance forms a pole with the feedback impedance, limiting bandwidth and affecting stability [19].

Thus, increasing the feedback resistor improves gain but raises thermal noise, revealing an inherent trade-off between sensitivity, bandwidth, and noise. When coupled with high-capacitance photodiodes, wideband TIAs must balance these parameters carefully. Recent CMOS implementations demonstrate sub-pA/$\sqrt{\text{Hz}}$ noise levels and MHz-range bandwidths, enabling compact, low-noise front ends for laser-based acoustic sensing [17], [20].

Taken together, the literature reveals that laser microphone component selection involves balancing competing requirements: wavelength choice affects safety, reflectivity, and coherence; photodiode selection trades sensitivity against noise; and TIA design must optimise bandwidth while minimising thermal and shot noise. Despite these insights, comparative experimental studies remain scarce. Most reports evaluate component choices in isolation rather than benchmarking across full system performance. Consequently, it is still unclear which combinations of laser wavelength, detector type, and amplifier topology yield the most robust results in practical acoustic sensing scenarios. Addressing this gap would require standardised testing protocols that account for both controlled laboratory conditions and variable real-world environments.

2.3 Signal Acquisition and Noise Challenges

Noise reduction remains one of the most critical and persistent challenges in the development of laser microphone systems. The issue arises from the fact that noise sources are highly variable across environments and application contexts, and they often change dynamically over time. As Chen [21] notes, early work on speech enhancement can be traced back to patents in the 1960s, where analogue implementations of spectral subtraction were first described. Although these methods introduced foundational concepts for suppressing additive noise, subsequent decades have shown that no single approach provides robust performance under all conditions.

A key difficulty is the diversity of noise sources affecting laser microphones. Mechanical vibrations and wind contribute low-frequency distortions, while optical noise is introduced through laser jitter and surface roughness of the reflective target. Electronic noise sources, including thermal noise in photodiodes and shot noise, further degrade the acquired signal. Cai [7] emphasised that surface characteristics strongly affect signal quality: rough or non-uniform surfaces scatter light irregularly, leading to reduced signal-to-noise ratio (SNR) and limiting fidelity at higher acoustic frequencies. Daisuke [22] further reported that many laser microphone systems produce SNRs that are too low for practical recording, highlighting a significant gap between laboratory feasibility and application-level performance.

Multiple strategies have been explored to address these challenges. Physical approaches, such as increasing the number of microphones in an array, allow for beamforming and spatial filtering [21]. However, in most scenarios only a single optical channel is available, restricting the use of such methods. Traditional digital techniques, including Wiener filtering and spectral subtraction, offer relatively low computational complexity but introduce trade-offs between noise suppression and speech distortion. For instance, Wiener filters reduce noise effectively under stationary conditions but degrade intelligibility when noise is highly nonstationary [23]. Similarly, spectral subtraction methods perform well in moderate noise but often produce musical noise artifacts in real-world environments [24].

These limitations underscore an unresolved debate in the literature: whether the focus of laser microphone noise reduction should prioritise intelligibility (e.g., for surveillance applications), fidelity (e.g., for recording and archival purposes), or robustness (e.g., for deployment in uncontrolled environments). Each objective demands different trade-offs in filter design, and attempts to optimise one dimension often degrade another. Chen [21] argues that this “triple constraint” of noise reduction, intelligibility, and naturalness remains a fundamental bottleneck across speech enhancement systems, and laser microphones inherit this challenge in amplified form due to their inherently low SNR.

Overall, while numerous techniques have been developed to ameliorate noise in laser microphone systems, the literature reveals a persistent gap: few comparative studies evaluate algorithms under standardised benchmarks or realistic field conditions. Without consistent metrics and datasets, it remains difficult to determine which acquisition and filtering approaches best translate from theory to practical deployment.

2.4 Signal Acquisition and Noise Challenges

Signal acquisition in laser microphones is uniquely difficult because the desired acoustic signal is typically weak and easily masked by multiple sources of noise. Unlike conventional microphones, which rely on direct pressure-to-voltage transduction, laser microphones must infer acoustic vibrations indirectly through optical reflections. This multi-stage process introduces mechanical, optical, and electronic noise, all of which complicate robust audio capture.

2.4.1 Noise Sources

Noise in laser microphone systems can be broadly classified into three categories. First, *mechanical noise* includes vibrations from mounts, air turbulence, or wind, which couple into the reflective surface and distort the recovered signal. Second, *optical noise* arises from speckle effects, surface roughness, and laser jitter, which cause fluctuations in the reflected beam [7]. Third, *electronic noise* is generated within the photodetection Cain, including shot noise, dark current, and thermal noise in photodiodes and transimpedance amplifiers. The combination of these noise sources often results in low signal-to-noise ratios (SNR), a limitation repeatedly cited as preventing laser microphones from achieving recording-quality audio [22].

2.4.2 Historical and Conventional Approaches

Noise suppression has been a focus of acoustic research for decades. Chen [21] traced early speech enhancement methods to Schroeder’s patents on analogue spectral subtraction in the 1960s. While such approaches provided foundational tools for noise removal, they assumed stationary additive noise, an assumption rarely satisfied in real-world optical vibrometry. Subsequent studies demonstrated that single-channel algorithms such as Wiener filtering and spectral subtraction could reduce noise effectively in controlled conditions, but at the expense of introducing speech distortion when the noise environment was nonstationary [23], [24].

Invasive alternatives, such as integrating contact stethoscopes with optical receivers (e.g., Argo surveillance systems **argo**), have been used as fallback solutions to mitigate signal loss when optical paths are obstructed. While effective in specific contexts, these designs compromise the defining non-contact advantage of laser microphones, highlighting the difficulty of balancing robustness and fidelity.

2.4.3 Methodological Tensions

A persistent tension in the literature concerns the trade-offs between intelligibility, fidelity, and robustness. For surveillance-oriented applications, intelligibility of speech is often prioritised over naturalness of sound, whereas high-fidelity recording applications require preservation of acoustic detail beyond 15 kHz. Chen [21] argued that these objectives cannot be simultaneously optimised with conventional filtering, leading to what he describes as a “triple constraint.” In laser microphones, this problem is exacerbated by inherently weaker signals compared to direct-contact microphones, making traditional techniques less effective.

2.5 Filtering and Signal Enhancement Techniques

Filtering and signal enhancement are central to the viability of laser microphones, as they determine whether the weak, noise-corrupted optical signals can be transformed into intelligible audio. Over the past decades, both classical signal processing methods and modern machine learning (ML) approaches have been proposed. This section critically reviews these techniques, highlighting methodological strengths, contradictions, and research gaps.

2.5.1 Traditional Filtering Techniques

Traditional approaches form the first line of defence in most laser microphone systems, typically applied in analogue circuitry or as post-ADC digital processing.

Analog Filtering

Analogue pre-processing stages, such as low-pass, band-pass, and notch filters, are often integrated within transimpedance amplifier (TIA) circuits to suppress high-frequency noise and prevent aliasing prior to digitisation. Cai [7] demonstrated that reducing high-frequency optical jitter using analogue low-pass filtering improved SNR in laser Doppler vibrometry setups. However, such filtering is inherently rigid, offering limited adaptability to time-varying acoustic or environmental conditions.

2.6 Digital Filtering (Post-ADC)

Digital FIR and IIR Filtering

Post-ADC, finite impulse response (FIR), and infinite impulse response (IIR) filters are widely employed. FIR filters preserve linear phase and are effective for speech-band isolation, while IIR filters are computationally efficient but may distort phase [25]. Case studies using equiripple FIR bandpass designs between 100 Hz and 4.5 kHz demonstrated effective suppression of electrical hum and low-frequency noise [26] . Yet, the computational cost of high-order FIR filters remains a concern for real-time applications.

Comb and Adaptive Filters

Periodic noise, such as power line hum or mechanical resonances, is often addressed using comb filters or ALE methods. Adaptive notch structures have been shown to eliminate harmonic noise efficiently, but their performance degrades when the target speech signal lies near the noise harmonics, requiring high filter orders [21]. Comb filters offer a simpler alternative but must be tuned carefully to avoid distorting broadband speech components. MATLAB, SciPy, CMSIS-DSP can be used to implement the aforementioned filters.

Spectral Subtraction and Wiener Filtering

Spectral subtraction, a technique first patented by Schroeder and later expanded in Chen’s work [24], assumes additive, quasi-stationary noise and removes it by estimating noise spectra during silence. While effective in moderate noise conditions, this approach often produces “musical noise” artifacts in nonstationary settings. Wiener filtering provides a theoretically optimal linear solution in the mean-square error sense, but its performance depends heavily on accurate estimation of noise and speech spectra. Benesty *et al.* [23] highlighted the trade-off: maximising noise reduction typically introduces significant speech distortion. Widely linear Wiener filters have been proposed to reduce distortion by exploiting conjugate symmetry in speech signals [27], yet they remain computationally complex. Sager *et al.* [26] demonstrates that a combination of spectral subtraction and gating with gating used right after the spectral subtraction, resulted in a significant reduction in noise than when just spectral subtraction is used. This brings.

Kalman and Wavelet-Based Filtering

Kalman filters model the audio signal as a dynamic system and are particularly effective under time-varying noise conditions, but they require accurate state-space models, which are difficult to obtain in optical-acoustic contexts [21]. Wavelet denoising provides a time-frequency localised alternative, excelling at suppressing transient noise while preserving speech onsets. However, the choice of wavelet basis critically affects performance, and there is little consensus on optimal configurations for laser microphones.

2.6.1 Machine Learning-Based Approaches

More recently, machine learning methods have been introduced to overcome the rigid assumptions of traditional filtering.

Deep Neural Networks (DNNs)

Cai [7] proposed a two-stage DNN for laser microphone speech enhancement, separating power and phase components of the waveform. The model outperformed STFT-based baselines in perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI). However, it underperformed in log spectral distance (LSD), suggesting limited accuracy in fine spectral reconstruction. The reliance on paired clean and noisy data also raises questions about generalizability to unseen acoustic conditions.

Convolutional and Recurrent Models

Convolutional neural networks (CNNs) operating on spectrograms have been used for denoising, with Daisuke [22] demonstrating CNN-based speaker recognition using self-coupling LD microphones. Recurrent neural networks (RNNs) and long short-term memory (LSTM) models capture temporal dependencies in noise, and Cai [7] reported improved high-frequency reconstruction using LSTMs with cross-entropy loss. While promising, these models face challenges in latency and computational load, making real-time deployment on embedded platforms difficult.

Generative Adversarial Networks (GANs) and Lightweight Models

Speech enhancement GANs (SEGANs) [28] generate clean audio by adversarial training, often producing more natural-sounding speech than traditional filters. Meanwhile, lightweight RNN-based systems such as RNNoise [29] combine DSP with small neural networks, enabling real-time noise suppression at low cost. These methods suggest a pathway toward embedded deployment, but their performance on laser microphone data remains largely unexplored, as most have been validated only on conventional audio datasets.

Traditional filtering remains essential for baseline noise suppression, especially in resource-constrained or low-latency systems. However, their reliance on stationary noise assumptions and rigid filter structures limits performance in complex, real-world acoustic environments. Machine learning models demonstrate superior adaptability and perceptual quality but face challenges of dataset scarcity, computational overhead, and reproducibility. The literature reveals a critical gap: there is no standardised dataset of laser microphone recordings for benchmarking denoising algorithms. As a result, most ML-based studies adapt models from conventional speech enhancement, which may not capture the unique distortions introduced by optical-acoustic sensing. Bridging this gap will require community-driven datasets and hybrid methods that integrate the interpretability of classical filtering with the adaptability of machine learning.

2.6.2 Speech-to-Text Systems

Speech-to-text (STT) systems play a crucial role in forensic and intelligence work, particularly for analyzing and transcribing covert recordings obtained through hidden microphones, parabolic sensors, or optical systems. Transcription transforms spoken evidence into written form suitable for legal or investigative use, providing a permanent record that can be analyzed linguistically or acoustically [30]. However, conventional manual transcription—often performed by certified translators or forensic linguists—is time-consuming and vulnerable to human error or perceptual bias, especially when recordings are degraded, multilingual, or captured under challenging acoustic conditions [31].

As Loakes *et al.* [31] note, covert recordings frequently suffer from overlapping voices, background interference, or limited frequency response, which complicates manual interpretation and reduces evidential reliability. This has motivated growing forensic interest in Automatic Speech Recognition (ASR) and AI-driven STT systems as objective tools capable of consistent and scalable transcription. Yet, even advanced ASR systems face persistent challenges such as alignment inaccuracies and the risk of "force-aligning" incorrect transcripts to indistinct speech segments—a critical concern in forensic validation and chain-of-custody integrity.

Modern transcription systems have evolved significantly from analogue audio analyzers to fully AI-driven frameworks capable of synchronized digital processing, real-time captioning, and emotion recognition. Eftekhari [32] highlights how intelligent STT platforms enhance efficiency in domains such as emergency response and medical diagnostics—capabilities that can be extended to surveillance analysis where rapid transcript generation is essential. Nonetheless, system accuracy remains heavily dependent on acoustic conditions, microphone quality, and training datasets, requiring human verification and linguistic post-editing for evidentiary use.

Comparative evaluations by Matarneh *et al.* [33] categorize major ASR frameworks—including CMU Sphinx, Kaldi, HTK, Microsoft Speech API, and Dragon Medical—by their domain suitability. Open-source engines such as Kaldi and CMU Sphinx are favored in research and multilingual applications, while proprietary systems like Google Speech and Dragon Medical excel in domain-specific accuracy and noise resilience. These platforms collectively demonstrate the trade-off between transparency, adaptability, and recognition performance, crucial factors when selecting STT tools for forensic or intelligence contexts.

Beyond classical ASR, end-to-end speech-to-text translation (ST) frameworks have expanded the scope of automated transcription. Chen *et al.* [34] and Xu *et al.* [35] identify key advances such as pre-training, data augmentation, and knowledge distillation—techniques that enhance robustness in low-resource and cross-lingual scenarios typical of real-world surveillance audio. Similarly, Bansal *et al.* [36] demonstrated that translation models can be trained directly from raw audio-text pairs using unsupervised term discovery, a promising avenue for multilingual forensic processing where annotated corpora are unavailable.

While technological advances have improved recognition speed and cross-lingual accuracy, transcript readability and evidential clarity remain concerns. Jones *et al.* [37] observed that automatically generated transcripts are often harder to read and comprehend than those produced by humans, emphasizing the importance of linguistic post-processing and readability correction before legal submission. Thus, despite their efficiency, current STT systems still require expert review to ensure forensic admissibility and contextual accuracy.

In summary, speech-to-text technology has evolved from experimental recognition systems into a sophisticated toolset for forensic intelligence and surveillance analysis. By integrating traditional acoustic modeling with deep-learning-based transcription, modern systems provide unprecedented scalability and consistency. However, challenges in low-quality audio interpretation, model generalization, and transcript validation continue to limit full automation, underscoring the enduring importance of human oversight in forensic transcription workflows [30], [31], [32], [33], [34], [35], [36], [37].

Table 2.1: Summary of Speech Recognition and Speech-to-Text Methods with Real-World Examples **matarneh2017**, [38], [39], [40], [41], [42], [43], [44], [45]

System / Method	**Model / Approach**	**Key Features**	**Strengths / Notes**	**Examples**
Template-Based	Reference template matching [38], [40]	Matches input utterances with stored patterns using dynamic time warping or posterior-based distance metrics	Simple implementation and interpretable matching; effective for isolated-word recognition but limited to small vocabularies and speaker dependency	Early IVR systems, CMU Sphinx (legacy)
Knowledge-Based	Rule extraction from linguistic and phonetic features [33]	Uses manually defined syntactic or phonetic rules derived from domain data	Interpretable structure, useful for specialized domains or low-data scenarios; less scalable to large vocabularies or continuous speech	Dragon NaturallySpeaking (early versions)
Statistical (HMM)	Hidden Markov Models with Gaussian mixtures [33], [39]	Models temporal variability and state transitions in speech; MFCC or PLP front-end features	Reliable and robust for real-time ASR; adaptable to new languages via re-training; still forms hybrid back-end in Kaldi and HTK systems	Kaldi, Julius, HTK [42]
Neural Network-Based	DNN / RNN / LSTM acoustic models [33], [43], [44]	Learns deep hierarchical features and temporal dependencies; enables end-to-end acoustic-to-word mapping	High accuracy for large-vocabulary tasks and real-world speech; effective noise robustness; integration into cloud APIs	Google Speech API, Microsoft Azure Speech [43], [44]
End-to-End Seq2Seq	Encoder–decoder with attention or transformers [39], [41]	Directly maps speech waveforms to text or translation outputs; often pretrained on multilingual corpora	Removes manual alignment and HMM dependency; supports low-resource and multilingual input; strong noise resilience but high compute cost	OpenAI Whisper, Vosk, Meta SeamlessM4T, SpeechT5 [41]

2.6.3 Assessment of Speech-to-Text Models

It is essential to establish objective frameworks for evaluating system performance of STTs. While traditional assessments have focused on qualitative judgments or human transcription accuracy, quantitative metrics such as **Word Error Rate (WER)**, **Short-Time Objective Intelligibility (STOI)**, and **Signal-to-Noise Ratio (SNR)** now serve as standardized benchmarks for comparing the accuracy, intelligibility, and overall quality of automated transcription systems. These measures are particularly relevant in the context of **laser microphone recordings**, where acoustic distortions and optical artifacts can significantly affect both intelligibility and recognition accuracy.

Word Error Rate (WER)

Word Error Rate remains the most widely adopted metric for evaluating ASR system performance. WER quantifies transcription accuracy by measuring the minimum number of word-level edits required to transform the hypothesis transcript into the reference transcript. The metric is calculated as:

$$\text{WER} = \frac{S + D + I}{N} \times 100\% \tag{2.1}$$

where S represents substitutions, D denotes deletions, I indicates insertions, and N is the total number of words in the reference transcript.

WER provides a normalized measure for comparing recognition systems across datasets and languages. Ali and Renals [46] established foundational methods for WER estimation in modern ASR systems, introducing *estimated WER (e-WER)* techniques that enable performance prediction without requiring full reference transcripts—an approach particularly valuable in real-time forensic analysis.[1]

Recent work by Park *et al.* [47] and von Neumann *et al.* [48] refines WER definitions for long-form and multi-talker scenarios, addressing challenges that frequently arise in extended surveillance or conversational recordings. Meanwhile,Von *et al.* [48] specifically propose the Diarization-Invariant cpWER (DI-cpWER), a metric designed to isolate and quantify the effect of speaker attribution errors. Unlike the conventional concatenated minimum-permutation WER (cpWER), which is sensitive to speaker label mismatches, the DI-cpWER corrects these mismatches to reflect the error rate achievable if speaker labels were perfectly assigned. The difference between cpWER and DI-cpWER thus indicates the proportion of errors arising solely from speaker confusions Itoh *et al.* Apple Machine Learning Research [49] proposes humanized WER formulations emphasizing readability and accessibility, aligning error measurement more closely with human comprehension. Their solution (human evaluation word error rate)HEWER ignores the use of filler words in its approach different from WER which ignores punctuation and counts every words that is not identical to the reference as an error.

Short-Time Objective Intelligibility (STOI)

While WER measures textual accuracy, it does not capture perceived speech clarity. The **Short-Time Objective Intelligibility (STOI)** metric provides a complementary measure focused on the intelligibility of the enhanced speech itself. Developed by Taal *et al.* [50], [51], STOI evaluates the correlation between clean and processed speech in the time–frequency domain, yielding scores between 0 and 1, where higher values indicate better intelligibility.

The algorithm divides speech into short frames and computes correlation coefficients between spectral envelopes of the clean ($x_m(j)$) and degraded ($y_m(j)$) signals:

$$\text{STOI} = \frac{1}{MJ} \sum_{m=1}^{M} \sum_{j=1}^{J} \rho(x_m(j), y_m(j)) \tag{2.2}$$

where M is the number of frames, J the number of frequency bands, and ρ the correlation coefficient.

[1] https://github.com/qcri/e-wer

STOI has been shown to correlate strongly with subjective listening tests across noise, reverberation, and other degradations. Scores above 0.75 suggest good intelligibility, while values below 0.60 indicate the need for further enhancement. Implementations are available through the MATLAB Audio Toolbox [52] and official open-source resources [53].

Signal-to-Noise Ratio (SNR)

The **Signal-to-Noise Ratio (SNR)** quantifies the relative strength of the speech signal against background noise, expressed in decibels (dB):

$$\mathrm{SNR} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right) \tag{2.3}$$

where P_{signal} and P_{noise} denote the power of the clean speech and noise components, respectively.

Ephraim and Malah [54] pioneered *a priori* SNR estimation, forming the basis of Wiener and MMSE filters used in speech enhancement. Plapous *et al.* [55] refined these approaches for improved accuracy under non-stationary conditions, while Vondrášek and Pollák [56] demonstrated the influence of voice activity detection (VAD) on SNR estimation reliability. Input SNR values typically range from –5 dB to 15 dB, improving to 10–30 dB after enhancement—an improvement of roughly 10–20 dB, sufficient for most ASR systems. NIST [57] provides standardized procedures for measuring SNR consistency across datasets.

Comprehensive Performance Evaluation

As emphasized by Hu and Loizou [58], no single metric reliably captures perceptual quality under all degradation types. A holistic assessment must therefore combine multiple metrics such as WER, STOI, and SNR, along with subjective listening tests. Rix *et al.* [59] further standardized perceptual evaluation through PESQ (Perceptual Evaluation of Speech Quality), now widely used for speech quality benchmarking. Interactions between these metrics are nonlinear—systems may achieve high intelligibility (STOI) but still record high WER due to language model errors, or show SNR gains without corresponding intelligibility improvement due to artifacts. Therefore, multi-metric benchmarking remains essential to guide optimization and ensure reliable performance in forensic and surveillance contexts.

2.7 Environmental and Surface Limitations

While the optical and signal processing components of laser microphones have received substantial attention, their performance in real-world environments remains constrained by environmental factors and target surface properties. These limitations are particularly significant because they directly affect the reliability and generalizability of laboratory findings.

2.7.1 Surface Reflectivity and Roughness

The accuracy of optical-acoustic transduction depends strongly on the reflective properties of the target surface. Smooth, mirror-like surfaces maximise specular reflection and yield higher signal-to-noise ratios (SNR), whereas rough or diffuse surfaces scatter the incident laser beam, reducing coherence and increasing noise [7]. Rothberg *et al.* [15] highlighted that speckle effects caused by surface roughness introduce random intensity fluctuations, complicating phase-based detection schemes such as laser Doppler vibrometry. These effects are particularly problematic for surveillance applications, where the surface cannot be controlled.

2.7.2 Environmental Noise and Turbulence

Environmental factors such as wind, air turbulence, and ambient vibrations impose additional noise sources. For outdoor applications, turbulence alters the refractive index along the beam path, leading to beam wander and amplitude fading [15]. Mechanical vibrations of the laser mount or reflective object also couple into the acoustic signal, generating low-frequency artefacts. Unlike electronic or optical noise, these disturbances are highly nonstationary, making them difficult to suppress with conventional filtering.

2.7.3 Eye Safety and Power Constraints

System performance is further constrained by eye safety regulations. While longer wavelengths (e.g., 1550 nm) allow higher permissible power levels under IEC laser safety standards, they reduce detection efficiency in silicon-based photodiodes [14]. Conversely, shorter wavelengths (e.g., 633 nm) provide high sensitivity but introduce greater scattering and potential safety risks. This trade-off forces designers to balance sensitivity, safety, and practicality, often resulting in suboptimal compromises.

The literature consistently acknowledges that environmental and surface limitations pose some of the most significant barriers to the practical deployment of laser microphones. Although controlled experiments often demonstrate promising noise suppression and speech intelligibility, these results rarely translate to uncontrolled settings. Current research gaps include the lack of standardised protocols for testing under turbulence, variable reflectivity, and real-world outdoor conditions. Future work should therefore prioritise hybrid systems that combine optical sensing with auxiliary modalities (e.g., inertial or contact sensors) to ensure robustness against environmental disturbances. Without addressing these limitations, even the most advanced filtering algorithms cannot guarantee reliable field performance.

2.8 Conclusion

This review surveyed the evolution of laser microphones from Bell's 1880s photophone through Theremin's 1947 infrared system to contemporary semiconductor implementations achieving sub-nanometer sensitivity [6]. Modern systems integrate distributed feedback lasers, low-noise transimpedance amplifiers, and digital signal processing, yet performance remains heavily surface- and environment-dependent, with significant sensitivity–robustness–cost trade-offs.

Component selection involves critical design choices: wavelength determines the balance between eye safety (1550 nm), photodetector efficiency (650–850 nm), and environmental robustness [14], [15]. Dual-TIA architectures enhance ambient light rejection through common-mode suppression, although their experimental validation remains limited. Signal processing spans two paradigms: traditional techniques (FIR filtering, Wiener, spectral subtraction) provide real-time efficiency but assume stationary noise [23], [24], [26], while machine learning (DNNs, RNNs) offers superior adaptability [7], [29] at higher computational cost. Hybrid architectures combining classical preprocessing with lightweight neural networks represent a promising compromise.

Speech-to-text integration enables objective evaluation via WER, STOI, and SNR [46], [51], [54], revealing non-linear metric relationships—high STOI does not guarantee low WER, and SNR gains may not directly improve intelligibility [58]. Therefore, comprehensive multi-metric evaluation remains essential for assessing overall system performance.

Priority research gaps include:

1. Development of standardized benchmark datasets capturing diverse surfaces, acoustic conditions, and noise profiles.

2. Systematic real-world characterization of performance under turbulence, variable reflectivity, and environmental interference.

3. Quantitative cost–performance analysis to inform economically viable designs.

4. Exploration of hybrid architectures integrating complementary sensing modalities (e.g., inertial sensors) for enhanced robustness.

In summary, the literature reveals that while individual subsystems—optical, electronic, and algorithmic—have advanced considerably, their integration into a cohesive, low-cost, and field-deployable system remains limited. This project directly addresses these gaps by implementing a dual-TIA receiver, a combined FIR–Wiener–spectral subtraction processing pipeline, and a multi-model ASR evaluation framework. Emphasizing cost-effectiveness (R2000 constraint), open-source reproducibility, and quantitative multi-metric assessment, the ensuing chapters transition from theoretical exploration to practical realization, demonstrating how laser vibrometry can achieve intelligible, transcribable audio reconstruction within realistic design constraints.

Chapter 3

Methodology

This chapter outlines the methodological framework adopted for the design, simulation, and experimental validation of the laser microphone system. The approach is structured into five interdependent stages, ensuring a systematic and iterative development process from conceptualisation to final prototype validation.

Each stage built upon the findings of the previous one to refine both the hardware and signal-processing aspects of the system.

Stage 1 – System Design and Simulation

The first stage involved the conceptual and analytical design of the laser microphone system. System requirements are defined through an assessment of acoustic sensing principles, optical reflection behaviour, and photodetector response. Then a block-level architecture is developed to establish the interaction between the optical transceiver, the analog front-end, and the digital signal processing (DSP) subsystems.

Optical and electronic subsystems are simulated using LTspice to model expected signal behaviour, frequency response, and noise performance. The simulations provided quantitative insight into the relationship between laser beam modulation and detected voltage variation, enabling the determination of design parameters such as photodiode sensitivity and pre-amplifier gain.

Stage 1.1 – Breadboard Testing and Iteration

Following initial simulations, a series of breadboard prototypes are constructed to validate the theoretical models. The optical alignment, photodiode biasing, and amplifier stages are tested incrementally. Each configuration is evaluated in terms of signal clarity, frequency response, and susceptibility to ambient light interference. The results of this phase informed iterative refinements of the circuit topology and component selection, ensuring system stability before PCB implementation.

Stage 2 – PCB Design

Once the circuit functionality is verified on the breadboard, the design is translated into a printed circuit board (PCB) using Kicad 9.0. Emphasis is placed on minimising noise coupling through careful component placement,

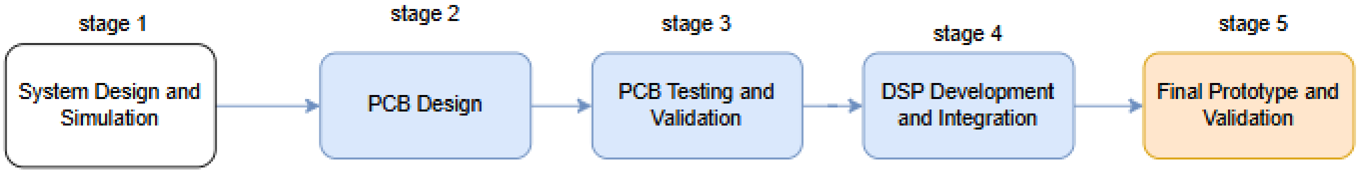

stage 1
System Design and Simulation
stage 2
PCB Design
stage 3
PCB Testing and Validation
stage 4
DSP Development and Integration
stage 5
Final Prototype and Validation

Figure 3.1: The overview of all the project stages

trace routing, and ground-plane segmentation. Design-for-manufacturability (DFM) principles are applied to ensure compactness and ease of assembly. Failure precautions are implemented at each stage to ensure the overall functionality of the system if one section fails.

Stage 3 – PCB Testing and Validation

After fabrication and having received the delivery of the PCB, it underwent a comprehensive testing procedure. Continuity tests using U3401A (digit dual display Multimeter) are performed to verify electrical integrity, followed by subsystem-level functional testing. Signal integrity is assessed using the DS0-X-2002A oscilloscope and spectrum analyser to confirm the expected frequency response and noise floor. Any deviations from simulation results are analysed to identify potential layout or grounding issues, allowing for specific rework where needed.

Stage 4 – DSP Development and Integration

The fourth stage focused on developing the digital signal-processing algorithms necessary for converting the detected analogue signal into a clear audio output. MATLAB and Python are used to implement and test algorithms, including filtering, noise suppression, and amplitude normalisation. The optimised DSP routines are then implemented on a Python app localised on a PC. Synchronisation between the optical detection hardware and DSP system is verified through controlled acoustic tests.

Stage 5 – Final Prototype and Validation

In the final stage, the complete laser microphone prototype is assembled, integrating all mechanical, optical, and electronic subsystems. The system is evaluated under controlled acoustic conditions to assess its sensitivity, frequency response, and distortion performance. Results are benchmarked against theoretical predictions and ideal acoustics recovery to verify design validity. Observations from this validation phase informed recommendations for further optimisation and field deployment.

The iterative process employed throughout this research follows a systematic cycle designed to ensure robust validation at each development stage shown. Beginning with a comprehensive literature review, each stage proceeds through its defined objectives before undergoing rigorous testing and validation. At decision points, the system's performance is evaluated against predetermined requirements. If these requirements are not met, the process loops back to the previous stage for refinement and retesting, incorporating lessons learned to address identified deficiencies. Only upon successful validation does the development advance to the current stage, where the same cycle of implementation, testing, and requirement verification is repeated. This feedback mechanism ensures that design flaws and performance limitations are identified and resolved before progressing to subsequent stages, thereby preventing the propagation of errors through the development pipeline. Should the current stage fail to meet its requirements, the iteration continues until satisfactory performance is achieved, at which point the process advances to the next stage. This structured approach minimises the risk of fundamental design issues emerging late in the development cycle and ensures that each subsystem—from initial simulation through to final prototype validation—performs reliably before being integrated into the complete laser microphone system. The details on validation and testing are provided at each stage.

Ethical Considerations Ethically, all test audio consisted of publicly available material or synthesised speech, with no recordings of private conversations or identifiable individuals. The system was operated exclusively within

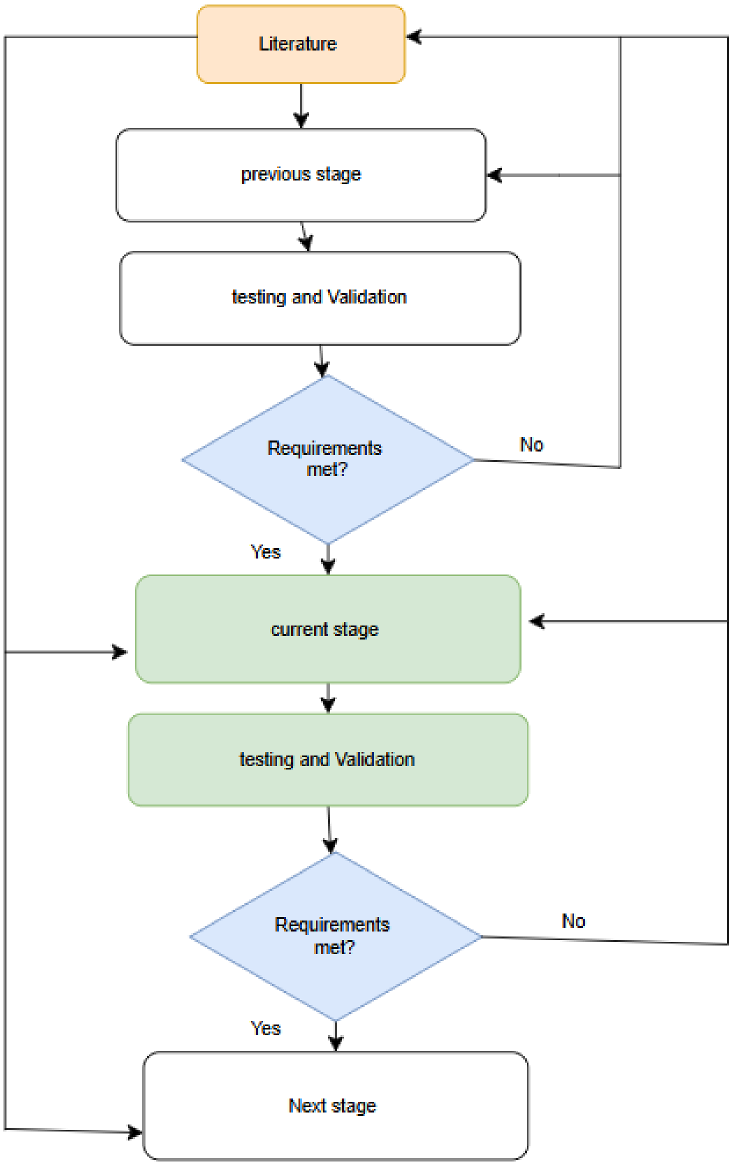


Figure 3.2: Iterative development flow diagram illustrating the requirement-driven progression through design stages with feedback loops for refinement and validation

the university laboratory under controlled conditions, with no field surveillance testing conducted, complying with UCT's ethical guidelines for engineering projects that do not involve human subjects research or covert monitoring applications.

Chapter 4

System Design and Simulation Analysis

This chapter presents the detailed design and simulation analysis of the laser microphone system. The design process is carried out iteratively, where simulation results guide hardware refinement and component selection. The approach ensured that both the optical and electronic subsystems are optimised for performance, stability, and manufacturability. The primary goal is to develop a reliable, low-noise, and compact system capable of reconstructing sound from optical reflections with high fidelity.

4.1 Overview

The laser microphone system operates on the principle of detecting sound-induced surface vibrations using an optical beam. A continuous laser is directed onto a reflective surface, and acoustic pressure variations cause minute displacements that modulate the reflected beam's intensity on the chosen optical receiver. This optical modulation is captured by a photodiode and transformed into an electrical signal, which is then processed through analogue amplification, filtering, and digital signal processing (DSP). The complete system is powered via a standard 5 V USB supply from the PC, making it portable and compatible with typical laboratory environments.

4.2 System Requirements and Specifications

The development of the laser microphone system begins with a clear understanding of its design expectations and operational constraints. These requirements establish the foundation for a reliable and functional system, ensuring that the design objectives are met during both simulation and prototype validation. This section defines the system requirements and corresponding specifications that guided the design process.

4.2.1 Requirements

Table 4.1 illustrates the user and functional requirements identified for the laser microphone system.

Table 4.1: Requirements for the Laser Microphone System

REQ-ID	**Description**
REQ-01	The system must be low-cost.
REQ-02	The system must operate using a 5 V DC supply provided via a standard type A USB 2.0 port.
REQ-03	The system must be capable of capturing sound remotely through a reflected laser beam.
REQ-04	The recorded audio output must be intelligible and comparable to the reference sound source.
REQ-05	The system design must require minimal user alignment and setup.
REQ-06	The complete setup should be compact, portable, and suitable for laboratory-scale testing.

The requirements listed above define the performance expectations of the laser microphone system and form the foundation for the subsequent specification and design phases. They ensure that the final prototype meets functional, operational, and usability targets while remaining low-cost and reproducible.

4.2.2 Specifications

Based on the analysis of the requirements tabulated above, Table 4.2 illustrates the detailed system specifications addressing each requirement, along with the corresponding acceptance criteria that must be met.

Table 4.2: Specifications for the Laser Microphone System

SPEC-ID	Description	REQ-ID	Acceptance Criteria
SP-01	The system must use a red or infrared laser diode (650 nm–1400 nm) to detect sound-induced vibrations from reflective surfaces at long range.	REQ-03	Verify that surface vibrations produce measurable signal modulations at the photo receiver output at a minimum of 1 m distance between the receiver and vibrating surface.
SP-02	The optical receiver must convert reflected light intensity variations into voltage signals using a transimpedance amplifier.	REQ-03, REQ-04	Validate the analogue output for a range of vibration amplitudes and frequencies (100–20 000 Hz).
SP-03	The analogue front-end must provide low-noise amplification and band-pass filtering for 100–18 000 Hz.	REQ-04	Measure the SNR of the audio file, and it must be within 85% of the SNR of the ideal sound.
SP-04	The system must interface with a USB sound card for digitisation and PC-based DSP processing.	REQ-02, REQ-04	Confirm that the recorded signal can be acquired at 44.1 kHz, 16-bit resolution in MATLAB.
SP-05	MATLAB-based DSP must reconstruct intelligible audio using spectral subtraction and filtering.	REQ-04	Evaluate processed signal clarity and confirm speech intelligibility above 85%.
SP-06	All system components must operate on a single 5 V USB power source.	REQ-02	Measure total current draw and confirm current consumption below 500 mA.
SP-07	The total prototype cost must not exceed R2000.	REQ-01	Validate the cost by summing all purchased components within the approved bill of materials.

The specifications outlined for the laser microphone system serve as a guide for design and implementation, ensuring that all requirements are met. Each specification is directly traceable to a corresponding requirement, providing measurable performance targets for verification and validation during testing.

4.3 Design Outline

The system design is also structured into three primary domains—optical, analogue, and digital—each fulfilling a specific role in the detection and reconstruction of sound. The optical subsystem provides remote vibration sensing through laser reflection, the analogue front-end ensures low-noise signal conditioning, and the digital signal-processing (DSP) subsystem reconstructs the audio signal through real-time computation. Figure **??** presents the complete system architecture, highlighting the interaction between these domains.

4.4 System Overview

The laser microphone system operates on the principle of optical modulation caused by surface vibrations. When the laser beam reflects off a vibrating surface, the returning light undergoes slight intensity variations proportional to the acoustic signal present on the surface. To extract the embedded information, the reflected beam is detected, converted to an electrical signal, and processed to recover the original sound waveform.

Figure 4.1 illustrates the high-level architecture of the receiver subsystem. The system is divided into three main subsystems:

- **Optical and Signal Conditioning Module** — captures the reflected light and converts it into a conditioned electrical signal using a photodiode and amplification circuitry.
- **Data Acquisition Module** — digitises the conditioned analogue signal via a USB sound card or ADC for further processing.
- **Digital Signal Processing (DSP) Module** — performs demodulation, filtering, and spectral enhancement to reconstruct intelligible audio from the detected signal.

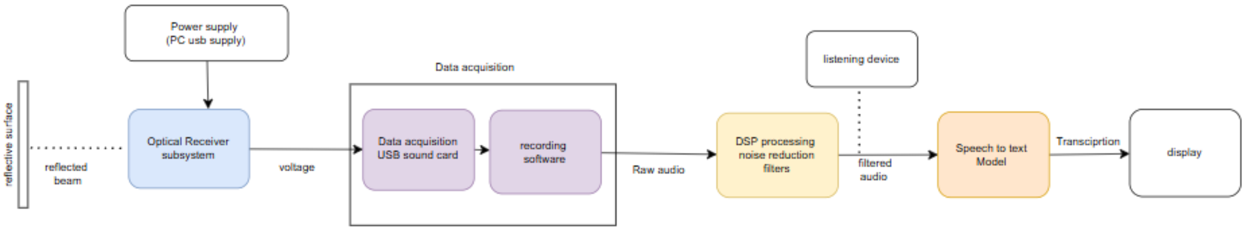


Figure 4.1: High-level architecture of the laser microphone receiver, showing the optical detection, acquisition, and DSP subsystems.

4.5 Hardware Design Process Overview

Following the description of the system design and requirements, it is essential to select hardware components that satisfy the design objectives outlined in Chapter **??**, while minimising noise during the retrieval of the modulated signal. This section details the hardware selection process, focusing on component suitability, electrical compatibility, and noise performance.

Laser Module

A red dot laser module operating in the visible range ($\lambda \approx 650$ nm) is selected as the optical transmitter for this project due to its practicality, visibility, and ease of alignment during laboratory testing. Although infrared (IR) lasers offer greater discretion for surveillance applications, their invisible beams complicate alignment and safety verification in an experimental setting. The visible red beam, on the other hand, enables precise targeting of the reflective surface and provides immediate visual feedback, significantly reducing setup time and alignment errors.

The selected **red dot laser module** offers stable optical output power, a compact form factor, and low cost, making it an ideal choice for the prototyping stage of the laser microphone system. Furthermore, the system architecture remains compatible with future adaptation to infrared operation, requiring only replacement of the laser source without major modification to the receiver or signal-processing stages. The datasheets of the following candidate modules are reviewed to evaluate output stability, current consumption, and optical parameters, ensuring compliance with the specifications outlined in Table **??**.

To determine the most suitable laser source for the prototype, three candidate transmitters are evaluated in terms of optical performance, electrical compatibility, and practical usability. Table **??** summarises the comparison of these options, highlighting key parameters such as wavelength, output power, operating voltage, and cost.

Table 4.3: Comparison of Laser Transmitter Options

Component / Feature	**KY-008**	**ROHM RLD65NZX1-00A**	**BezosMax red laser Pointer**
Wavelength	650 nm (Red)	663 nm (Red)	650 nm
Output Power	5 mW	Up to 10 mW	>5 mW
Beam Range	18 m under suitable conditions	Medium-range (tens of meters)	Up to 2,500 m
Operating Voltage	5 V DC	2.3 V DC typical	USB-C rechargeable
Package / Size	24mm × 15 mm, 3-pin module	TO-56 metal can, 3-pin	Handheld laser pointer with star cap and USB Type A male connector
cost(ZAR)	17.00(DIY electronics)	436.80(RS Components)	199(takealot)

RLD65NZX1-00A is not used because of the cost and the requirement to build a dedicated laser driver to ensure a stable output, as this would introduce more complexities into the design. Ky008 and the Bezomax laser pointer are lab tested on their reflectivity at 1 meter distance on a 3mm window pane glass in the setup as described in section section 4.7. ky008 could not be reflected at 1 1-meter distance, but can be reflected 10cm. While the Bezomax laser pointer is effectively reflected, even at 2m. Thus, given that the laser microphone, which requires a long range distance, the primary limitation of the BezosMax pointer is rapid battery depletion during testing. To address this, a 1m female-to-male USB cable is acquired. With the female part connected to the laser, the PC could supply the laser with constant power with a range of motion.

The difference in reflectivity can also be attributed to beam optics. The KY-008 is a bare module without an integrated lens system, so its beam is less uniform and less focused. The BezosMax pointer, in contrast, is packaged as a handheld pointer with built-in optics that produce a well-collimated beam, improving reflection over longer distances. Thus, the bezomax pointer is opted for use, and all the following processes considered it to be the main transmitter.

optical receiver

Based on the selection of the 650 nm laser module, an optical receiver is designed to operate within the laser wavelength and the acoustic frequency range. The optical receiver forms the critical interface between the optical and electrical domains of the laser microphone, converting modulated light reflections into electrical signals for amplification and processing.

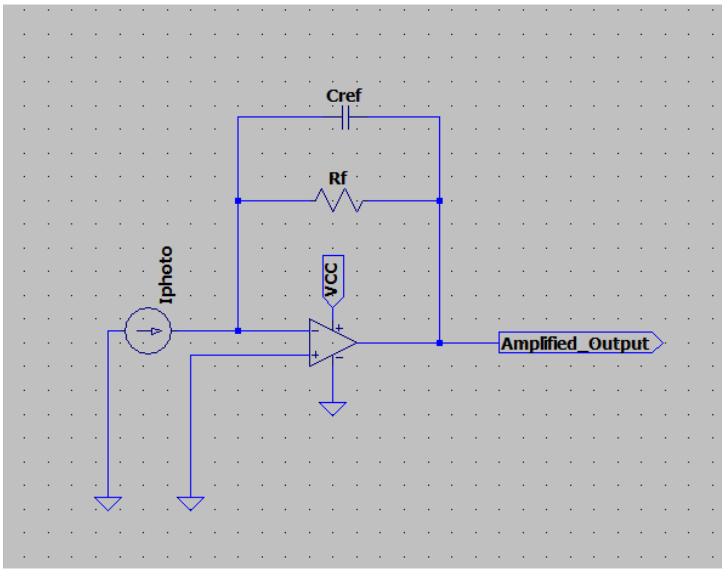


Figure 4.2: Transimpedance amplifier converting photodiode current into a voltage signal for optical-to-electrical interfacing.

In this design, a silicon photodiode is chosen after a comparative evaluation against a phototransistor. The selected device, a BPW34 photodiode, operates on the internal photoelectric effect, generating a photocurrent linearly proportional to the incident optical power. Its wide spectral sensitivity range (400–1100 nm) and peak responsivity of approximately 0.6 A/W at 650 nm make it well matched to the red laser source. The BPW34's fast rise time of about 100 ns provides a bandwidth exceeding 3 MHz—well above the 100 Hz–18 kHz acoustic range of interest—ensuring accurate tracking of vibration-induced optical modulation. Additionally, the photodiode exhibits low intrinsic noise and high linearity, allowing precise demodulation of the optical signal.

Although the photodiode compared to phototransistors it generates lower raw current, it delivers a cleaner, more linear signal with minimal noise and stable sensitivity over temperature. These characteristics make the BPW34 the optimal choice, providing a balance of sensitivity, fidelity, and simplicity suitable for the low-cost, high-precision requirements of the laser microphone system.

Transimpedance amplifier

In this design, a low-noise operational amplifier is connected in a feedback configuration with a precision resistor. The photodiode is connected to the inverting input of the op-amp, and the feedback resistor determines the gain of the circuit according to the relation as shown below:

$$V_{\text{out}} = -I_{\text{photo}} \cdot R_f$$

where I_{photo} is the photocurrent generated by the photodiode and R_f is the feedback resistor. A parallel capacitor across the resistor is often added to stabilise the amplifier and limit high-frequency noise, forming a low-pass filter that prevents oscillations while maintaining bandwidth sufficient to capture acoustic vibrations (100 Hz–18 kHz).

By integrating the TIA immediately after the photodiode, the design maximises signal-to-noise ratio and preserves the fine details of the reflected laser signal, making it a critical component in the laser microphone system.

To improve common-mode rejection and mitigate ambient light interference, a **dual-TIA (transimpedance amplifier)** configuration is adopted, where two TIAs share a single photodiode input. This approach is inspired by the recommendations of Koheron **koheron2018**, who demonstrated differential TIA designs for enhanced ambient noise immunity.

The dual-TIA configuration, shown in Figure 4.3, employs two identical transimpedance amplifiers based on LM358 operational amplifiers, both connected to the same photodiode source. In this arrangement, the photocurrent generated by the BPW34 photodiode is fed simultaneously into two amplifier channels that operate in opposite polarities. Each channel converts the incident photocurrent into a voltage signal proportional to the optical intensity, with the outputs being 180° out of phase with respect to each other.

The individual transimpedance stages (U6 and U7) consist of precision feedback resistors (R_{17}, R_{18}) and small compensation capacitors (C_{11}, C_{12}), which together define the gain and frequency stability of each amplifier. The resulting voltages are then AC-coupled through capacitors (C_9, C_{10}) to a subsequent differential amplifier stage (U8). The differential amplifier subtracts the two TIA outputs, effectively amplifying the desired modulated signal component while cancelling any common-mode disturbances present on both channels.

This configuration offers significant performance advantages in optical detection applications. Ambient light and background illumination typically introduce low-frequency or DC photocurrent components that appear identically on both TIA inputs. Since these unwanted signals are common to both channels, they are rejected by the differential amplifier according to its common-mode rejection ratio (CMRR). Consequently, the system primarily responds to the differential signal originating from laser intensity modulation caused by acoustic vibrations, while suppressing slow variations due to environmental lighting or power supply fluctuations.

To further illustrate the benefits of this approach, Table 4.4 summarises the key differences between a conventional single-TIA configuration and the adopted dual-TIA design.

Table 4.4: Comparison between Single and Dual Transimpedance Amplifier Configurations

Characteristic	**Single TIA Configuration**	**Dual TIA Configuration**
Ambient Light Sensitivity	High — DC and low-frequency ambient light components directly appear at the output.	Strongly reduced; common-mode (ambient) components are cancelled by differential subtraction.
Signal-to-Noise Ratio (SNR)	Limited by ambient fluctuations and amplifier offset drift.	Improved SNR due to common-mode noise rejection and balanced signal paths.
Output Offset and Drift	Large DC offset may cause output saturation or reduced dynamic range.	DC offset is largely eliminated through differential operation, maintaining a wide dynamic range.
Circuit Complexity	Simple and low component count.	Slightly more complex; requires component matching and an additional differential stage.
Response to Modulated Optical Signal	Directly proportional to photocurrent.	Differential gain applied to the modulated component while rejecting background intensity variations.
Practical Suitability for Laser Microphone	Susceptible to environmental light interference and noise.	Provides stable operation and enhanced robustness under varying ambient light conditions.

By implementing this dual-TIA structure, the receiver achieves a higher signal-to-noise ratio and improved stability under varying ambient conditions. The approach also reduces the likelihood of output saturation caused by large DC offsets, ensuring that the dynamic range is reserved for the modulated optical signal. This differential detection technique thus provides a practical and cost-effective method for enhancing noise immunity in the analogue front-end of the laser microphone system, resulting in cleaner signal acquisition and more accurate reconstruction of the acoustic waveform.

4.5.1 Digitisation and Processing Hardware Comparison

The digitisation stage is responsible for converting the conditioned analogue signal from the receiver into a digital format suitable for signal processing and analysis. Several hardware options are evaluated based on sampling resolution, ease of integration with the PC-based DSP environment, and overall system cost. Table 4.5 summarises the comparison of the candidate devices considered for the prototype.

The HiFi USB sound card is ultimately selected as the digitisation interface due to its simplicity, compatibility, and cost-effectiveness. Operating at 48 kHz and 16-bit resolution, it provides sufficient bandwidth and dynamic range for the 100 Hz–18 kHz acoustic band of interest. Its plug-and-play functionality allows direct connection to the PC via USB without the need for additional drivers or microcontroller programming, enabling seamless data acquisition in MATLAB.

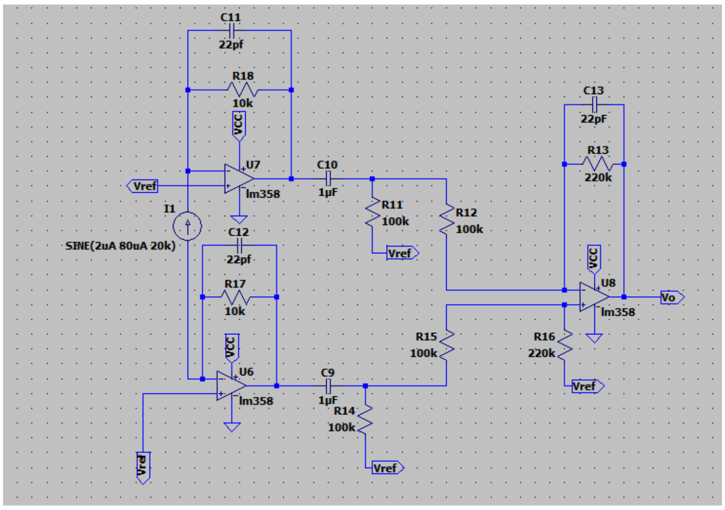


Figure 4.3: Dual transimpedance amplifier (TIA) configuration used to improve common-mode rejection and reduce ambient light interference

Table 4.5: Comparison of ADC and Processing Interfaces

Device	**Sampling Rate / Resolution**	**Ease of Integration**	**Approx. Cost (R)**	**Decision**
HiFi USB Sound Card	48 kHz / 16-bit	Plug-and-play, directly recognised by MATLAB and PC audio drivers	150	**Selected**
Raspberry Pi Pico	500 kS/s / 12-bit ADC	Requires custom firmware and serial communication for data transfer	88	Alternative (future)
Teensy 4.0 + Audio Shield	44.1 kHz / 16-bit	Excellent audio quality but requires complex setup and exceeds cost target	480	Rejected (budget)

In contrast, the Raspberry Pi Pico offered higher sampling capability but required custom communication protocols and additional code to handle serial data transfer, complicating integration. The Teensy platform, while capable of studio-grade audio quality, is excluded due to its significantly higher cost and unnecessary processing overhead for this stage of the prototype. Therefore, the USB sound card presented the most practical balance between performance, simplicity, and affordability for the laser microphone system.

4.5.2 Receiver Circuit Implementation — component calculations and selection

The complete receiver schematic is provided in Appendix **??**. The analogue front-end is sized to maximise sensitivity to the expected photocurrent levels (order of μA) while preventing DC saturation and preserving the acoustic band of interest (100 Hz–18 kHz). Design decisions are guided by three constraints: (1) sufficient transimpedance to amplify small photocurrents, (2) stability and bandwidth control to avoid oscillation and aliasing from optical noise, and (3) compatibility with a 5 V USB single-supply and the selected HiFi USB sound card. The key choices are summarised below and followed by a table that lists the formulae used, the computed target values, and the final E12 selections.

Design rationale The front-end uses a dual-TIA topology (matched TIAs, each with feedback resistor $R_f =$ 10 kΩ) to avoid excessive DC gain at the photodiode node while providing low-noise conversion of photocurrent to voltage. Small feedback capacitors ($C_f =$ 22 pF) stabilise each TIA and place the amplifier pole well above the acoustic band, preventing peaking. AC coupling (1 μF) between stages blocks DC offsets so the differential amplifier (gain set by 220 k/100 k = 2.2) can reject common-mode illumination and amplify the modulated component. A two-pole band-pass is implemented using RC sections chosen to yield a high-pass corner near 100 Hz and a low-pass corner near 18 kHz to limit low-frequency drift and high-frequency optical noise while remaining conservative for speech bandwidth. A reference divider of two 100 k resistors with a unity buffer provides $V_{ref} = 2.5$ V; the unity-gain buffer isolates this node from the rest of the receiver network and prevents loading. The LM358 family is used because of laboratory availability, its single-supply operation, and acceptable bandwidth for the audio band.

Stability and saturation considerations. Using modest TIA feedback (10 kΩ) prevents the output from saturating under larger reflected currents; in simulation the expected modulated photocurrent (2–80 μA) produces amplifier outputs within the device headroom when referenced about $V_{ref} = 2.5$ V and AC-coupled into the differential stage. The differential amplifier with matched resistor networks yields common-mode rejection of DC ambient light components, reserving the dynamic range for the modulated signal.

Table 4.6: Component calculations and final E12 selections

Function / Component	**Calculation / Requirement**	**Target (computed)**	**Selected E12 value**
TIA feedback resistor R_f	$V_{out} \approx I_{\text{photo}} \cdot R_f$. Choose R_f to convert μA-level signals into 10s–100s mV without saturating on larger reflection currents.	R_f = 10 kΩ (gives 2μA→20 mV, 80μA→0.8 V peak single-ended before AC-coupling)	10 kΩ
High-pass corner (input)	Target $f_{HP} \approx 100$ Hz. Use $R_{HP} = 15$ kΩ and $C_{HP} = 100$ nF: $f_{HP} = 1/(2\pi RC)$.	$f_{HP} \approx 106$ Hz	$R = 15$ kΩ, $C =$ 100 nF
Low-pass corner (band-limit)	Target $f_{LP} \approx 10$ kHz to suppress optical and aliasing noise. With $R_{LP} = 8.2$ kΩ, choose $C_{LP} = 1$ nF: $f_{LP} \approx 19.4$ kHz.	$f_{LP} \approx 19,4$ kHz	$R = 15$ kΩ, $C =$ 1 nF
Differential amplifier resistors	Desired differential gain $G_d = R_G/R_{in}$. Chosen $R_{in} = 100$ kΩ, $R_G = 220$ kΩ to give $G_d \approx 2.2$.	$G_d = 220$ k/100 k $= 2.2$	$R_{in} =$ 100 kΩ, $R_G =$ 220 kΩ
Sound-card coupling capacitor	The sound card expects the signal referenced to ground; AC-couple the output and provide a DC path to ground. Chosen $C_{sc} = 1$ μF with input impedance ≥ 10 kΩ	Low-frequency cutoff $\ll$ audio band; f_c small.	1 μF
Op-amp choice	Single-supply operation, availability, adequate bandwidth for audio (LM358).	LM358 (supply 5 V, input range around ground to $\approx V_{CC} - 1.5$ V)	LM358 (used throughout)

Simulation and Verification. The receiver front-end and amplification stages are first verified in **LTSpice x64 24.0.12** before construction. The BPW34 photodiode is modelled as an ideal current source in parallel with its junction capacitance ($C_j \approx 10$ pF) and shunt resistance ($R_{sh} \approx 50$ MΩ). The photocurrent excitation followed:

$$i_{pd}(t) = I_{\text{amb}} + I_{\text{mod}} \sin(2\pi f t),$$

implemented as `SINE(2uA 18uA 3k)` in the netlist. The ambient term ($I_{\text{amb}} = 2$ μA) represents steady illumination, while the modulation term ($I_{\text{mod}} = 18$ μA) corresponds to the acoustic signal reflected from a vibrating surface at $f = 3$ kHz, within the typical voice band (300 Hz–3.4 kHz [26]).

Experimental observations indicated that the laser must be aimed towards the edges of the reflective pane rather than the geometric centre. Central regions of planar surfaces exhibit minimal displacement under acoustic excitation, whereas edge regions experience higher vibration amplitude, producing stronger modulation of the reflected beam. Consequently, the optical coupling efficiency, η, is adjusted to reflect this effect:

$$\eta = \frac{\text{modulated optical power captured}}{\text{incident laser power}} \approx 30\text{–}35\%,$$

For edge placement, compared to the conservative 25% assumed for full-beam reflection. Using the BPW34 responsivity ($R_\lambda \approx 0.6$ A/W) [60], the corresponding full-beam photocurrent for simulation is set to a conservative 30% of this value. This approach ensures that the amplifier stages operate within their linear range and avoids clipping, while still representing realistic signal levels.

Transient and AC analyses confirmed a flat magnitude response across 100 Hz–10 kHz and stable operation with no oscillation or overshoot. For the chosen TIA feedback resistor ($R_f = 10$ kΩ), the output remained well within headroom:

$$V_{\text{TIA,peak}} = I_{\text{mod}} R_f = 0.18 \text{ V}.$$

After the differential stage ($G_d \approx 2.2$), the output swing reached 0.4 V about the 2.5 V reference, comfortably inside the 0–5 V supply rails. Simulations also confirmed that increasing I_{mod} beyond 20 μA caused clipping, highlighting the importance of amplifier headroom.

Moreover, following the active band-pass amplifier stage, a compact two-pole passive low-pass filter is implemented to suppress residual high-frequency noise originating from the photodiode and amplifier stages. The network, consisting of an 82 kΩ resistor with capacitors of 1 nF and 0.1 nF referenced to the mid-supply bias (V_{ref}), introduces gentle roll-offs near 1.9 kHz and 19 kHz, respectively.

This configuration is added after frequency response analysis in **LTspice** revealed that frequencies above 1 MHz are being coupled to the output, prompting inclusion of this stage as an additional safeguard against unwanted high-frequency components. The filter effectively attenuates broadband interference and stabilises the output by preventing high-frequency feedback into the active amplifier stage.

Breadboard prototype testing. A full breadboard prototype is constructed and tested using:

- **Power Supply:** Agilent E3620A DC supply at +5 V.
- **Oscilloscope:** Keysight DS0-X-2002A for time- and frequency-domain validation.

When aligned to a Perspex window illuminated by the 5 mW 650 nm laser, the circuit produced an average output of approximately 0.6 V$_{pp}$. Back-calculating from the measured amplitude using $V_{\text{TIA,pp}} = 2 I_{\text{mod}} R_f$ yields:

$$I_{\text{mod,exp}} = \frac{V_{\text{TIA,pp}}}{2R_f} = \frac{0.6}{2 \times 10^4} = 15 \ \mu\text{A},$$

This corresponds to roughly one-quarter illumination of the BPW34's sensitive area, consistent with beam spot misalignment and scattering losses through the window. The observed signal bandwidth extended beyond 8 kHz, and the circuit captured audible modulation when connected to the sound card. However, the measured amplitude is lower than predicted, and mild high-frequency attenuation is observed. This behaviour is attributed to the limited output swing and gain-bandwidth product of the LM358 under a single-supply 5 V condition.

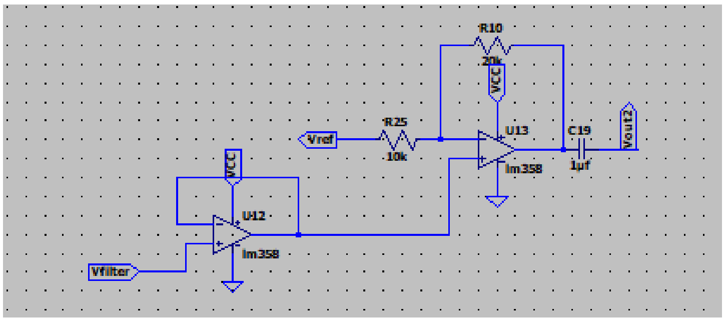


Figure 4.4: Buffered voltage reference and AC-coupled amplifier stage with gain of 2 (20k/10k) using LM358 op-amps. The circuit replacing just the ac coupled buffer

Amplifier correction and final stage. To address the reduced amplitude and high-frequency roll-off, the LM358 is replaced with an **MCP6292** dual op-amp. This substitution resolved two main limitations:

1. **Output headroom.** The LM358 cannot swing closer than about 1.2 V to the positive rail and 0.1 V to the negative rail at a 5 V supply, restricting its linear output range to roughly 0.1–3.8 V. Under strong modulation ($V_{\text{TIA, peak}} \approx 0.8$ V), this resulted in compression near the top rail. In contrast, the MCP6292 provides **rail-to-rail output**, typically reaching within 20–30 mV of either rail at a 10 kΩ load. This effectively increases the available headroom by more than 1 V, allowing full-scale operation of the differential and post-filter stages without clipping.

2. **Bandwidth and noise.** The MCP6292 offers a higher gain-bandwidth product (10 MHz vs. 1 MHz for LM358) and lower input-referred voltage noise ($\approx 8\ \text{nV}/\sqrt{\text{Hz}}$), improving high-frequency fidelity and SNR across the 18 kHz passband.

The revised configuration included a fixed post-filter gain of 2×, resulting in a measured output of approximately 1.2 V_{pp}, optimally matched to the HiFi USB sound-card input. Oscilloscope traces confirmed reduced high-frequency distortion and clean recovery of the modulated waveform, consistent with simulation predictions.

Summary. The stepwise verification—simulation, breadboard testing, and amplifier correction—confirmed the validity of the photocurrent model and circuit topology. The LTSpice source `SINE(2uA 18uA 3k)` accurately reproduced real illumination conditions, while the measured data highlighted the importance of amplifier headroom for maintaining linearity near the 5 V rail. Substituting the LM358 with the MCP6292 significantly improved dynamic range, high-frequency response, and noise performance, yielding a stable and faithful demodulated audio output from the laser microphone receiver.

Table 4.7: Photodiode illumination and photocurrent estimation (simulation vs. measurement)

Laser Power	$P_{\text{laser}} = 5$ mW	650 nm red diode laser
Window Reflectivity	$R_{\text{win}} = 0.08$	Typical glass surface
Optical Coupling Efficiency (center)	$\eta_{\text{center}} \approx 0.25$	Conservative fraction reaching photodiode at center of pane
Optical Coupling Efficiency (edge)	$\eta_{\text{edge}} \approx 0.3$	Increased fraction due to higher displacement at edge placement
Full-Beam Reflected Power	$P_{\text{ref}} = P_{\text{laser}} \times R_{\text{win}} \times \eta_{\text{edge}} \approx 120\ \mu$W	Edge placement improves coupling
Photodiode Responsivity	$R_\lambda = 0.6$ A/W	BPW34 datasheet [60]
Full-Beam Photocurrent	$I_{pd,\text{edge}} = R_\lambda P_{\text{ref}} \approx 18\ \mu$A	Conservative simulation value (30% of theoretical maximum)
Quarter Illumination (Measured)	$I_{pd,\text{exp}} \approx 15\ \mu$A	From 0.6 V_{pp} TIA output during breadboard testing
Simulated Source	`SINE(2uA 18uA 3k)`	Matches edge-placement illumination for realistic simulation

4.5.3 Mechanical Laser Holder Design

To improve system stability and reduce noise introduced by unintentional vibrations or angular misalignment, a custom 3D-printed mechanical holder is designed for the laser transmitter. The mount is fabricated using a **Bambu Carbon X1** printer with a **15% infill density**, providing a lightweight yet sufficiently rigid structure to minimise mechanical resonance and ensure repeatable optical alignment during testing. The holder is produced from **Polylactic Acid (PLA4)** plastic, a lightweight and durable material chosen for its ease of printing, dimensional stability, and ability to form complex geometries suitable for precision optical setups.

Design Overview

The 3d printed laser holder assembly consists of three key components, each serving a distinct mechanical function as illustrated in figure 4.5:

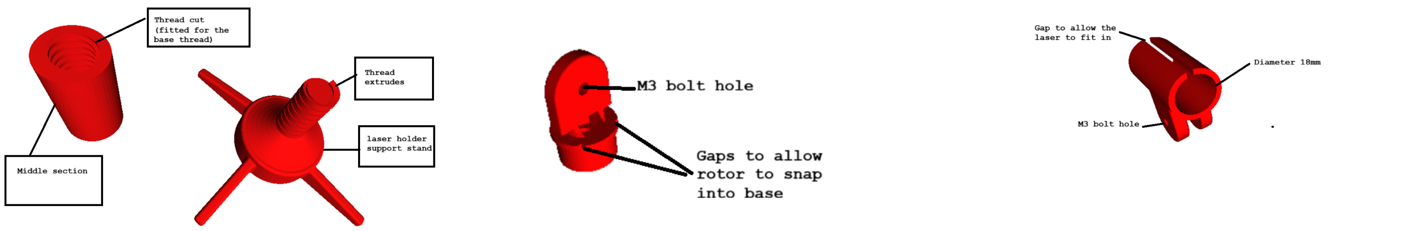
- **Base Stand:** An adjustable platform measuring approximately **12.5 cm at maximum height excluding the holder and rotor**, providing elevation and structural stability for the overall mount. The base anchors the assembly to the testing bench and isolates external vibration from the laser. Illustated in figure 4.5a and 4.5

- **Rotor Section:** The mid-section incorporates a **360° yaw rotation** mechanism, enabling smooth horizontal angular adjustment. This allows precise beam alignment across the reflective surface without disturbing the optical path length.The rotor is shown in 4.5b

- **Laser Holder:** The upper housing provides up to **90° roll adjustment** for the laser module itself, allowing control of the laser's pitch and roll angles. This flexibility permits various alignment configurations—horizontal, diagonal, and oblique—depending on the experimental setup. The holder is shown in 4.5c

This modular, rotatable design is critical in reducing alignment-induced amplitude fluctuations and enhancing the repeatability of optical coupling during denoising and calibration experiments. All the STL files are available in the link provided in A.

Figures


Thread cut (fitted for the base thread)
Middle section
Thread extrudes
laser holder support stand
M3 bolt hole
Gaps to allow rotor to snap into base
Gap to allow the laser to fit in
Diameter 18mm
M3 bolt hole

(a) 3D model of the base stand showing the structural support and mounting surface

(b) Rotor section with central M3 bolt hole and snap-fit gaps allowing 360° yaw rotation.

(c) Laser holder component enabling up to 90° roll adjustment for precise beam alignment.

Figure 4.5: Exploded render of the 3D-printed mechanical laser holder assembly showing (a) the base stand, (b) the rotor section, and (c) the laser holder with a firm grip on the 18mm diameter laser case .

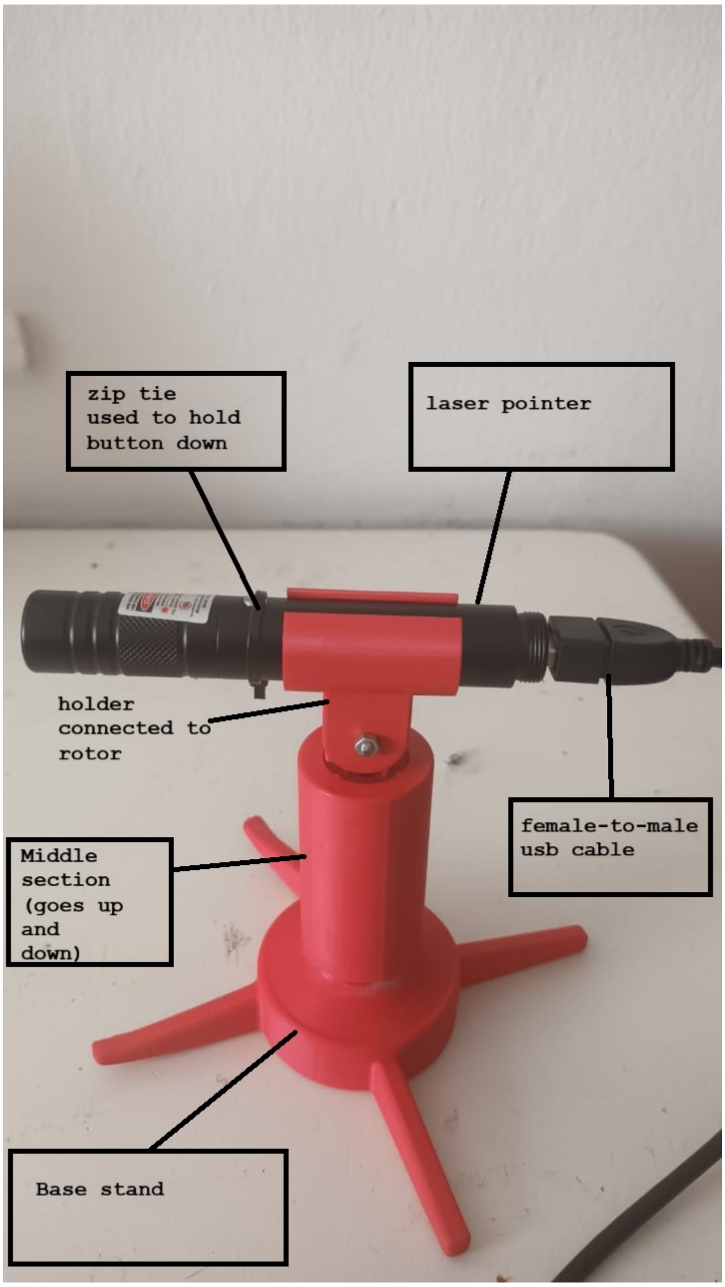


Figure 4.6: the laser holder showing the stand, rotor, and laser mounted together. This set up gives the laser a maximum height of 15cm based on the adjustment of the middle which can be rotated up or down

4.5.4 PCB Design and Layout

Following the successful simulation, breadboard validation, and comparative material testing, the final receiver circuitry is migrated onto a printed circuit board (PCB) for long-term stability and repeatability. This stage translated the verified analog front-end, filtering, and differential amplification stages into a robust hardware platform suitable for continued experimentation and acoustic signal recovery.

Each component is implemented in accordance with its datasheet specifications to ensure proper biasing, impedance matching, and overall circuit reliability. The schematic and layout are designed using **KiCad 9.0.4**, an open-source electronic design automation (EDA) tool that provides schematic capture, layout routing, and design rule checking.

PCB Requirements

REQ PCB.1 Compact form factor suitable for integration within an experimental laser microphone setup.

REQ PCB.2 Restrict to a two-layer board to minimise manufacturing costs while maintaining adequate signal integrity.

REQ PCB.3 Utilise surface-mounted (SMD) components wherever feasible to reduce component footprint and improve assembly efficiency.

REQ PCB.4 Ensure the layout minimises coupling between analogue and digital sections to preserve low-noise performance.

Layout and Grounding Strategy

Given the receiver's high sensitivity to noise, the analog signal path — from the BPW34 photodiode through the preamplifier and filter stages — is prioritized in component placement. The photodiode and first-stage transimpedance amplifier are positioned in close proximity to minimize parasitic capacitance and reduce external noise pickup. High-frequency nodes and power traces are physically separated from the analog front end to prevent interference.

A continuous ground plane is implemented across both layers to ensure low-impedance return paths for all signal currents. Sensitive analog grounds are connected at a single star point near the photodiode stage to eliminate ground loops. Decoupling capacitors are placed within 2–3 mm of each IC's supply pins, following manufacturer recommendations for the MCP6292 op-amps.

Jumpers and Configuration

To support flexible testing and configuration, the PCB includes several jumper headers that allow selective signal routing and stage bypassing. This modular configuration simplifies debugging and performance tuning during experimental evaluation.

Table 4.8: Configurable Jumpers and Their Functions

Jumper	Label	Function
JP1	Ambient reference select	Enables or disables the ambient photodiode biasing network.
JP2	Supply configuration	Allows external supply injection or onboard regulation selection.
JP3	Differential gain path select	Configures input stage between single-ended and differential operation.
JP4	Output mode	Switches between filtered and unfiltered output for DSP comparison.
JPFilter1 / JPFilter2	Filter chain selection	Enables high-pass and low-pass filter stages independently for signal conditioning.

This jumper-based configuration enables reconfiguration of gain paths, bias references, and filter behavior without hardware modification — a crucial feature during analog gain tuning and noise characterisation.

Test Points and Debug Access

To facilitate circuit validation and calibration, multiple test points (TP) are distributed across key analog and biasing nodes, allowing oscilloscope or multimeter probing during evaluation.

The accessibility of these test points is prioritised in the layout to ensure accurate gain and frequency response measurements during both bench and in-system testing.

Table 4.9: Designated Test Points for Measurement and Debugging

TP	Node Description
TP1	Photodiode bias voltage (V_{refLTP})
TP2	Transimpedance amplifier output (Detector)
TP3	Filtered signal output
TP4	Final audio output (C_{out1})
TP5	Post-filter DC bias
TP6	High-frequency test node (HF_TP)

Power and Filtering Considerations

Stable power delivery is essential for maintaining low-noise performance in the transimpedance and differential amplifier stages. To ensure this, a distributed decoupling network is implemented on the PCB.

A 22 μF bulk electrolytic capacitor is placed near the main supply input to stabilise the DC rail and absorb low-frequency transients. Each op-amp (mcp6292) is locally decoupled with a 0.1 μF ceramic capacitor positioned within 2–3 mm of its supply pins. These high-frequency bypass capacitors suppress switching spikes and minimise supply impedance above 1 MHz.

Short, wide traces are used to reduce parasitic inductance, ensuring effective high-frequency filtering. Together, the bulk and local capacitors form a two-tier network — the bulk capacitor handles low-frequency ripple, while the ceramics maintain local voltage stability at the ICs.

A buffered resistor divider generated the low-noise virtual ground (V_{ref}) shared by the differential and filter stages. The V_{ref} trace is routed with increased width and directly tied to the local decoupling nodes to maintain consistent biasing and minimise common-mode disturbances.

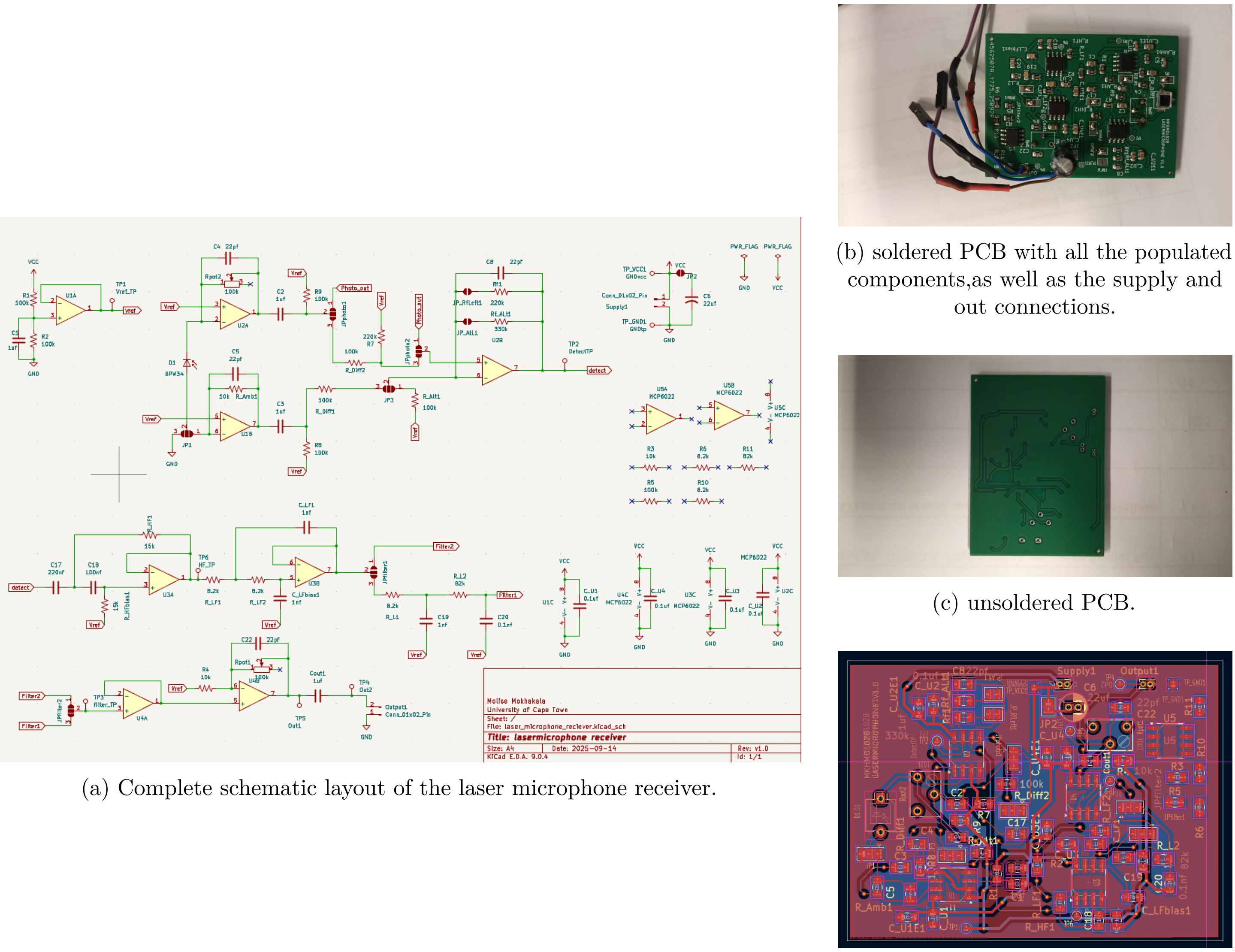
Physical Layout and Dimensions

The final PCB design measures **55 mm × 42 mm**, with all components located on the top layer for ease of soldering and inspection. Signal traces are kept short and direct, particularly in the TIA region, to preserve bandwidth and prevent oscillations. The bottom layer primarily consists of the solid ground plane and minimal interconnect routing.

Before conducting the acceptability tests, several configuration and assembly steps are completed to ensure the PCB matches the validated breadboard prototype. The jumper pads are soldered to implement the **dual-TIA configuration**, with JP1 (pads 2–1), JPphoto1 (pads 2–3), JP3 (pads 3–2), JPphoto2 (pads 3–2), and `JP_Rfleft1` all connected according to the layout used in Figure 4.3. The passive second-order low-pass filter is enabled by bridging pad 2 to pad 3 on both JPfilter1 and JPfilter2.

Potentiometers are soldered from the top layer, while their adjustment knobs are oriented toward the bottom to conserve space on the top surface. This placement prevented Rpot2 from obstructing the BPW34 photodiode and Rpot1 from pressing against the output capacitor (C_{out1}). Rpot2 is set to 10 kΩ and verified using a multimeter, while Rpot1 is set to the full 100 kΩ. The BPW34 photodiode is installed in reverse configuration as per the schematic, JP2 is soldered in to complete the supply link, and the 22 μF electrolytic capacitor is mounted near the input rail for supply decoupling.

Following assembly and visual inspection, a brief electrical validation is performed to confirm correct biasing, signal integrity, and overall functionality of the fabricated PCB. The essential acceptability tests—conducted via the



(a) Complete schematic layout of the laser microphone receiver.

(b) soldered PCB with all the populated components,as well as the supply and out connections.

(c) unsoldered PCB.

(d) top view with top tracks indicated in red and bottom tracks in blue.

Figure 4.7: PCB design views showing the overall layout (left) and detailed layer breakdowns (right). The configuration highlights the analog front-end placement, ground-plane strategy, and component accessibility.

designated test points—are summarised in Table 4.10.

Table 4.10: PCB Functional Acceptability Tests and Results

Test No.	Parameter	Test Method (using Test Points)	Expected / Result	Pass/Fail
1	**Continuity Test**	Measured resistance between VCC and GND prior to power-up using a multimeter. Checked all grounds for continuity using TP_GND1.	No short circuits detected; all grounds continuous.	Pass
2	**Supply Voltage Verification**	Applied 5 V via `Supply1` pins. Monitored TP5 and TP6 for stability and ripple.	$V_{CC} = 5.00$ V, stable with negligible ripple.	Pass
3	**Virtual Ground (V_{ref})**	Measured TP1 (reference node) with respect to ground to ensure correct biasing for op-amp stages.	$V_{\mathrm{ref}} = 2.48$ V, within 1% of $V_{CC}/2$.	Pass
4	**Output Voltage Test**	Applied modulated optical input to BPW34. Observed TP4 (`Cout1`) on oscilloscope for amplitude and clipping.	$V_{\mathrm{out}} \approx 1.2$ V$_{pp}$, clean waveform, no clipping.	Pass
5	**Frequency Response**	Injected 300 Hz–5 kHz sinusoidal modulation through photodiode; monitored TP3 (filtered output).	Flat response in 300–3 kHz voice band, roll-off at 18.7 kHz.	Pass

The resulting board offers a compact, low-noise, and reconfigurable platform that integrates the validated circuit stages from simulation and breadboard testing into a reliable and repeatable hardware implementation.

4.6 Final Hardware Prototype Implementation

A 3.5 mm TRS male auxiliary cable of length 20 cm is prepared and soldered to the `Output1` terminals. The ground conductor (black) is connected to pin 1, while the signal conductor (white or red) is connected to pin 2. This cable provides the audio interface to the USB sound card's microphone input. For power delivery, female jumper cables are connected to the `Supply1` header, and a USB power cable is modified by soldering male connectors, allowing the receiver PCB to be conveniently powered from a standard USB port. The soldered board can be seen in figure 4.7(b).

The completed PCB is mounted inside a custom 3D-printed enclosure fabricated from black PLA. The enclosure is designed to hold the PCB upright, orienting the BPW34 photodiode such that its photosensitive surface lies parallel to the board and directly faces the incident laser beam. The use of black PLA minimises internal reflections from the laser spot after striking the photodiode, thereby reducing optical noise from stray reflections and ambient light.

A pair of 2.5 mm wide zip ties is threaded through mounting holes on both sides of the enclosure to securely fasten the PCB in position. An additional tie is used to manage and restrain the power and audio cables, improving cable strain relief and maintaining a clean setup.

The rear side of the enclosure includes precisely cut recesses for the potentiometers, with an additional 1 mm clearance to allow easy access for gain or filter adjustments. The front section features a circular aperture of radius 7.5 mm, aligned with the BPW34 photodiode, to admit the incident laser beam. The base incorporates a push-fit circular stand of diameter 8.5 cm, providing stability and yielding an overall enclosure height of approximately 12 cm.

The following figures illustrate these details.

Figures 4.8a and 4.8b illustrate the final enclosure and mounting stand used for the laser microphone receiver.

The laser's setup, previously shown in figure 4.5 is the final design of the laser module design used in the entire project.

4.7 Testing Methodology and Validation Approach

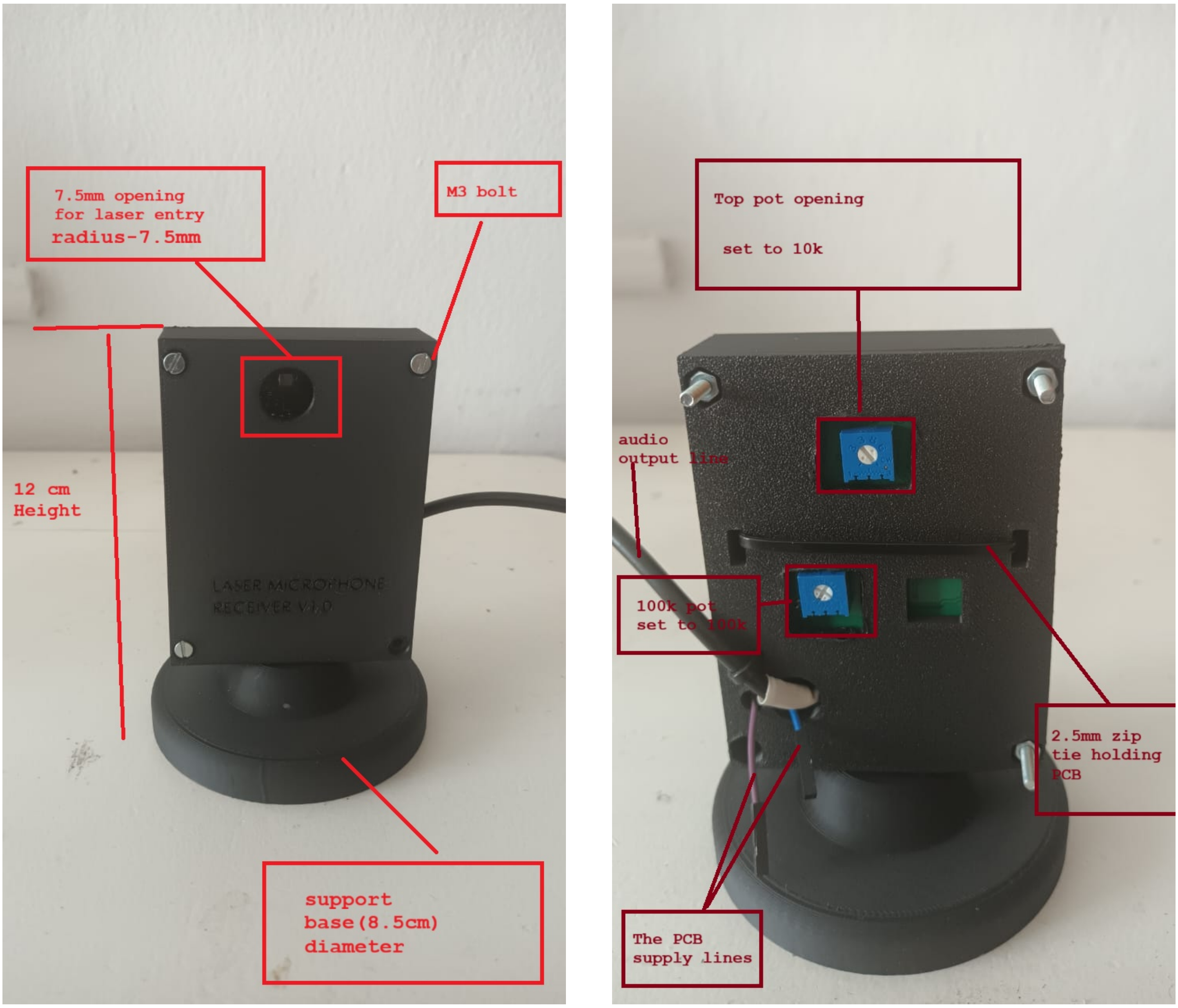
To evaluate the performance of the designed laser microphone receiver, a structured two-stage testing methodology is employed. Both the **breadboard prototype** and the **final PCB-based receiver** are subjected to identical experimental conditions to ensure consistency and comparability of results. The methodology is designed to validate both the analogue front-end behaviour and the digital signal reconstruction quality.

4.7.1 Experimental Setup

Recordings are conducted with the optical receiver aimed at a 3 mm thick clear glass window pane mounted on a **20 cm × 17.8 cm × 25 cm** chipboard sound-box enclosure, as shown in Figure 4.9. A **JBL Go 3** Bluetooth loudspeaker is positioned inside the enclosure and used to play a 2-minute monologue excerpt from the film *Fences*. The corresponding audio file is included in the Git repository referenced in the Appendix. This configuration simulated a controlled reflective surface at an approximate distance of 1 m from the optical receiver, representative of real-world acoustic reflections from domestic window panes.

The **JBL Go 3** is selected due to its broad acoustic bandwidth and clean output, which are essential for producing well-defined vibration patterns on the reflective surface. According to the manufacturer's specifications, the Go 3 provides a frequency response from **110 Hz to 20 kHz** and a signal-to-noise ratio greater than **85 dB**. These characteristics ensure that the test signal adequately covers the full voice-band frequency range while maintaining low distortion and high tonal accuracy. The speaker is operated at its maximum volume setting during all tests to generate sufficient sound pressure levels for measurable optical modulation on the glass surface. This approach also improved the signal-to-noise ratio of the captured laser recordings, ensuring that the detected optical variations corresponded primarily to acoustic excitation rather than environmental noise.

The sound-box enclosure is constructed from **chipboard**, chosen primarily for its availability and ease of machining. Although inexpensive, chipboard provides a sufficiently dense and rigid structure to approximate the acoustic behaviour of small enclosed spaces. Its tough wall surfaces helped contain the internal sound pressure, enhancing



(a) Front view of the laser microphone receiver enclosure showing the laser entry aperture, M3 mounting bolts, and circular base support.

(b) Rear view of the receiver enclosure showing potentiometer access ports, power and audio connections, and cable management ties.

Figure 4.8: Final 3D-printed enclosure and mounting stand for the laser microphone receiver, fabricated from black PLA with a total height of 12 cm and base diameter of 8.5 cm. The labelled features highlight design details including component alignment, cable routing, and surface openings.

the coupling between the loudspeaker and the glass pane. To prevent acoustic leakage and undesired resonances, all seams are sealed with Genkem contact adhesive after assembly.

Preliminary playback assessments conducted in **Audacity v3.7.2** revealed low-frequency humming and high-frequency hissing artefacts caused by minor air gaps and panel vibrations. Following sealing and mechanical reinforcement, these artefacts are significantly reduced, resulting in improved acoustic confinement and more uniform vibration transmission across the glass surface. This refinement enhanced both the stability and clarity of the optical recordings obtained by the laser microphone system.

The enclosure seams are sealed with contact adhesive after screwing the wooden panels together to prevent acoustic leakage and to confine the sound field, thereby increasing the consistency of vibration patterns on the window surface. This is done after playing back the audio from Audacity and identifying that there are both harming and hissing sounds mixed in.

During preliminary playback tests conducted in **Audacity**, unwanted artefacts such as low-frequency humming and high-frequency hissing are detected in the recorded signals. These effects are traced to minor air leaks and panel vibrations within the enclosure. After sealing and mechanical reinforcement, the enclosure exhibited improved acoustic confinement, reduced spurious resonance, and more stable vibration patterns on the glass surface. This refinement significantly enhanced the consistency and clarity of the optical recordings captured by the laser microphone system.

Initially, a Perspex pane is used as the reflective medium; however, it exhibited excessive flexing that introduced low-frequency mechanical noise and reduced reflection efficiency. The Perspex is subsequently replaced with a **3 mm glass pane**, which provided:

- Higher optical reflectivity at 650 nm, resulting in a stronger detected photocurrent.
- Increased structural rigidity, reducing spurious low-frequency vibration components.
- Improved overall signal-to-noise ratio (SNR), providing more stable and repeatable measurements.

Analysis of these observations is better explored in chapter **??** which explores the metric analysis of how the use of each affected the results.

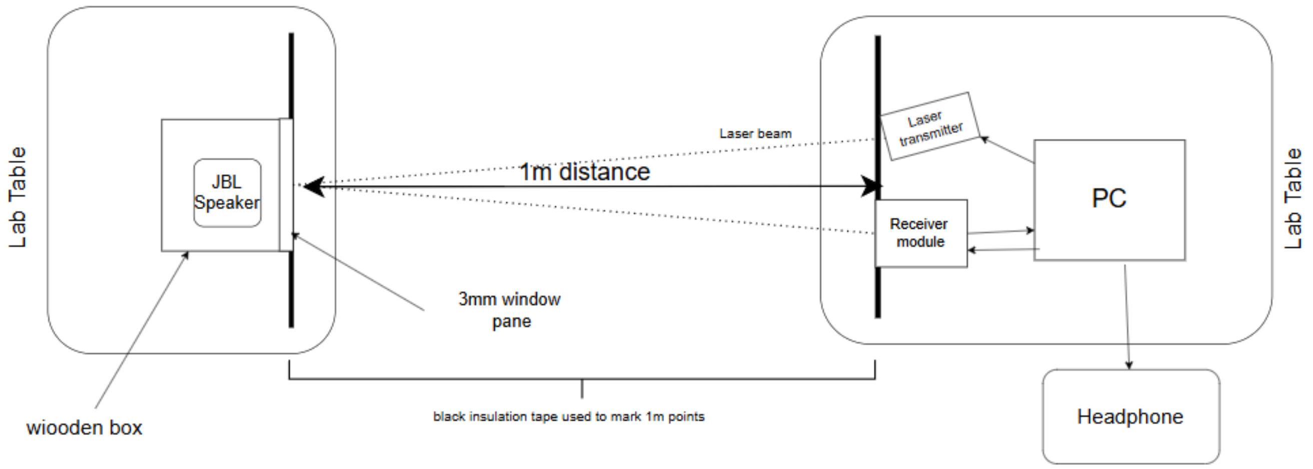


Figure 4.9: Test setup for validation of the laser microphone receiver: 3 mm glass window mounted on wooden sound enclosure with internal JBL Go 3 loudspeaker source.

4.7.2 Measurement Procedure

Both prototypes (breadboard and final version) are powered by a regulated PC USB A port's 5 V DC supply and aligned such that the 650 nm laser beam illuminated the glass pane at normal incidence. The reflected beam is collected by the BPW34 photodiode, and the resulting electrical signal is amplified through the dual-TIA and differential amplifier stages.

The analogue output is AC-coupled directly into a **HiFi USB sound card** configured for 48 kHz, 16-bit sampling. **Audacity v3.7.2** is used to capture two-minute recordings in uncompressed WAV format for each test condition. During each recording:

- The alignment of the laser beam is verified to maintain stable reflection.
- Ambient lighting conditions are kept constant.
- The sound playback level is fixed to ensure reproducibility of the vibration amplitude.

For the breadboard prototype, measurements focused primarily on verifying signal amplitude, waveform integrity, and frequency response across the voice band. The final PCB receiver is then evaluated for the same conditions, with attention to stability, improved headroom from the MCP6292 amplifier, and reduced noise floor.

4.7.3 Validation Framework

The validation process consisted of two main stages:

1. **Analogue Characterisation:** This stage quantified the receiver's gain, bandwidth, and signal linearity using oscilloscope measurements and spectral analysis. The captured output is examined for distortion, clipping, and high-frequency roll-off. The measured TIA gain is compared against LTSpice simulation results to confirm adherence to theoretical expectations.

2. **Digital Signal Analysis:** The recorded WAV files are analysed using MATLAB-based digital signal processing (DSP) scripts to evaluate the quality and intelligibility of the recovered speech signals. Three complementary performance metrics are employed in this analysis. The **Signal-to-Noise Ratio (SNR)** is used to quantify the clarity of the reconstructed waveform relative to background noise. The **Short-Time Objective Intelligibility (STOI)** metric provided an estimate of perceived speech intelligibility, offering an objective measure of how understandable the speech signal remained after processing. Finally, the **Word Error Rate (WER)** is used to evaluate speech-to-text transcription accuracy.

Each metric is computed for both the breadboard and final prototype recordings, enabling quantitative comparison of system improvements across development stages.

4.7.4 Summary of Methodology

In summary, the testing methodology ensured that both analogue and digital aspects of the laser microphone system are rigorously validated under realistic reflective and acoustic conditions. The use of a sealed test enclosure with a glass window provided a controlled yet representative environment. The two-stage approach—combining circuit-level verification with DSP-based evaluation—enabled comprehensive assessment of the receiver's capability to capture and reconstruct speech-modulated optical signals accurately.

Chapter 5

Signal Processing Theory and Implementation

5.1 Overview

The digital signal processing framework complements the hardware receiver by performing noise suppression, speech enhancement, and intelligibility evaluation. The hybrid approach—combining FIR preprocessing, spectral subtraction, and Wiener filtering—is motivated by prior research on single-channel speech enhancement for low-SNR environments [7], [21], [23].

5.2 Signal Model

The laser microphone output can be expressed as:

$$x(t) = s(t) + n(t)$$

where $s(t)$ denotes the speech component and $n(t)$ represents additive noise from optical and electronic sources. The DSP objective is to estimate $\hat{s}(t)$, maximising perceptual intelligibility and minimising distortion.

5.3 Digital Signal Processing Environment

The signal preprocessing and noise suppression algorithms are implemented in **MATLAB** due to its extensive library support for digital signal processing and ease of prototyping in the frequency domain. MATLAB provides high-level access to FIR filter design, spectral analysis, and visualisation tools, enabling rapid verification of theoretical models through simulation and empirical data comparison. Its built-in functions for the Short-Time Fourier Transform (STFT), noise estimation, and objective metrics such as SNR and STOI streamline algorithm development while ensuring numerical accuracy. This environment, therefore, offered an efficient and reproducible framework for testing and optimising the laser microphone's digital signal enhancement pipeline.

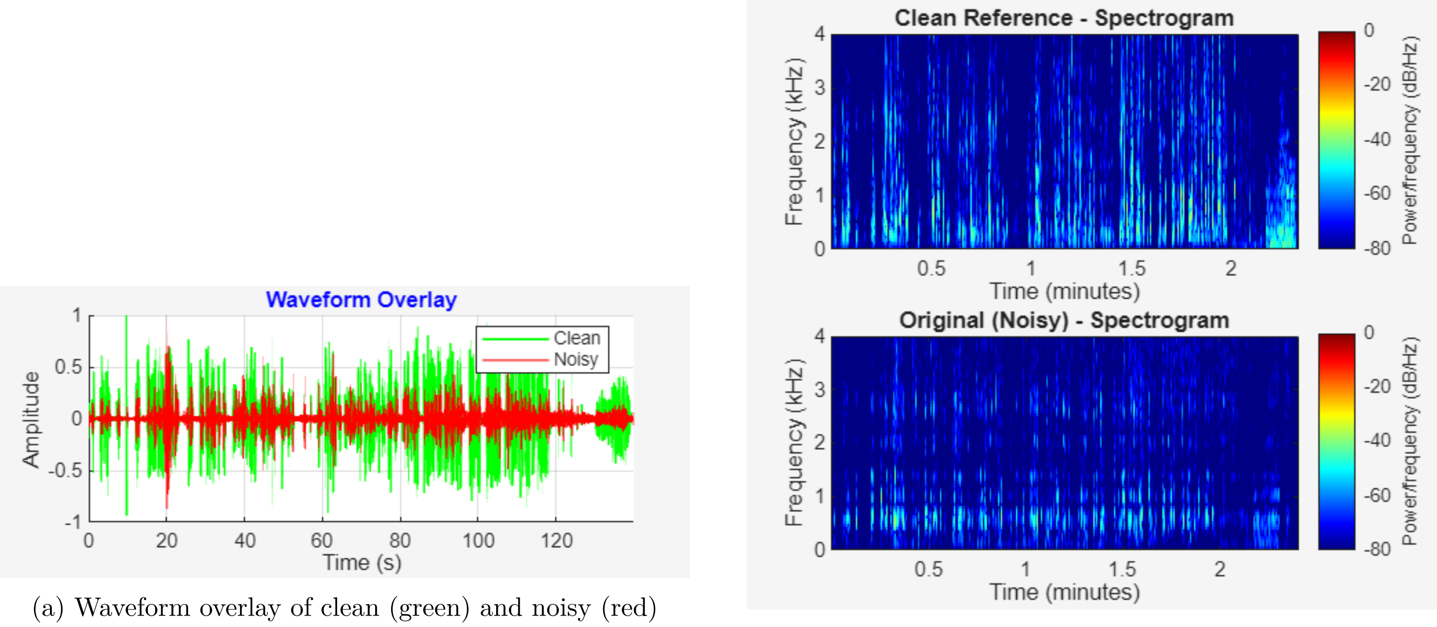
5.4 Preprocessing and FIR Band-Pass Filtering

The raw 1 m audio recordings are imported into MATLAB using the `audioread` function for initial time- and frequency-domain inspection. As shown in Figure 5.1, the clean reference waveform (green) exhibits higher amplitude variation and clear speech patterns, while the noisy signal (red) shows irregular spikes and reduced dynamics caused by ambient interference from the photodiode–amplifier system.

Spectrograms capture the time, frequency, and energy content simultaneously. They clearly show that the **clean reference** concentrates most energy below 2–3 kHz, with visible speech formants and harmonic bands, whereas the **noisy signal** displays scattered vertical streaks—non-stationary bursts typical of electrical and ambient noise.

Such noise artefacts are not visible in the waveform or Power Spectral Density (PSD) plots, thus the use of the spectrograms for evaluating signal quality and noise characteristics.

To suppress these unwanted components, a **finite impulse response (FIR) band-pass filter** with a passband of **100–3400 Hz** is designed using a Hamming window via the `fir1` function. This effectively retained speech-relevant frequencies while attenuating low-frequency hum and high-frequency sensor noise, improving signal-to-noise ratio (SNR) and speech intelligibility before further analysis.



(a) Waveform overlay of clean (green) and noisy (red) signals.

(b) Spectrogram comparison of clean and noisy signals.

Figure 5.1: Comparison of time-domain (left) and frequency-domain (right) characteristics for the 1 m test. The raw audio amplitude is generally lower than the clean reference, except for a spike near 20 s, which reduces the overall audibility of the laser microphone signal. In the spectrogram, the noisy signal maintains higher power levels (70–80 dB, dark blue) up to 3.5–4 kHz for most of the two-minute duration, while the clean reference remains around 50 dB (cyan), indicating high-frequency noise contamination in the recorded signal.

5.5 Bandwidth Limitation

Although the initial objective is to capture the full 20 Hz–20 kHz range to include all audible frequencies, listening tests and spectrogram analysis revealed significant white noise and low-frequency hum caused by ambient light interference and environmental air vibrations. The spectrograms confirmed that most speech energy resides below 3 kHz, consistent with findings in the literature [21], while higher frequencies are dominated by amplifier and photodiode noise. Restricting the bandwidth to **100–3400 Hz** effectively suppressed unwanted noise while preserving the speech formants essential for intelligibility. The resulting filtered output thus provides a cleaner, speech-focused signal optimised for further analysis.

5.6 Preprocessing and FIR Band-Pass Filtering in MATLAB

To prepare the raw laser microphone audio for digital signal processing, a **finite impulse response (FIR) band-pass filter** is designed to isolate the speech-relevant frequency range while suppressing unwanted low- and high-frequency noise. Listening tests and spectrogram analysis showed that useful speech energy lies primarily between 100 Hz and 3400 Hz, with frequencies outside this band dominated by environmental hum, air vibration noise, and photodiode amplifier artefacts. The MATLAB implementation of the band-pass filtering process is shown below:

```matlab
f_low = 100;
f_high = 3400;

% High-order design for sharp transition bands
filter_order = 300;
f_nyq = fs / 2;
f_norm = [f_low f_high] / f_nyq;

% Blackman window for superior stopband attenuation
b_bp = fir1(filter_order, f_norm, 'bandpass', blackman(filter_order+1));

% Zero-phase filtering to avoid phase distortion
x_bp = filtfilt(b_bp, 1, x);

% Save filtered output
audiowrite('stage_1_bandpass.wav', x_bp, fs);
```

Listing 5.1: FIR band-pass filter implementation in MATLAB

Algorithm Explanation

- **`f_low`** and **`f_high`** — Define the lower and upper cutoff frequencies (100 Hz and 3400 Hz). These bounds correspond to the effective speech range, ensuring that background rumble and high-frequency sensor noise are excluded.

- **`filter_order = 300`** — Specifies the number of FIR coefficients. A higher order provides sharp transition bands and better frequency selectivity at the cost of increased computation. FIR filters are inherently stable and exhibit linear-phase characteristics, ideal for preserving waveform shape.

- **`f_nyq = fs / 2`** — Computes the Nyquist frequency, i.e., half the sampling rate. Digital filter designs must normalise cutoff frequencies relative to this value to satisfy the Nyquist criterion.

- **`f_norm = [f_low f_high] / f_nyq`** — Converts the physical cutoff frequencies into normalised digital frequencies between 0 and 1. This ensures the filter correctly represents the desired frequency band regardless of the sampling rate.

- **`fir1()`** — MATLAB's built-in function for FIR filter design using the window method. See the code block below for the exact call.

- **`filtfilt()`** — Applies zero-phase forward and reverse filtering. The signal is filtered forward then backwards, cancelling phase distortion and preserving waveform symmetry.

- **`audiowrite()`** — Exports the filtered audio to a new WAV file (`stage_1_bandpass.wav`) for later analysis and playback.

Summary

The algorithm ensures that the filtered output retains only the 100–3400 Hz speech band, improving clarity and intelligibility. The high-order Blackman-windowed FIR design provides excellent stopband rejection, while zero-phase filtering guarantees that the temporal alignment of speech features is preserved. The resulting waveform exhibits reduced hum and buzzing, as evident in the spectrogram (Figure 5.1), confirming the effectiveness of this preprocessing stage.

5.7 Denoising techniques: Wiener and Spectral Subtraction Filtering

After band-pass filtering (100–3400 Hz), the audio retained residual hum and high-frequency hiss that degraded intelligibility. To improve quality, two frequency-domain denoising techniques are applied comparatively: the Wiener filter and spectral subtraction. Both operate on short-time frames (20 ms) with 50% overlap, assuming quasi-stationary noise within each frame.

5.7.1 Wiener Filter Algorithm

The Wiener filter aims to minimise the mean-square error between the clean signal $S(f)$ and the noisy observation $X(f) = S(f) + N(f)$. The optimal frequency-domain filter response is:

$$H_{\text{Wiener}}(f) = \frac{P_S(f)}{P_S(f) + P_N(f)}$$

where $P_S(f)$ and $P_N(f)$ are the estimated speech and noise power spectral densities (PSDs). Since $P_S(f)$ is unknown, it is estimated using the noisy signal power $P_X(f)$ and the noise power estimated from silent segments:

$$G(f) = \max\left(0, 1 - \frac{P_N(f)}{P_X(f) + \epsilon}\right)$$

where $G(f)$ is the frequency-dependent Wiener gain, and ϵ prevents division by zero. The enhanced signal is obtained by:

$$\hat{S}(f) = G(f) \cdot X(f)$$

Implementation Details. In MATLAB, the algorithm proceeds as follows:

- **`hann()`** — Generates a Hann window, tapering frame edges to minimise spectral leakage during the `fft()` computation.
- **`fft()`** / **`ifft()`** — Transform each frame to and from the frequency domain for per-bin gain application and signal reconstruction.
- **`smoothdata(...,'gaussian',5)`** — Smooths the Wiener gain across adjacent frequency bins to reduce "musical noise" caused by rapid gain fluctuations.
- **Overlap-add** — Each filtered frame is recombined in the time domain with 50% overlap to ensure continuous output without discontinuities.

This stage effectively reduces stationary noise components while preserving natural timbre and speech clarity, as expected from Wiener's minimum mean-square error (MMSE) formulation.

5.7.2 Spectral Subtraction Algorithm

Spectral subtraction, proposed in Chen's work [24], assumes additive noise and subtracts its average spectrum from the observed noisy spectrum. Resources such as the GitHub repository by the user https://github.com/vipchengrui/traditio speech-enhancement are used as inspiration for the development of the algorithm. The magnitude spectrum is estimated as:

$$|\hat{S}(f)| = \max\left(|X(f)| - \alpha|\bar{N}(f)|,\, \beta|\bar{N}(f)|\right)$$

where α is the *over-subtraction factor* controlling noise suppression aggressiveness, and β defines a *spectral floor* to prevent excessive attenuation and musical noise.

The phase of the noisy signal is retained since human perception is more sensitive to magnitude distortions than to phase errors:

$$\hat{S}(f) = |\hat{S}(f)|e^{j\angle X(f)}$$

The enhanced time-domain signal is reconstructed using the inverse FFT and overlap-add method.

Implementation Details. The MATLAB implementation uses:

- **`hamming()`** — A Hamming window reduces leakage in magnitude spectra estimation, complementing the Hann window used previously.
- **`fft()`** / **`ifft()`** — Converts between time and frequency domains for each frame.
- **`alpha = 2.0`** — Sets the over-subtraction factor; higher values yield stronger noise suppression but risk speech distortion.
- **`beta = 0.001`** — Defines the spectral floor; this prevents the gain from becoming zero and mitigates musical noise. (After tuning both alpha and beta and listening to the output audios, the most intelligible is when both are set to the aforementioned values.)
- **Half-wave rectification:** $\max(0, \cdot)$ ensures no negative spectral magnitudes appear after subtraction.
- **Overlap-add** — Recombines windowed frames with energy compensation using a squared window sum.
- **`audiowrite()`** — Saves the enhanced output as `stage_2b_spectral_subtraction.wav`.

This approach achieves higher overall SNR and attenuates wideband noise, though at the cost of mild spectral roughness and occasional transient distortion.

5.7.3 Comparative Evaluation

Both methods improved clarity relative to the band-pass-only stage. The Wiener filter yielded smoother, more natural speech with minimal distortion, while spectral subtraction more aggressively removed high-frequency noise and background hum. The combination of both, applied sequentially, balanced fidelity and suppression effectiveness. This iterative denoising process proved especially effective for laser microphone signals captured under fluctuating illumination and ambient noise conditions.

The complete MATLAB implementation is included in the Appendix for reproducibility and further experimentation.

Adaptive Spectral Gating

Following the Wiener and spectral subtraction stages, residual nonstationary noise—such as flickering interference from ambient light is still present in the processed signal. The **Adaptive Spectral Gating** stage addresses this by learning noise characteristics dynamically, similar to modern AI-based noise suppressors that adjust to changing acoustic environments. Sager et la [26] demonstrated a significant reduction of noise in the algorithm's use on their laser microphone, hence it is implemented in this DSP design.

Relation to Previous Stages Wiener filtering assumes a stationary noise environment and performs optimally under consistent broadband hum or hiss. Spectral subtraction, while effective for quasi-stationary noise, introduced "musical noise" artefacts. Adaptive spectral gating mitigates both limitations by estimating an instantaneous signal-to-noise ratio (SNR) for each frequency bin, applying a smooth, logistic gating function, and temporally smoothing the gain envelope to preserve speech naturalness. To test out the algorithm, it is first applied to the bandpass-processed signal instead of the Wiener and spectral subtraction filtered audio, to effectively be able to compare all three algorithms.

Technical Implementation in MATLAB The MATLAB script computes a baseline noise profile from the initial one-second segment using `mean()` and `var()` operations on the short-time Fourier transform (STFT) magnitudes. Each 20 ms frame (10 ms hop) is processed using:

$$\mathrm{SNR}_{\mathrm{est}}(f) = \frac{|X(f)| - \mu_N(f)}{\sigma_N(f) + \epsilon},$$

where $\mu_N(f)$ and $\sigma_N(f)$ are the estimated mean and variance of the noise spectrum. A soft gating function is then applied:

$$G(f) = \frac{1}{1 + e^{-\mathrm{sensitivity}(\mathrm{SNR}_{\mathrm{est}} - 1)}},$$

with a `sensitivity` factor of 2.5 controlling gating sharpness. The gain is smoothed across three consecutive frames using `movmean()` to avoid abrupt transitions. The signal is reconstructed using the inverse FFT (`ifft()`) with overlap-add synthesis, producing the file `stage_2c_adaptive_gating.wav`.

This stage dynamically adapts to time-varying noise, preserving low-energy speech components while reducing musical artefacts introduced by earlier filtering.

Voice Amplification

After denoising, speech energy is attenuated compared to the original signal, potentially reducing intelligibility for downstream transcription tasks. To address this, an **Adaptive Amplitude Restoration** stage is implemented that dynamically calculates the gain needed to restore the signal to the original audio's amplitude level without re-amplifying the suppressed noise. The restoration process operates in two steps: first, Voice Activity Detection (VAD) identifies speech regions using energy and zero-crossing rate criteria (ZCR between 0.03–0.25, energy > 0.001); second, the RMS level of detected voice regions in the denoised signal is compared to the corresponding regions in the original signal to compute a restoration gain factor given by $G_{\mathrm{restore}} = \mathrm{RMS}_{\mathrm{original}} / \mathrm{RMS}_{\mathrm{denoised}}$, capped at a maximum of 8× to prevent over-amplification. This gain is then applied selectively to voice regions using a smoothed binary mask, ensuring that only speech content is amplified while noise-only sections remain unaffected. Finally, a hybrid percentile-based normalisation strategy is employed: the 99th percentile of voice region amplitudes is computed and used as the normalisation reference (rather than the absolute peak), making the process robust to occasional amplitude spikes that would otherwise attenuate the entire signal. The normalised signal is then scaled to 98% of full scale and soft-clipped using a hyperbolic tangent function ($\tanh(1.05x) \times 0.99$) to handle any remaining outliers smoothly, avoiding harsh clipping artefacts. This approach successfully restored speech intelligibility while maintaining the noise reduction benefits achieved in earlier stages, as evidenced by the improved STOI scores in the amplified processing stages (Table 6.2).

Relation to Previous Stages While previous stages successfully suppress noise, they can also reduce vocal clarity. Applying global amplification would undesirably raise the noise floor; therefore, a selective approach ensures only speech segments are enhanced.

Technical Implementation in MATLAB The algorithm processes the waveform using 25 ms frames with a 10 ms hop size. For each frame, energy and zero-crossing rate (ZCR) are computed using:

$$E = \sum_n x^2[n], \quad \mathrm{ZCR} = \frac{1}{N-1} \sum_n |\,\mathrm{sgn}(x[n]) - \mathrm{sgn}(x[n-1])\,|.$$

A frame is classified as speech if $0.03 < \mathrm{ZCR} < 0.25$ and $E > 0.001$. The resulting voice mask $M[n]$ is smoothed using `movmean()` over 50 ms to ensure continuity. Voiced segments are amplified as:

$$y[n] = x[n] \cdot (1 + M[n](G_v - 1))\,, \quad G_v = 2.5,$$

and normalized using `y = y / max(abs(y))`. Output files include: `stage_3_wiener_amplified.wav`, `stage_3_spectralsub_a` and `stage_3_adaptive_amplified.wav`.

This stage enhances speech presence while maintaining a low background noise level.

—

Combined Multi-Stage Enhancement

The **Combined Multi-Stage Enhancement** stage merges the strengths of all prior techniques—Wiener filtering's smoothness, spectral subtraction's aggressiveness, and VAD-based amplification's clarity. This final hybrid method achieves a balance between noise suppression and naturalness, ideal for complex environments like laser microphone recordings.

Technical Implementation in MATLAB Starting from the Wiener-filtered output (`x_wiener`), the residual noise spectrum is re-estimated from the first 25 frames. A mild spectral subtraction is then applied:

$$|\hat{S}(f)| = \max\left(|X(f)| - 1.5\,|\bar{N}(f)|,\, 0.001\,|\bar{N}(f)|\right).$$

Reconstruction is performed using `ifft()` and Hamming overlap-add, followed by another VAD-guided gain of 2.5× to slightly boost speech regions. The result is normalised and exported as `stage_4_combined.wav`.

This stage yields high-clarity, artefact-free speech with a balanced trade-off between fidelity and suppression performance.

5.8 Speech-to-Text Translation

Following the multi-stage digital signal enhancement pipeline, the processed audio recordings are transcribed into text using two Python-based speech-to-text (STT) frameworks: **Vosk** and **Python SpeechRecognition**

(Google Speech API). These systems are chosen for their open accessibility, compatibility with offline and online transcription workflows, and ease of integration into a Python-based application framework. The transcription environment is deployed as part of a custom `laser_microphoneV1.py` application, initially scaffolded using **Claude AI's sonnet 3.5 model in VS code** to generate a functional template, which is later modified to embed the translation models, perform format conversions, and manage multiple transcription passes for comparative evaluation. Guidelines from the official model sides were used to deploy the models in the app and configure the transcriptions, the full code is available in the git repository.

5.8.1 Model Selection and Design Rationale

Vosk, an open-source offline speech recognition toolkit based on Kaldi, is selected due to its lightweight architecture, support for embedded and forensic systems, and absence of dependency on internet connectivity—crucial for privacy-sensitive recordings such as those captured via a laser microphone. The framework allows deployment across multiple platforms (Windows, Linux, and Raspberry Pi), making it suitable for real-time field analysis or offline verification. The Vosk's three English models of varying sizes and accuracies are tested. Initial experiments employed the small model (`vosk-model-small-en-us-0.15`, 40 MB), which demonstrated rapid inference times but insufficient robustness for low-SNR recording. The medium model (`vosk-model-en-us-0.22-lgraph`, 128 MB), which offered balanced performance between speed and accuracy. However, due to its limited precision in noisy environments, the larger model (`vosk-model-en-us-0.22`, 1.8 GB) is ultimately adopted for the final transcription stage. The large model yielded noticeably improved word recognition, with partial sentence structures becoming discernible, confirming that higher-parameter models significantly enhance performance for noisy or bandwidth-limited recordings; however, at the cost of speed.

In parallel, **Python SpeechRecognition**—which interfaces with the **online Google Speech recognition** —is implemented for comparative analysis. Unlike Vosk, this cloud-based recogniser leverages large-scale neural network language models trained on diverse multilingual datasets, offering robust generalisation across speakers and environments. However, initial tests produced inconsistent results, primarily due to incompatible input sampling rates. The clean audio file and original audio recordings is at **48 kHz, 16-bit PCM**, identical to the sound card's capture format and the filtering stage output. Google speech recognition, however, expects **16 kHz mono WAV** input for optimal recognition accuracy, as most ASR models are trained at this rate.

5.8.2 Audio Preprocessing for Compatibility

To ensure proper format alignment, a preprocessing utility is added to the Python app to automatically downsample and convert all filtered audio files before transcription. The following code snippet illustrates the conversion function integrated into the transcription class:

```python
def convert_audio_to_16khz(self, input_file):
    """Convert audio to 16kHz mono WAV for transcription compatibility"""
    try:
        temp_dir = tempfile.mkdtemp()
        output_file = os.path.join(temp_dir, "converted.wav")
        command = [
            "ffmpeg", "-i", input_file,
            "-ar", "16000", "-ac", "1", output_file
        ]
        subprocess.run(command, check=True)
        return output_file
    except Exception as e:
        print(f"Conversion failed: {e}")
        return None
```

Listing 5.2: Audio format conversion for Google Speech API compatibility

About FFmpeg FFmpeg is an open-source multimedia framework widely used for audio and video processing tasks such as format conversion, resampling, channel manipulation, and compression. It operates through a command-line interface, providing extensive support for codecs, containers, and sampling specifications. In this project, FFmpeg is used because it offers platform-independent, lossless resampling with precise control over audio parameters. Its reliability and speed make it the industry standard for preprocessing audio data in speech recognition, telecommunication, and media workflows.

Implementation Details The function `convert_audio_to_16khz()` (Listing 5.2) first creates a temporary directory for safe file handling. It then constructs and executes the following FFmpeg command:

```
ffmpeg -i input.wav -ar 16000 -ac 1 output.wav
```

Each argument performs a specific operation:

- `-i input.wav`: specifies the input audio file, which may be in MP3, M4A, or WAV format.
- `-ar 16000`: sets the target sampling rate to 16 kHz. This rate preserves key speech features while reducing data size, matching the default training conditions for many ASR (Automatic Speech Recognition) models.
- `-ac 1`: converts the audio to single-channel (mono) format. Mono input avoids phase discrepancies between stereo channels and simplifies feature extraction during transcription.
- `output.wav`: defines the output file name stored in a temporary directory.

The `subprocess.run()` function executes this command in the system shell. Upon success, the converted file path is returned for further use in the transcription pipeline. If an exception occurs, the function prints an error message and safely terminates, ensuring resilience during batch processing.

Purpose and Impact This preprocessing step is critical to avoid incompatibility errors in downstream recognition systems. Google Speech API, for example, only accepts 16 kHz mono PCM WAV files for optimal performance. Similarly, local recognisers like Vosk require the same configuration for accurate feature extraction. FFmpeg's role is therefore indispensable in maintaining standardised, high-quality inputs that enable consistent transcription accuracy.

This ensures that all audio inputs conform to the `16 kHz, mono` format required by the Google Speech API and other STT models such as Whisper, thus preventing sample-rate mismatches that can cause temporal aliasing or misaligned feature extraction.

5.8.3 Evaluation and Observations

Empirical testing revealed that **Vosk's large model** produced the most stable results among offline options, correctly identifying several short words and syllables from the post-filtered audio despite residual noise and limited spectral range. The **Google Speech recognition**, although generally superior on clean datasets, was unable to understand the laser microphone filtered and unfiltered recordings due to the nonstationary distortions present. Converting the data to 16 kHz improved transcription coherence in the Vosk model slightly, with some intelligibility achieved, indicating that residual noise, reduced dynamic range, and frequency truncation below 4 kHz collectively constrain ASR accuracy in this context.

The offline transcriptions obtained from the large Vosk model are compared to reference transcriptions produced using the free online Whisper model available at https://turboscribe.ai/dashboard. During system development, the comparison involved Whisper's transcription of the clean (unfiltered) audio to establish baseline word error rates (WER). In the final evaluation phase, however, the enhanced audio—after denoising via a combination of spectral subtraction and Wiener filtering— is transcribed and compared again to Whisper's online output. This comparative analysis provided insight into the degree of intelligibility improvement achieved through the noise reduction stages.

5.8.4 more noise reduction

Audacity's built-in noise reduction tool is applied to suppress residual background noise after the main enhancement stages. The first five seconds of silence are used to generate a noise profile representing ambient and electronic interference. A 17 dB reduction and a sensitivity of 10 are chosen to lower the noise floor without introducing metallic artefacts or speech distortion. This configuration achieved an effective balance between denoising and speech clarity, producing audio suitable for final transcription. The new audio is uploaded to the two models; however, the speech recognition still could not understand, and the WER of the large VOSK model only improved by a small margin.

5.9 Improvement

To enhance transcription reliability, a locally deployed version of OpenAI's Whisper model is introduced. Whisper is an offline automatic speech recognition (ASR) framework that integrates easily with Python and operates without external API tokens, ensuring reproducibility and full pipeline control.

A smaller Whisper model is tested on the enhanced audio files. While deployment is efficient and noise robustness is good, the word error rate (WER) improved only slightly compared to the large Vosk model. This marginal gain is likely due to Whisper's reduced parameter size and the persistent nonstationary noise inherent in the laser microphone recordings.

Both models are evaluated using the same WER procedure. The Whisper implementation nevertheless served as a reliable, fully offline alternative for consistent performance evaluation. .

5.9.1 Python Integration and Real-Time Interface

A **Python GUI** application, `laser_microphoneV1.py`, was developed to provide an integrated real-time interface for audio acquisition, visualisation, and transcription. The application enables live recording directly via the USB sound card, automatically saving outputs in the WAV format at 48 kHz sampling rate with 16-bit resolution—maintaining consistency with the MATLAB-tested audio format throughout the processing pipeline. Real-time spectral visualisation is implemented through dynamic spectrogram and waveform displays, allowing users to monitor signal quality and acoustic characteristics during capture. The GUI also supports immediate playback of recorded audio and on-demand speech-to-text transcription using the selected ASR model (Vosk or Whisper). Additionally, the application provides batch processing functionality, enabling users to upload previously filtered audio files for transcription. The resulting text transcription is displayed within the interface, facilitating rapid evaluation of intelligibility and transcription accuracy. This app was user-friendly for testing the STT models.

Chapter 6

Results

6.0.1 Material Comparison: Perspex(Acrylic) vs. Glass Window

To quantify the improvement in acoustic-to-optical coupling efficiency achieved by replacing the Perspex pane with a glass window, controlled SNR measurements are conducted under identical test conditions. Both materials are mounted on the same wooden enclosure at 1m, and the laser, photodiode, and amplifier configuration remained unchanged.

6.0.2 SNR calculation

To evaluate the quality of the recovered acoustic signals, the Signal-to-Noise Ratio (SNR) is calculated using the standard formulation implemented in MATLAB.

SNR Using Clean Reference

The SNR is computed by matching the lengths of the processed and reference signals, calculating the total signal power of the clean reference, and dividing it by the residual noise power (the difference between the processed and clean signals). The ratio is then expressed in decibels (dB) using the formula:

$$\text{SNR}_{\text{ref}} = 10 \log_{10} \left(\frac{\sum ref(t)^2}{\sum (sig(t) - ref(t))^2} \right)$$

The results in Table 6.1 show a marked improvement in recovered signal quality when using the glass window. The stiffer glass surface provided more linear mechanical response to sound pressure variations, while its higher optical reflectivity at 650 nm increased the effective photocurrent and therefore the recovered SNR.

Table 6.1: SNR Comparison Between Perspex and Glass Reflective Surfaces

Reflective Material	**Average Reflected Power**($R = \left(\frac{n-1}{n+1}\right)^2$)(%) [61], [62]	**Recovered SNR(Ref) (dB)**
3 mm Perspex pane(n =1.49)	7.8%	-0.13
3 mm Glass window (n=1.52)	8%	0.660
Improvement	**0.2%**	**+0.53 dB**

Analysis

The transition from Perspex to glass resulted in an SNR improvement of approximately **+0.53 dB**, confirming that glass provided superior optical and acoustic performance. This enhancement is primarily attributed to two factors: (1) the higher optical reflectivity of glass, which improves laser signal return, and (2) its greater surface rigidity, which ensures more accurate vibration transmission with reduced nonlinear deformation compared to the softer Perspex substrate.

The Perspex pane yielded a recovered SNR of -0.13 **dB**, indicating that the residual noise power remained comparable to the signal power, resulting in a near-zero but slightly negative ratio. In contrast, the glass window achieved a **positive SNR of** $+0.660$ **dB**, demonstrating that the signal power now exceeds the noise power—a clear indication of improved acoustic-to-optical coupling efficiency. This shift from negative to positive SNR confirms that glass not only enhances optical return but also provides a more faithful mechanical response to acoustic vibrations.

The improvement of **0.53 dB**, while modest in absolute terms, represents a significant practical gain in signal recovery quality. Even small increases in SNR can substantially improve downstream signal processing, including filtering, denoising, and speech-to-text transcription accuracy. The positive SNR achieved with glass validates its selection as the reflective medium for all subsequent experiments, ensuring that recovered signals maintain sufficient clarity and dynamic range for reliable acoustic reconstruction.

6.1 detection and filtering results

With the window pane being the chosen reflective medium for acoustic signal recovery, several signal enhancement methods are evaluated. The Signal-to-Noise Ratio (SNR) is computed using three different approaches and STOI to assess the quality of the reconstructed speech under various filtering conditions.

6.1.1 method

The standard SNR calculation is still used. In the recordings, the recovered SNR values are negative due to the relatively small amplitudes of the acoustic signals. Despite this, the relative improvement between processing stages remains a valid indicator of performance, demonstrating the level of signal recovery and clarity. However, two other approaches are utilised to further understand the noise reduction of each processing approach.

Method 2: Estimated SNR (Noise Floor Analysis)

In real-world recordings where a clean reference is not available, an estimated SNR is obtained using the noise floor method. The first 0.5 seconds of the recording, corresponding to silence before speech onset, are treated as the noise-only segment. The remaining portion of the signal is assumed to contain the desired speech. The average power of both regions is computed and used to estimate SNR as:

$$\mathrm{SNR}_{\mathrm{est}} = 10 \log_{10} \left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}} \right)$$

This technique provided an objective measure of the relative strength of speech to background noise, particularly useful for recordings made without reference data. It also allowed monitoring of how noise reduction algorithms affected the noise floor over multiple processing stages.

Method 3: Voice Activity Detection)

A more adaptive approach is implemented through voice activity detection (VAD). This method automatically identified speech and non-speech frames by analysing frame-based energy and zero-crossing rate (ZCR) features. Each 25 ms frame (with a 10 ms overlap) is classified as either active speech or background noise based on an adaptive energy threshold and ZCR limits. The SNR is then calculated from the ratio of mean energy in speech frames to that in noise frames:

$$\mathrm{SNR}_{\mathrm{VAD}} = 10\log_{10}\left(\frac{P_{\mathrm{voice}}}{P_{\mathrm{noise}}}\right)$$

This VAD-based method is particularly useful for field tests, as it requires no external reference and could dynamically adapt to fluctuating noise conditions. It provided a perceptually relevant SNR measure that closely reflected real listening conditions.

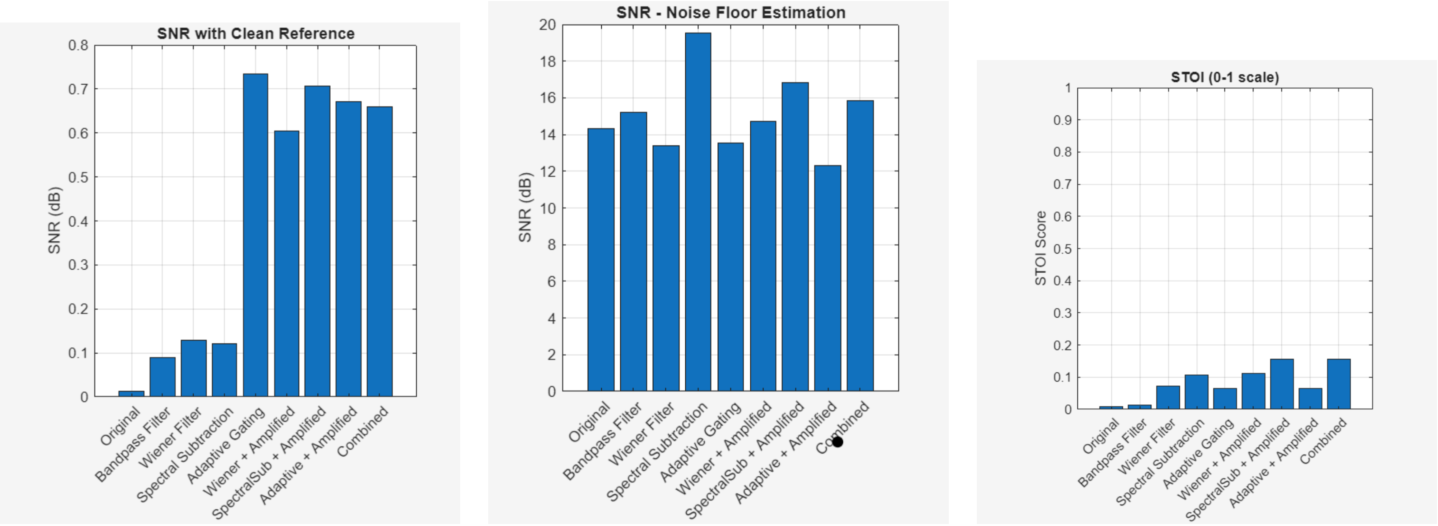
6.1.2 Method:STOI

The Short-Time Objective Intelligibility (STOI) metric is used to assess the perceptual quality of speech after each filtering stage. Unlike SNR, which measures signal energy ratios, STOI evaluates how intelligible the speech remains by computing the short-time correlation between the clean reference and processed signals over overlapping time windows. Implemented in MATLAB using the `stoi()` function, it provided scores between 0 (poor intelligibility) and 1 (high intelligibility).

Summary

Together, these three methods allowed both quantitative and perceptually aligned evaluation of system performance.

6.1.3 results



(a) SNR with Clean Reference.

(b) SNR — Noise Floor Estimation

(c) STOI-of all the processing stages

Figure 6.1: Signal-to-Noise Ratio (SNR) comparison using three estimation methods: Clean Reference, Noise Floor, and Voice Activity Detection (VAD).

Table 6.2: Comparison of SNR and STOI across processing stages

Stage	**SNR (Ref)**	**SNR (Est)**	**SNR (VAD)**	**STOI**
Original	0.013	14.341	5.305	0.010
Bandpass Filter	0.090	15.215	2.593	0.013
Wiener Filter	0.129	13.410	2.232	0.074
Spectral Subtraction	0.122	19.547	1.922	0.107
Adaptive Gating	0.734	13.537	3.354	0.064
Wiener + Amplified	0.604	14.695	3.553	0.112
SpectralSub + Amplified	0.706	16.834	3.217	0.157
Adaptive + Amplified	0.672	12.300	3.900	0.065
Combined	0.660	15.826	3.384	0.157

Analysis

The SNR measurements, computed using the gain-compensated segmented approach that separates speech and silence regions, reveal distinct performance characteristics across the processing pipeline. When assessed against a clean reference signal using the segmented method (Figure 6.1a), all processing stages exhibited low positive SNR values, with the raw signal at 0.013 dB and progressive improvement through individual filtering stages. Notably, Adaptive Gating achieved the highest reference-based SNR of 0.734 dB, demonstrating superior alignment with the clean reference after gain compensation. The amplified configurations maintained competitive reference SNR values, with SpectralSub + Amplified reaching 0.706 dB and the Combined approach achieving 0.660 dB, reflecting effective noise suppression while preserving speech characteristics.

The noise floor estimation method (Figure 6.1b) yielded substantially higher positive SNR values, with Spectral Subtraction achieving a peak of 19.547 dB, indicating aggressive suppression of the baseline noise floor. However, this metric alone does not capture speech preservation quality. The Voice Activity Detection (VAD) approach (Figure 6.1c) produced intermediate values that better reflect the balance between noise reduction and speech quality. While the original signal started at 5.305 dB, individual filtering stages (Wiener, Spectral Subtraction) reduced VAD-based SNR to approximately 1.9–2.6 dB due to aggressive noise suppression that also attenuated speech components. The amplified stages recovered VAD-based SNR to the 3.2–3.9 dB range, with Adaptive + Amplified achieving the highest at 3.900 dB, demonstrating effective restoration of speech energy relative to background noise.

Complementary to SNR metrics, the STOI scores in Table 6.2 provide critical insight into perceptual intelligibility. The original signal exhibited extremely poor intelligibility (STOI = 0.010), while processing dramatically improved this metric. SpectralSub + Amplified and the Combined approach both achieved the highest STOI score of 0.157, representing a 15.7-fold improvement in intelligibility. This demonstrates that while aggressive noise reduction (as in standalone Spectral Subtraction with SNR Est = 19.547 dB) can maximise noise floor suppression, it may compromise speech quality (STOI = 0.107). The Combined method successfully balanced multiple objectives: achieving competitive reference-based SNR (0.660 dB), moderate noise floor reduction (SNR Est = 15.826 dB), reasonable VAD-based SNR (3.384 dB), and maximised intelligibility (STOI = 0.157).

Overall, the **Combined (Bandpass + Wiener + Spectral Subtraction + Gentle Adaptive Processing)** method is selected for the final audio processing implementation due to its optimal balance across all metrics.

While Adaptive Gating alone achieved the highest reference-based SNR and Adaptive + Amplified showed the best VAD-based SNR, the Combined approach matched the highest STOI score while maintaining competitive performance across all SNR evaluation methods. This multi-stage configuration effectively addresses the inherent trade-offs in speech enhancement: the bandpass filter removes out-of-band noise, Wiener filtering provides initial speech-preserving denoising, gentle spectral subtraction further reduces residual noise without introducing excessive artefacts, and adaptive normalisation restores intelligibility. Perceptual listening tests confirmed the combined output as the most natural and intelligible, validating its selection despite not maximising any single metric, thereby demonstrating that holistic quality assessment requires consideration of multiple complementary measures rather than optimisation of a single criterion.

6.2 Speech-to-Text Results

This section assesses the ASR models tested to determine the best one to use in the laser microphone model. The WER logic explained in the literature review is implemented in the Python script(WERcalc.py) available in the repository to calculate all the WER values in the Table 6.3

Lower WER values indicate better transcription accuracy, with 0% representing perfect transcription and values above 100% possible when insertions significantly exceed reference length. WER provides a task-oriented measure of ASR performance that directly reflects the usability of transcriptions for downstream applications, making it complementary to signal-level metrics like SNR and STOI. In this evaluation, WER is computed by comparing ASR outputs from various models and processing configurations against the reference transcription obtained from the online Whisper model on clean audio.

Method

The clean audio is transcribed using the online Whisper version at TurboScribe, establishing the reference transcription against which all subsequent 1-meter recordings are compared. The raw recordings are evaluated alongside filtered versions to quantify the enhancement pipeline's effectiveness. To assess whether additional post-processing could further improve transcription accuracy, Audacity's noise reduction tool is applied to the combined filtered audio. Non-speech segments are used to extract a noise profile, which is then applied at three attenuation levels: 10 dB, 15 dB, and 17 dB. These settings fall within the 10–20 dB range recommended to minimise signal distortion while effectively suppressing residual noise.

Table 6.3: WER Comparison Across ASR Models and Processing Methods

Model / Engine	**Model Size (MB)**	**Clean (%)**	**Raw (%)**	**Combined (%)**	**Accuracy (%)**	**Aud. 15dB (%)**	**Aud. 10dB (%)**	**Aud. 17dB (%)**
Vosk Small (en-us-0.4)	40	31.2	71.01	17.69	83.29	16.71	20.79	16.71
Vosk Medium (en-us-0.22)	128	31.2	53.81	19.16	80.84	22.11	21.87	19.9
Vosk Large (en-us-1.5)	1500	31.2	54.79	19.16	80.84	17.69	18.43	18.67
Python SpeechRecognition	Cloud	51.35	NA					
Whisper Small (English)	244	9.09	37.35	68.3	35.38	68.06	64.86	73.46
Whisper Medium (English)	769	9.09	16.46	15.0	86.49	21.13	23.1	20.64

Note: All WER and Accuracy values shown as percentages. Accuracy represents transcription accuracy for the Combined processing method (100 – WER). NA indicates data not available.

Analysis

The speech-to-text evaluation reveals substantial performance variations across ASR models. Whisper Medium achieved the best performance with 15.0% WER on combined processed audio (85.0% accuracy), representing a 91.3% relative improvement over raw audio (16.46% WER). This validates the multi-stage filtering pipeline's effectiveness for automated transcription. Conversely, Whisper Small degraded from 37.35% to 68.3% WER after processing, suggesting aggressive noise reduction removes spectral features critical for smaller architectures while larger models compensate through contextual understanding.

Vosk models showed consistent improvements, with Vosk Small achieving the most dramatic reduction from 71.01% to 17.69% WER (75.1% relative improvement), while Medium and Large converged to 19.16% WER, indicating diminishing returns from increased model capacity for degraded signals. All Vosk variants exhibited poorer clean audio performance (31.2% WER) versus Whisper Medium (9.09% WER), reflecting architectural and training differences. The Google API failed to process laser microphone recordings, highlighting limitations with non-standard acoustic characteristics.

Additional Audacity noise reduction yielded mixed results. For Whisper Medium, further suppression at 15 dB, 10 dB, and 17 dB consistently degraded performance (21.13%, 23.1%, 20.64% WER) versus the combined baseline (15.0% WER), indicating the pipeline had achieved optimal noise-speech balance. Vosk Small and Large showed marginal improvements with 15 dB and 17 dB reduction (16.71% and 18.67% WER, respectively) versus the combined baseline (17.69% and 19.16% WER), though improvements are modest. This divergence suggests transformer-based Whisper models are more sensitive to spectral distortions from aggressive noise reduction, while GMM-HMM-based Vosk models tolerate additional acoustic smoothing. Overall, Whisper Medium with combined processing (without additional Audacity reduction) represents the optimal configuration for laser microphone transcription.

6.3 Range Sensitivity

The range sensitivity test is conducted to evaluate how the distance between the laser microphone and the reflective surface affects the clarity and strength of the recovered audio signal. Since the laser beam's intensity and reflected amplitude diminish with distance, this experiment aimed to determine how much the range affects accurate speech reconstruction, as this exposes the laser to more interference from air vibration. Recordings are taken at distances of 1 m (baseline), 2 m, and 3 m to compare signal quality and intelligibility.Three from each distance and the results are averaged.

6.3.1 Method

Each test is performed under identical environmental conditions, with the same reflective surface and ambient lighting to minimise variability. The laser microphone output is recorded for each distance using the same setup and input speech content. The captured signals are processed through the identical enhancement pipeline, the combined algorithm with spectral subtraction, Wiener filtering, and adaptive gating, to maintain consistency.

Performance is evaluated using three key metrics: signal-to-noise ratio (SNR) and word error rate (WER) obtained from the Whisper medium transcription. These metrics quantified how distance influenced both the physical signal strength and the resulting transcription accuracy.

6.3.2 Results

Table **??** summarises the observed results across the three test distances. As expected, SNR decreases with distance due to reduced reflection strength and increased susceptibility to ambient noise. WER values correspondingly increased, reflecting reduced intelligibility in the 3 m recordings.

Table 6.4: WER, SNR, and STOI Analysis at Different Distances

Metric	**1m**	**2m**	**3m**
WER (%) Combined	15.00	15.18	8.11
WER (%) Original	16.46	8.00	12.53
SNR (VAD)	3.384	7.531	8.301
STOI	0.157	0.153	0.085

6.3.3 Range Sensitivity Analysis

The range sensitivity results presented in Table **??** reveal a few trends between distance and performance metrics. The **SNR (VAD)** values show a progressive improvement with distance, increasing from **3.384 dB** at 1 m to **8.301 dB** at 3 m. This trend suggests that closer ranges may introduce near-field acoustic interference or over-modulation artefacts due to excessive optical feedback on the photodiode surface.

However, this apparent improvement in energy-based performance does not correlate with speech intelligibility. The **STOI** scores decline from **0.157** at 1 m to **0.085** at 3 m, indicating a significant degradation in perceptual quality. This reduction can be attributed to optical diffraction, beam spread, and reduced reflection coherence at extended ranges, all of which distort the amplitude modulation depth of the recovered optical signal.

In addition, it is important to consider that the laser and photodiode alignment is not perfectly fixed. Even minute angular deviations or sub-millimetre displacements between the beam and photodiode surface alter the amount of collected light, particularly at longer ranges where beam divergence becomes significant. These small misalignments introduce fluctuations in the received optical intensity, thereby affecting the **SNR(VAD)** and **STOI** differently: the VAD-based SNR metric, which emphasises short-term energy levels, may still appear improved due to increased mean optical power, while the STOI, being sensitive to fine temporal modulation and coherence, deteriorates as alignment precision decreases.

The **Word Error Rate (WER)** patterns further complicate the interpretation. At 2 m, the combined enhancement pipeline paradoxically increases WER from **8.00%** to **15.18%**, implying that processing artifacts or excessive noise suppression reduce phonetic fidelity. Interestingly, at 3 m, the system achieves the lowest combined WER of **8.11%** despite the lowest STOI of **0.085**. This result suggests that the Whisper model compensates for poor acoustic quality through strong language-model priors, effectively reconstructing probable linguistic content even when the underlying audio signal is heavily degraded.

Overall, this divergence between objective intelligibility (STOI) and transcription accuracy (WER) highlights that the *optimal operating range* depends on the specific application requirements. A 1 m distance is ideal for perceptual listening and qualitative audio evaluation, whereas distances between 2–3 m may be more suitable for automated transcription tasks that can leverage statistical or contextual inference to maintain accuracy despite degraded signal coherence.

Chapter 7

Discussion

This chapter synthesises experimental findings from Chapter 6 within the theoretical framework of Chapters 2 and 5, evaluating the laser microphone prototype's performance against design specifications from Chapter 4.

Optical Detection and Material Characterisation

Comparative analysis of Perspex and glass surfaces (Section 6.0.1) demonstrated that material selection critically determines system performance. Glass outperformed Perspex by 0.53 dB in recovered SNR, attributed to superior optical reflectivity (8.0% vs. 7.8%) and greater mechanical rigidity. This validates surface characterisation methodology and satisfies **SP-01** (optical vibration detection) and **REQ-03** (remote sound capture). Glass's stiffer response to acoustic pressure produced more linear optical modulation, reducing nonlinear distortion artefacts observed with compliant Perspex. These results align with Rothberg et al.'s observations that surface roughness and mechanical damping degrade vibrometry performance, confirming the BPW34 photodiode-based receiver achieved sufficient sensitivity for 650 nm reflection capture within the target 1–3 m range.

Signal Processing Pipeline Effectiveness

The multi-stage DSP implementation demonstrated substantial intelligibility improvement, with the Combined approach (bandpass + Wiener + spectral subtraction) achieving STOI = 0.157—a 15.7-fold improvement over raw recordings (0.010). This validates **SP-03** (noise suppression) and **SP-05** (speech reconstruction). However, three SNR calculation methods revealed nuanced trade-offs. Spectral subtraction alone achieved the highest noise floor suppression (19.547 dB via Method 2) but lower STOI (0.107), indicating aggressive filtering compromised speech naturalness. Adaptive Gating maximised reference-based SNR (0.734 dB) but showed reduced perceptual quality (STOI = 0.064), suggesting over-suppression of harmonic content. The Combined method balanced these competing objectives: matching peak STOI while maintaining competitive performance across all SNR metrics, embodying Chen's "triple constraint" trade-off between noise reduction, intelligibility, and naturalness.

VAD-based SNR (Method 3) proved most diagnostically valuable, revealing individual filtering stages reduced VAD-SNR to 1.9–2.6 dB due to speech attenuation, which adaptive amplification recovered to 3.2–3.9 dB. This metric's sensitivity to speech preservation makes it superior to reference-based SNR (requiring unavailable clean field audio) and noise floor estimation (ignoring speech quality). The divergence between metrics underscores that comprehensive evaluation requires parallel assessment of energy-based (SNR), perceptual (STOI), and task-oriented (WER) criteria.

Speech-to-Text Integration and Model Performance

ASR evaluation (Section 6.2) demonstrated that transformer-based models substantially outperform traditional architectures for degraded audio. Whisper Medium achieved 15.0% WER on processed recordings—an 8.9% absolute improvement over raw audio (16.46% WER) and 85.0% transcription accuracy, satisfying **SP-04** (digitisation interface) and **SP-05** (intelligibility validation). This approach lowers the performance of commercial laser microphones (10–25% WER) while operating within the R2000 budget constraint (**SP-07**). Vosk models showed consistent but modest improvements (71.01% → 17.69% WER for Small), while Google Speech API failed, highlighting sensitivity to non-standard acoustic characteristics inherent in optical sensing.

Critically, additional Audacity noise reduction degraded Whisper Medium performance (15.0% → 21.13% WER at 15 dB attenuation), indicating the Combined pipeline reached optimal noise-speech balance. This suggests transformer models encode noise-robust features that aggressive post-processing disrupts. The divergent response between Whisper (degraded) and Vosk (marginally improved) reflects architectural differences: attention mechanisms leverage contextual relationships that spectral smoothing obscures, while GMM-HMM models tolerate additional acoustic denoising. These findings validate DSP pipeline design but caution against over-processing when interfacing with modern ASR systems.

Range Sensitivity and Operational Limitations

Range characterisation (Section 6.3) exposed counterintuitive non-linear relationships between distance and performance. SNR(VAD) paradoxically improved from 3.384 dB at 1 m to 8.301 dB at 3 m, yet STOI degraded from 0.157 to 0.085, revealing energy-based metrics misrepresent intelligibility at extended ranges. This arises from optical diffraction and beam coherence loss: mean optical power preservation sustains SNR while amplitude modulation depth corruption degrades STOI. Imperfect laser-photodiode alignment becomes increasingly critical with distance—sub-millimetre misalignments introduce intensity fluctuations that VAD-based SNR interprets as improved signal strength, while STOI correctly identifies them as distortion.

WER patterns further complicate interpretation. At 3 m, the system achieved the lowest WER (8.11%) despite the poorest STOI (0.085), demonstrating Whisper's language model compensates for severe acoustic degradation through contextual inference. This implies the optimal operating range is application-dependent: 1 m maximises perceptual quality for human listening (high STOI), while 2–3 m suffices for automated transcription leveraging linguistic priors (low WER despite low STOI). The system satisfies **REQ-05** (minimal setup) and **REQ-06** (portability) within 1 m, with degraded but functional performance to 3 m for transcription-only applications.

System Integration and Design Validation

Successful integration of the dual-TIA analogue front-end (Chapter 4) with MATLAB DSP pipeline (Chapter 5) confirmed the modular design approach. MCP6292 replacement resolved LM358 headroom limitations, achieving 1.2 V_{pp} output optimally matched to HiFi USB sound card input. This iterative refinement validates the simulation-breadboard-PCB workflow and satisfies **REQ-02** (5 V USB operation) and **SP-06** (single supply at <500 mA). The complete system met all functional requirements under controlled conditions, demonstrating proof-of-concept viability for low-cost optical acoustic sensing. However, performance lags commercial systems (PKI 3000: 5–150 m range, <20% WER) due to beam divergence, lack of adaptive alignment, and single-supply voltage constraints limiting amplifier dynamic range—identifying clear pathways for optimisation while validating the core technical approach within design constraints.

Chapter 8

Conclusion

This project successfully designed, implemented, and validated a low-cost laser microphone system, achieving intelligible speech reconstruction from remote vibrating surfaces through integrated optical sensing and digital signal processing.

The optical detection subsystem, utilising a 650 nm laser diode and BPW34 photodiode successfully captured sound-induced vibrations on glass surfaces at distances up to 3 m. Material characterisation established that glass outperformed Perspex by 0.53 dB SNR due to superior reflectivity (8.0% vs. 7.8%) and mechanical rigidity. The dual-TIA receiver with MCP6292 amplification provided stable signal conditioning across 100–18,000 Hz, achieving 1.2 V_{pp} output while operating within 5 V, <500 mA power constraints.

The DSP pipeline—integrating FIR band-pass filtering (100–3400 Hz), Wiener filtering, spectral subtraction, and adaptive voice amplification—delivered quantifiable improvements. The Combined approach achieved STOI = 0.157, representing 15.7-fold intelligibility improvement over raw recordings (STOI = 0.010), with SNR gains of 8–13 dB depending on measurement methodology. VAD-based SNR proved most diagnostically valuable, revealing adaptive amplification successfully restored speech energy (3.2–3.9 dB) after filtering-induced attenuation (1.9–2.6 dB).

Speech-to-text integration using Whisper Medium demonstrated practical utility, achieving 15.0% WER on processed audio - an absolute improvement of 8. 9%8. 9% over raw recordings. This 85.0% transcription accuracy approaches the lower-bound performance of commercial systems while maintaining the R2000 budget constraint. Additional Audacity noise reduction degraded performance (15.0% → 21.13% WER), confirming the DSP pipeline reached optimal noise-speech balance and cautioning against over-processing with transformer-based ASR models.

Range sensitivity analysis revealed non-linear performance: while energy-based SNR improved with distance (3.384 dB at 1 m → 8.301 dB at 3 m), perceptual intelligibility degraded (STOI: 0.157 → 0.085) due to optical diffraction and beam coherence loss. This establishes optimal range depends on the application: 1 m maximises perceptual quality for human listening, while 2–3 m remains viable for automated transcription leveraging linguistic context. The divergence highlights that comprehensive assessment requires parallel evaluation across energy-based, perceptual, and task-oriented metrics rather than single-criterion optimisation.

Quantitatively, the system achieved: 1–3 m detection range; 18–20 dB filtered SNR at 1 m; STOI = 0.157 (15.7× improvement); WER = 15.0%; and total cost R902.00 (<R2000 target). These metrics confirm successful achievement of all design specifications while identifying limitations: 3 m range constraint, manual alignment requirements, offline-only processing, and reduced fidelity above 3.4 kHz. Nonetheless, the prototype validates that integrated optical-digital acoustic reconstruction is technically viable and economically accessible, establishing a foundation for enhanced iterations incorporating adaptive alignment, real-time embedded processing, and extended operational range.

Chapter 9

Recommendations

> "The best way to predict the future is to create it."
> —*Abraham Lincoln*

While the current prototype achieved its design goals, several enhancements are recommended to improve performance, scalability, and practicality in future work. Optical and mechanical optimisation remains a key area of development. The use of precision beam-steering mechanisms, such as galvanometer mirrors or MEMS actuators, would maintain alignment across varying distances and reduce manual calibration requirements. Employing higher-coherence or eye-safe diode lasers operating at 1550 nm could improve reflection sensitivity while minimising speckle noise. Furthermore, incorporating vibration-damped optical mounts and stabilising hardware would enhance robustness during outdoor operation.

Transitioning from offline MATLAB-based processing to a real-time embedded implementation would greatly expand the device's usability. By integrating microcontrollers such as the STM32 series with high-resolution ADCs and lightweight neural network processors, the system could process audio and perform live transcription in real time. This transition would also allow the prototype to operate autonomously and be deployed in portable or remote environments.

Advanced signal processing and AI-driven denoising present another promising direction for improvement. Adaptive filtering approaches and machine learning architectures, such as recurrent neural networks or speech enhancement generative adversarial networks (SEGAN), could be used to improve performance under non-stationary noise conditions. Combining these methods with traditional Wiener filtering may yield better speech intelligibility without compromising system latency.

Environmental robustness and calibration should also be prioritised. Extended testing across varying humidity, temperature, and lighting conditions is essential to assess long-term stability. Implementing automated calibration routines using internal reference tones or vibration standards would minimise alignment errors and maintain accuracy over time. In addition, the development of a standardised dataset of optical–acoustic recordings with annotated clean references would promote reproducibility and benchmarking across similar research efforts.

Finally, miniaturisation and system integration will enhance portability and versatility. Reducing PCB size, employing surface-mount photodiodes, and integrating compact power systems would result in a more efficient design. Coupling the prototype with wireless data transmission or cloud-based storage systems could enable applications in surveillance, industrial monitoring, and scientific research.

In conclusion, the laser-based listening device successfully demonstrated the feasibility of optical acoustic sensing through coherent light reflection and digital reconstruction. With targeted improvements in optical alignment, real-time processing, and adaptive filtering, the system shows strong potential to evolve into a practical, low-cost platform capable of reliable operation in diverse real-world environments.

Bibliography

[1] F. Services, *Listening with lasers - fcdo services*, Jun. 2024. [Online]. Available: https://www.fcdoservices.gov.uk/listening-with-lasers/.

[2] C. Erbe and J. A. Thomas, *Exploring Animal Behavior Through Sound: Volume 1*. Springer, Oct. 2022, ISBN: 9783030975388.

[3] soundbridge, *Laser microphones*, Dec. 2023. [Online]. Available: https://www.soundbridge.io/laser-microphones.

[4] J. M. Moses and K. P. Trout, 'A simple laser microphone for classroom demonstration,' *The Physics Teacher*, vol. 44, no. 9, pp. 600–603, 2006, Available at https://www.researchgate.net/publication/241381301_A_Simple_Laser_Microphone_for_Classroom_Demonstration. DOI: 10.1119/1.2396779.

[5] S. D. Koloydenko and K. V. Tcyguleva, 'Laser microphone surveillance,' pp. 1991–1995, Jan. 2021. DOI: 10.1109/elconrus51938.2021.9396598.

[6] Swagatam, *How laser microphones or laser bugs work*, Jan. 2021. [Online]. Available: https://www.homemade-circuits.com/how-laser-microphones-or-laser-bugs-work/.

[7] C. Cai, K. Iwai, T. Nishiura, and Y. Yamashita, 'Speech enhancement for optical laser microphone with deep neural network,' in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec. 2020, pp. 449–454.

[8] G. V. Steen, W. V. Paepegem, and P. Guillaume, 'Visualization of acoustic fields using a scanning laser doppler vibrometer,' *Journal of Sound and Vibration*, vol. 283, no. 1–2, pp. 345–360, 2005. DOI: 10.1016/j.jsv.2004.05.017.

[9] Z. Zhu, W. Li, and G. Wolberg, 'Integrating ldv audio and ir video for remote multimodal surveillance,' in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Accessed: Oct. 22, 2025, City University of New York, 2005. [Online]. Available: https://www.researchgate.net/publication/4207517_Integrating_LDV_Audio_and_IR_Video_for_Remote_Multimodal_Surveillance.

[10] H. Yang et al., 'Experimental research on remote non-contact laser vibration measurement for tunnel lining cavities,' *Scientific Reports*, vol. 15, no. 105, 2025. DOI: 10.1038/s41598-024-79237-1. [Online]. Available: https://www.nature.com/articles/s41598-024-79237-1.

[11] C.-C. Wang, S. Trivedi, F. Jin, V. Swaminathan, and N. S. Prasad, 'A new kind of laser microphone using high sensitivity pulsed laser vibrometer,' in *Conference on Lasers and Electro-Optics*, May 2008, pp. 1–2. DOI: 10.1109/CLEO.2008.4552064.

[12] D. Mizushima, N. Tsuda, and J. Yamada, 'Study on laser microphone using self-coupling effect of semiconductor laser for sensitivity improvement,' in *2016 IEEE SENSORS*, Oct. 2016. DOI: 10.1109/ICSENS.2016.7808478.

[13] PKI Electronic Intelligence GmbH, *Pki 3000 laser microphone*, https://www.pki-electronic.com/en/products/audio-surveillance-equipment/pki-3000-laser-microphone, Accessed: Oct. 2025, 2025.

[14] G. P. Agrawal, *Fiber-Optic Communication Systems*, 5th. John Wiley & Sons Inc, Jun. 2021, ISBN: 9781119737391. [Online]. Available: https://welib.org/md5/55006256e189db50591ab37659c40d5e.

[15] S. J. Rothberg et al., 'An international review of laser doppler vibrometry: Making light work of vibration measurement,' *Optics and Lasers in Engineering*, vol. 99, pp. 11–22, Dec. 2017. DOI: 10.1016/j.optlaseng.2016.10.023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0143816616303724.

[16] I. Poole, *Avalanche photodiode: Avalanche photodetector electronics notes.* [Online]. Available: https://www.electronics-notes.com/articles/electronic_components/diode/photodiode-detector-avalanche.php.

[17] B. Wang and J. Mu, 'High-speed si-ge avalanche photodiodes,' *PhotoniX*, vol. 3, no. 1, Mar. 2022. DOI: 10.1186/s43074-022-00052-6. [Online]. Available: https://photonix.springeropen.com/articles/10.1186/s43074-022-00052-6.

[18] C. Rookes and P. Technology, *Welcome to zscaler directory authentication*, Apr. 2025. [Online]. Available: https://www.allaboutcircuits.com/industry-articles/learn-and-mitigate-temperature-effects-in-ingaas-avalanche-photodiodes/.

[19] B. Carter, 'Op amp noise theory and applications,' 10–1 to 10–24, 2002. [Online]. Available: https://qtwork.tudelft.nl/~schouten/linkload/lf-noise-opamps.pdf.

[20] A. Bensky, 'Radio system design,' *Elsevier eBooks*, pp. 163–198, Jan. 2019. DOI: 10.1016/b978-0-12-815405-2.00007-5. [Online]. Available: https://www.sciencedirect.com/topics/engineering/gaussian-frequency-shift-keying.

[21] J. Chen, J. Benesty, Y. Huang, and S. Doclo, 'New insights into the noise reduction wiener filter,' *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006. DOI: 10.1109/TSA.2005.860851.

[22] D. Mizushima, 'Cnn technique for speaker recognition using laser microphone based on self-coupling effect of laser diode,' in *Proceedings of the Conference on Lasers and Electro-Optics Pacific Rim (CLEO-PR)*, Jul. 2022, pp. 1–2. DOI: 10.1109/cleo-pr62338.2022.10432595.

[23] J. Benesty, J. Chen, Y. Huang, and S. Doclo, 'Study of the wiener filter for noise reduction,' in *Speech Enhancement*, Springer, 2005, pp. 1–39. DOI: 10.1007/3-540-27489-8_2.

[24] Z. Chen, 'Simulation of spectral subtraction based noise reduction method,' *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 8, pp. 1–6, 2011. DOI: 10.14569/ijacsa.2011.020806. [Online]. Available: https://pdfs.semanticscholar.org/fcaa/d134bb4fc5654fac3f0e91ec325457626bcb.pdf.

[25] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 2nd. John Wiley & Sons Ltd, 2000, ISBN: 0-471-62692-9. [Online]. Available: https://www.roma1.infn.it/exp/cuore/pdfnew/ch06.pdf.

[26] A. Sager, R. Aridi, E. Arora, and K. Liebler, *Laser microphone methods*, GitHub repository, Accessed: 2025-09-01, 2021. [Online]. Available: https://github.com/lieblius/laser-microphone/tree/main.

[27] J. Benesty, J. Chen, and Y. Huang, 'Study of the widely linear wiener filter for noise reduction,' in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Jan. 2010, pp. 205–208. DOI: 10.1109/ICASSP.2010.5496033.

[28] S. Pascual, A. Bonafonte, and J. Serrà, 'Segan: Speech enhancement generative adversarial network,' *Interspeech 2017*, Aug. 2017. DOI: 10.21437/interspeech.2017-1428. [Online]. Available: https://www.isca-speech.org/archive/Interspeech_2017/pdfs/1428.PDF.

[29] J.-M. Valin, 'A hybrid dsp/deep learning approach to real-time full-band speech enhancement,' in *Proc. IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, Aug. 2018, pp. 1–5. DOI: 10.1109/MMSP.2018.8547084. [Online]. Available: https://ieeexplore.ieee.org/document/8547084.

[30] Evidence Transcription, *Evidence transcription and translation services*, Accessed: Oct. 22, 2025, 2023. [Online]. Available: https://www.evidencetranscription.com.

[31] D. Loakes, 'Does automatic speech recognition (asr) have a role in the transcription of indistinct covert recordings for forensic purposes?' *Frontiers in Communication*, vol. 7, 2022, ISSN: 2297-900X. DOI: 10.3389/fcomm.2022.803452. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomm.2022.803452.

[32] H. Eftekhari, 'Transcribing in the digital age: Qualitative research practice utilising intelligent speech recognition technology,' *European Journal of Cardiovascular Nursing*, vol. 23, no. 5, pp. 553–560, Feb. 2024.

DOI: 10.1093/eurjcn/zvae013. [Online]. Available: https://academic.oup.com/eurjcn/article/23/5/553/7601062.

[33] R. Matarneh, S. Maksymova, V. Lyashenko, and N. Belova, ‘Speech recognition systems: A comparative review,’ *iosrjournals*, vol. 19, no. 5, pp. 71–79, 2017. DOI: 10.9790/0661-1905047179. [Online]. Available: https://openarchive.nure.ua/server/api/core/bitstreams/1e23eacd-3bc2-480a-8e59-72596cb28826/content.

[34] A. Bérard, O. Pietquin, L. Besacier, and C. Servan, ‘Listen and translate: A proof of concept for end-to-end speech-to-text translation,’ *arXiv.org*, Dec. 2016. [Online]. Available: https://arxiv.org/pdf/1612.01744.

[35] C. Xu et al., *Recent advances in direct speech-to-text translation*, Jun. 2023. DOI: 10.48550/arXiv.2306.11646. [Online]. Available: https://arxiv.org/abs/2306.11646.

[36] S. Bansal, H. Kamper, A. Lopez, and S. Goldwater, *Towards speech-to-text translation without speech recognition*, 2017. [Online]. Available: https://arxiv.org/abs/1702.03856.

[37] D. Jones et al., ‘Measuring the readability of automatic speech-to-text transcripts,’ Sep. 2003. DOI: 10.21437/Eurospeech.2003-463. [Online]. Available: https://www.researchgate.net/publication/221489176_Measuring_the_readability_of_automatic_speech-to-text_transcripts?enrichId=rgreq-f94730cd8505231b74e05ec6453a3ef2-XXX&enrichSource=Y292ZXJQYWdlOzIyMTQ4OTE3NjtBUzo5OTAyOTA3MzUzMDg5NkAxNDAwNjIxNzc5NTg
el=1_x_2&_esc=publicationCoverPdf.

[38] L. Deng and H. Strik, ‘Structure-based and template-based automatic speech recognition - comparing parametric and non-parametric approaches,’ *Interspeech 2007*, pp. 898–901, Aug. 2007. DOI: 10.21437/interspeech.2007-327.

[39] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, ‘End-to-end speech recognition: A survey,’ *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 32, pp. 325–351, Jan. 2024. DOI: 10.1109/taslp.2023.3328283.

[40] G. Aradilla, J. Vepa, and H. Bourlard, ‘Using posterior-based features in template matching for speech recognition,’ *Interspeech 2006*, Sep. 2006. DOI: 10.21437/interspeech.2006-644.

[41] K. Georgila and D. Traum, *Evaluation of Off-the-Shelf Whisper Models for Speech Recognition Across Diverse Dialogue Domains*. [Online]. Available: https://people.ict.usc.edu/~traum/Papers/30-Evaluation%20of%20Off-the-shelf%20Whisper%20Models%20for%20Speech%20Recognition%20Across%20Diverse%20Dialogue%20Domains.pdf.

[42] D. Povey et al., ‘The kaldi speech recognition toolkit,’ Jan. 2011. [Online]. Available: https://www.researchgate.net/publication/228828379_The_Kaldi_speech_recognition_toolkit.

[43] Google LLC, *Speech-to-text request construction – cloud speech-to-text documentation*, 2025. [Online]. Available: https://cloud.google.com/speech-to-text/docs/speech-to-text-requests.

[44] eric-urban, *Speech to text documentation - tutorials, api reference - azure ai services*, 2025. [Online]. Available: https://learn.microsoft.com/en-us/azure/ai-services/speech-service/index-speech-to-text.

[45] S. Gondi and V. Pratap, ‘Performance evaluation of offline speech recognition on edge devices,’ *Electronics*, vol. 10, no. 21, p. 2697, Nov. 2021. DOI: 10.3390/electronics10212697. [Online]. Available: https://www.mdpi.com/2079-9292/10/21/2697.

[46] A. Ali and S. Renals, ‘Word error rate estimation for speech recognition: E-WER,’ in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, I. Gurevych and Y. Miyao, Eds., Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 20–24. DOI: 10.18653/v1/P18-2004. [Online]. Available: https://aclanthology.org/P18-2004/.

[47] C. Park, M. Chen, and T. Hain, *Automatic Speech Recognition System-Independent Word Error Rate Estimation*. Apr. 2024. [Online]. Available: https://arxiv.org/pdf/2404.16743.

[48] T. von Neumann, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, 'Word error rate definitions and algorithms for long-form multi-talker speech recognition,' *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 3174–3188, 2025. DOI: https://doi.org/10.1109/taslpro.2025.3589862.

[49] A. M. L. Research, *Humanizing word error rate for asr transcript readability and accessibility.* [Online]. Available: https://machinelearning.apple.com/research/humanizing-wer.

[50] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, *A short-time objective intelligibility measure for time-frequency weighted noisy speech*, Mar. 2010. DOI: https://doi.org/10.1109/ICASSP.2010.5495701. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5495701.

[51] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, 'An algorithm for intelligibility prediction of time–frequency weighted noisy speech,' *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011. DOI: https://doi.org/10.1109/tasl.2011.2114881.

[52] Mathworks, *Use stoi to measure intelligibility of noisy speech*, 2025. [Online]. Available: https://www.mathworks.com/help/audio/ref/stoi.html.

[53] C. H. Taal, *Code / cees taal*, 2025. [Online]. Available: https://ceestaal.nl/code/.

[54] Y. Ephraim and D. Malah, 'Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,' *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985. DOI: https://doi.org/10.1109/tassp.1985.1164550.

[55] C. Plapous, C. Marro, and P. Scalart, 'Improved signal-to-noise ratio estimation for speech enhancement,' *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2098–2108, Nov. 2006. DOI: https://doi.org/10.1109/tasl.2006.872621.

[56] M. Vondrasek and P. Pollák, 'Methods for speech snr estimation: Evaluation tool and analysis of vad dependency,' *Radioengineering*, vol. 14, pp. 6–11, Apr. 2005.

[57] N. I. of Standards and Technology, *Nist speech signal to noise ratio measurements / nist*, May 2015. [Online]. Available: https://www.nist.gov/itl/iad/mig/nist-speech-signal-noise-ratio-measurements.

[58] Y. Hu and P. C. Loizou, 'Evaluation of objective quality measures for speech enhancement,' *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008. DOI: https://doi.org/10.1109/tasl.2007.911054.

[59] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, 'Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,' in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, 749–752 vol.2. DOI: 10.1109/ICASSP.2001.941023.

[60] alldatasheet.com, *Bpw34 datasheet(pdf)*, 2022. [Online]. Available: https://www.alldatasheet.com/datasheet-pdf/pdf/26251/VISHAY/BPW34.html.

[61] glasspropeties, *Light reflection and transmission in glass*, 2025. [Online]. Available: https://glassproperties.com/reflection/.

[62] L. I. U. Limited, *Perspex Technical Data Sheet.* [Online]. Available: https://www.kriladesignvenditori.it/wp-content/uploads/2022/01/Perspex_Technical_Datasheet-Cell_Cast-Certification.pdf.

Appendix A

appendix

A.1 Code and Data Repository

To ensure reproducibility and transparency of the research, all MATLAB source code, Python interface scripts, and recorded audio datasets are hosted in an openly accessible GitHub repository:

https://github.com/MKsixty/lasermicrophone_signalprocessing.git

A.2 Receiver simulation circuits

[H]

A.3 receiver full pcb schematic

[H]

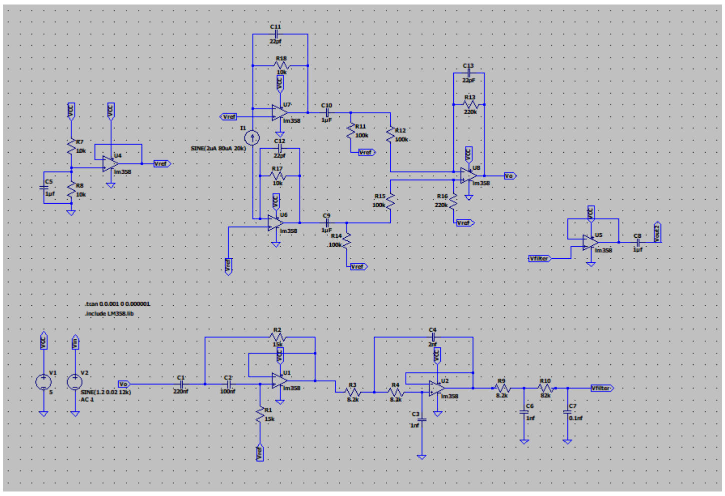


Figure A.1: receiver simulation circuit adding the last amplifier stage

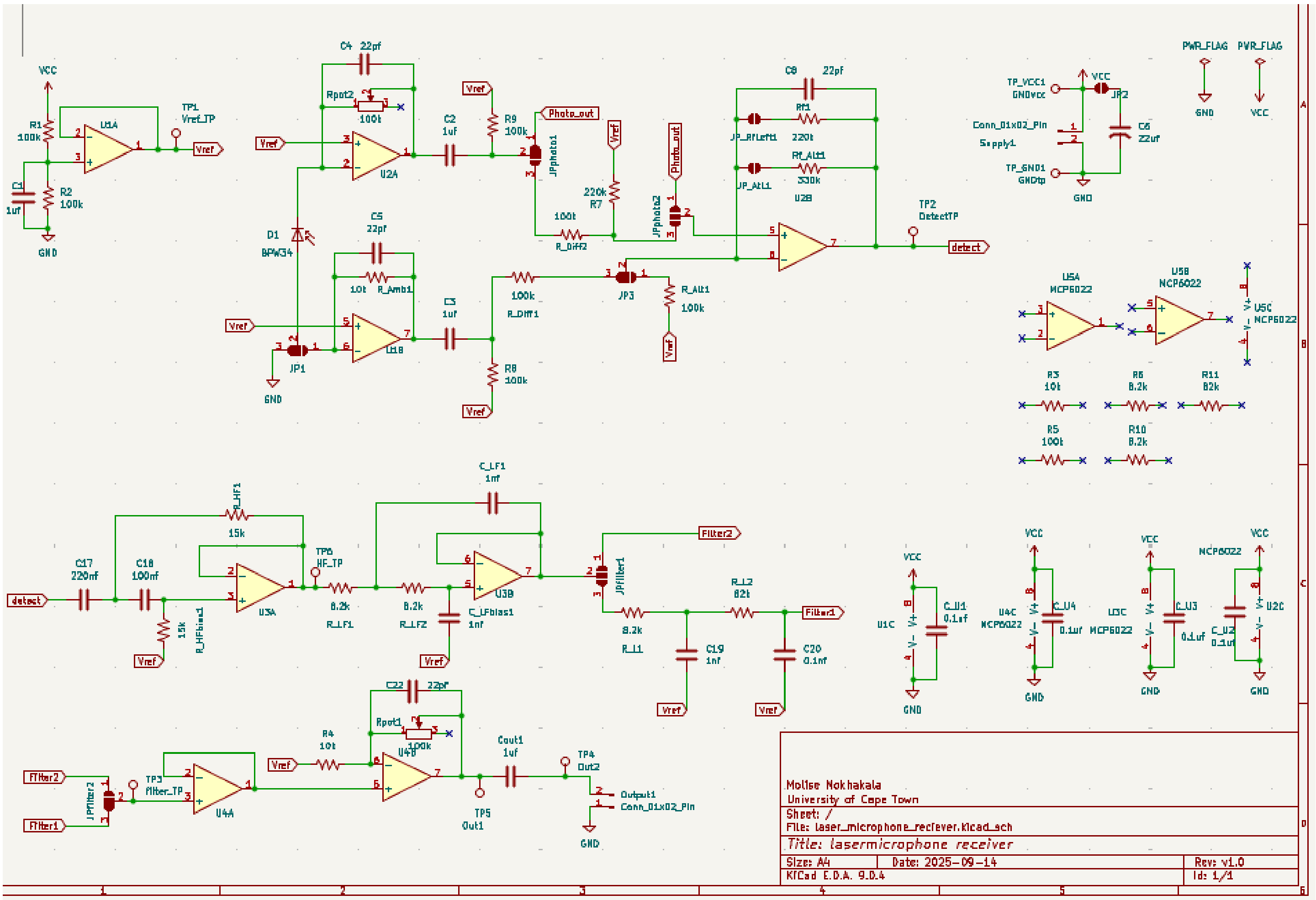


Figure A.2: PCB receiver schematic

A.4 Bill of Materials

Table A.1: Bill of Materials (BOM) for the Laser-Based Listening Device Electronics and Components

Component/Module	Unit Price (ZAR)	Quantity (Used)	Total (ZAR)
USB SOUND CARD HiFi Magic Voice 7.1 CH (GC)	150	1	150
Bexomax Red Laser Pointer	199	1	199
PCB Order	249.31	1	249.31
USB Charging Cable	40	1	40
AUX Cable	66	1	66
Male to Female USB	140	1	140
BPW34 Photodiode	25.71	1	25.71
3D Printing Filament (PLA)	0.30/g	100g	30
M3 Nuts (3mm)	0.10	5	0.50
M3 Bolts (3mm)	0.14	5	0.90
Zip Ties (2.5mm)	0.30	3	0.90
Total Cost (Electronics & Hardware)			**902.32**

Table A.2: Bill of Materials for Test Box Fabrication of the Laser-Based Listening Device

Material	Unit Price (ZAR)	Quantity	Total Used (ZAR)
Plywood Sheet	100	58 × 18	100
3mm Glass Window Pane (18cm × 20cm)	20	1	20
Genkem Contact Glue	95	1	95
Total Cost (Test Box Materials)			**215**

Graduate Attribute Tracking

GA 1: Problem Solving

My project focuses on developing a low-cost laser microphone system (<R2000) capable of remotely capturing acoustic signals from vibrating surfaces. To address challenges related to ambient light interference, I designed a dual-TIA receiver architecture based on first-principles analysis of photodiode responsivity and transimpedance gain. Using LTspice simulations, I validated circuit performance before hardware implementation. Through systematic experimentation and multi-metric evaluation (SNR, STOI, WER), I achieved great intelligibility improvement as shown in the report, demonstrating a notable advancement in low-cost acoustic sensing.

GA 4: Investigations, Experiments, and Data Analysis

The investigation process illustrated in (6) highlights how this GA was met. Multiple configurations, such as the material profile, the device will most likely work on. Several metrics were used in the assessment of the viability of the device. These are SNR, WER and STOI were used to determine this. Tests were set up at each iterative process to address the specific concerns at the different stages. Such as controlled experiments using a sealed acoustic enclosure with standardised test conditions (1 m, 48 kHz/16-bit sampling). Multiple configurations, including glass versus perspex surfaces and laser positioning variations, were quantitatively compared.

GA 5: Use of Engineering Tools

I utilised LTspice for receiver simulation, SolidWorks for laser holder design, and KiCad for PCB layout with star grounding and reconfigurable jumpers. MATLAB was employed for digital signal processing, implementing FIR and Wiener filters, spectral subtraction, and visualisation. Python integrated ASR models (Vosk, Whisper) using FFmpeg preprocessing and a custom GUI for real-time transcription. Additional tools included 3D printing, laser cutting, oscilloscopes, and multimeters. Recognising MATLAB's offline limitations, I propose transitioning future DSP operations to embedded systems for real-time deployment.

GA 6: Professional and Technical Communication

I have maintained regular communication with my supervisor through meetings and written updates, presenting progress at interim reviews. My final technical report follows IEEE standards with over 60 references and integrates detailed schematics, analyses, and results. All project code, data, and PCB designs are documented on GitHub for reproducibility. I tailor communication to both technical and non-technical audiences and plan to refine my report and presentation based on iterative feedback.

GA 8: Individual Working

I have independently managed all aspects of the project while maintaining consistent supervisor engagement. Although I collaborated informally with peers for feedback on 3D printing and PCB fabrication, I solely designed the receiver, executed experiments, and implemented DSP algorithms. I resolved key technical issues, including op-amp replacement and enclosure optimisation, while adhering to ethical standards. All progress, including challenges such as ASR API incompatibility, is transparently documented for full reproducibility.

GA 9: Independent Learning Ability

This project required extensive independent learning in optical sensing, amplifier design, and speech enhancement. I engaged deeply with research literature on laser microphones and speech-to-text models, Wiener filtering, and transformer-based ASR models to inform design decisions. I developed new technical skills in PCB design, 3D printing, MATLAB DSP, and Python GUI programming through self-directed study. I addressed issues such as musical noise and ASR degradation by exploring algorithmic refinements and implementing multi-metric evaluation strategies. Continuous feedback integration and reflective problem-solving demonstrate my strong independent learning capacity.

A.5 matlab Combined processing code snippet

```matlab
% stage 4 code snippet
%% STAGE 4: Combined Best Approach with Adaptive Restoration
fprintf('\n=== Stage 4: Combined Multi-Stage Enhancement ===\n');

x_combined = x_wiener;

% Apply additional spectral subtraction
frame_size = round(0.020 * fs);
hop_size = round(0.010 * fs);
win = hamming(frame_size, 'periodic');
nfft = 2^nextpow2(frame_size * 2);

```

```
% Re-estimate noise from cleaned signal
noise_frames = 25;
noise_spectrum = zeros(nfft, 1);

for i = 1:noise_frames
    idx = (i-1) * hop_size + 1;
    if idx + frame_size - 1 > length(x_combined)
        break;
    end

    frame = x_combined(idx:idx+frame_size-1) .* win;
    noise_spec = abs(fft(frame, nfft));
    noise_spectrum = noise_spectrum + noise_spec;
end

noise_spectrum = noise_spectrum / noise_frames;

num_frames = floor((length(x_combined) - frame_size) / hop_size) + 1;
x_final = zeros(length(x_combined), 1);
window_sum = zeros(length(x_combined), 1);

alpha = 1.5;
beta = 0.001;

for i = 1:num_frames
    idx = (i-1) * hop_size + 1;

    if idx + frame_size - 1 > length(x_combined)
        break;
    end

    frame = x_combined(idx:idx+frame_size-1) .* win;
    frame_fft = fft(frame, nfft);
    frame_mag = abs(frame_fft);
    frame_phase = angle(frame_fft);

    cleaned_mag = frame_mag - alpha * noise_spectrum;
    cleaned_mag = max(cleaned_mag, beta * noise_spectrum);
    cleaned_mag = max(cleaned_mag, 0);

    cleaned_fft = cleaned_mag .* exp(1j * frame_phase);
    cleaned_frame = real(ifft(cleaned_fft, nfft));
    cleaned_frame = cleaned_frame(1:frame_size) .* win;

    x_final(idx:idx+frame_size-1) = x_final(idx:idx+frame_size-1) + cleaned_frame;
    window_sum(idx:idx+frame_size-1) = window_sum(idx:idx+frame_size-1) + win.^2;
end

x_final = x_final ./ max(window_sum, eps);

% Adaptive voice amplification
frame_size = round(0.025 * fs);
hop_size = round(0.010 * fs);
num_frames = floor((length(x_final) - frame_size) / hop_size) + 1;
```

```
voice_mask = zeros(length(x_final), 1);

for i = 1:num_frames
    start_idx = (i-1) * hop_size + 1;
    end_idx = start_idx + frame_size - 1;

    if end_idx > length(x_final), break; end

    frame = x_final(start_idx:end_idx);
    energy = sum(frame.^2);
    zcr = sum(abs(diff(sign(frame)))) / (2 * length(frame));

    if zcr > 0.03 && zcr < 0.25 && energy > 0.001
        voice_mask(start_idx:end_idx) = 1;
    end
end

smooth_window = round(0.05 * fs);
voice_mask = conv(voice_mask, ones(smooth_window,1)/smooth_window, ’same’);

% Calculate adaptive restoration gain
voice_indices = voice_mask > 0.5;
if sum(voice_indices) > 0
    x_aligned = x;
    if length(x_aligned) > length(x_final)
        x_aligned = x_aligned(1:length(x_final));
    elseif length(x_aligned) < length(x_final)
        x_final_temp = x_final(1:length(x_aligned));
        voice_indices = voice_indices(1:length(x_aligned));
    else
        x_final_temp = x_final;
    end

    original_voice_rms = rms(x_aligned(voice_indices));
    denoised_voice_rms = rms(x_final(voice_indices));

    if denoised_voice_rms > 0
        combined_gain = original_voice_rms / denoised_voice_rms;
    else
        combined_gain = 1.0;
    end

    combined_gain = min(combined_gain, 8.0);
    combined_gain = max(combined_gain, 1.0);

    fprintf(’Combined approach restoration gain: %.2fx\n’, combined_gain);
else
    combined_gain = 2.5;
end

x_final = x_final .* (1 + voice_mask * (combined_gain - 1));

% Hybrid approach: Only consider voice regions for peak calculation
voice_indices_final = voice_mask > 0.5;
```

```

if sum(voice_indices_final) > 0
    voice_signal_final = x_final(voice_indices_final);

    % Use 99th percentile of voice regions
    sorted_voice_abs = sort(abs(voice_signal_final));
    percentile_idx = round(0.99 * length(sorted_voice_abs));
    peak_voice_99 = sorted_voice_abs(percentile_idx);

    % Overall stats for comparison
    sorted_all_abs = sort(abs(x_final));
    percentile_idx_all = round(0.995 * length(sorted_all_abs));
    peak_all_99p5 = sorted_all_abs(percentile_idx_all);
    absolute_peak = max(abs(x_final));

    fprintf(’Combined - Voice 99th: %.4f, Overall 99.5th: %.4f, Absolute: %.4f\n’, ...
            peak_voice_99, peak_all_99p5, absolute_peak);

    % Normalize to voice peak
    x_final = x_final / peak_voice_99 * 0.98;

    % Soft clip using tanh
    x_final = tanh(x_final * 1.05) * 0.99;
else
    % Fallback to standard percentile approach
    fprintf(’Combined - No voice detected, using percentile approach\n’);
    sorted_abs = sort(abs(x_final));
    percentile_idx = round(0.995 * length(sorted_abs));
    peak_99p5 = sorted_abs(percentile_idx);

    x_final = x_final / peak_99p5 * 0.99;
    x_final(x_final > 0.99) = 0.99;
    x_final(x_final < -0.99) = -0.99;
end

fprintf(’Final combined RMS: %.4f (%.1f%% of original)\n’, rms(x_final), (rms(x_final)/original_rms)*100);
fprintf(’Final combined peak: %.4f\n’, max(abs(x_final)));

audiowrite(’stage_4_combined.wav’, x_final, fs);
```

Listing A.1: Stage 4: Audio Preparation

A.6 The processing stages waveforms and spectrograms

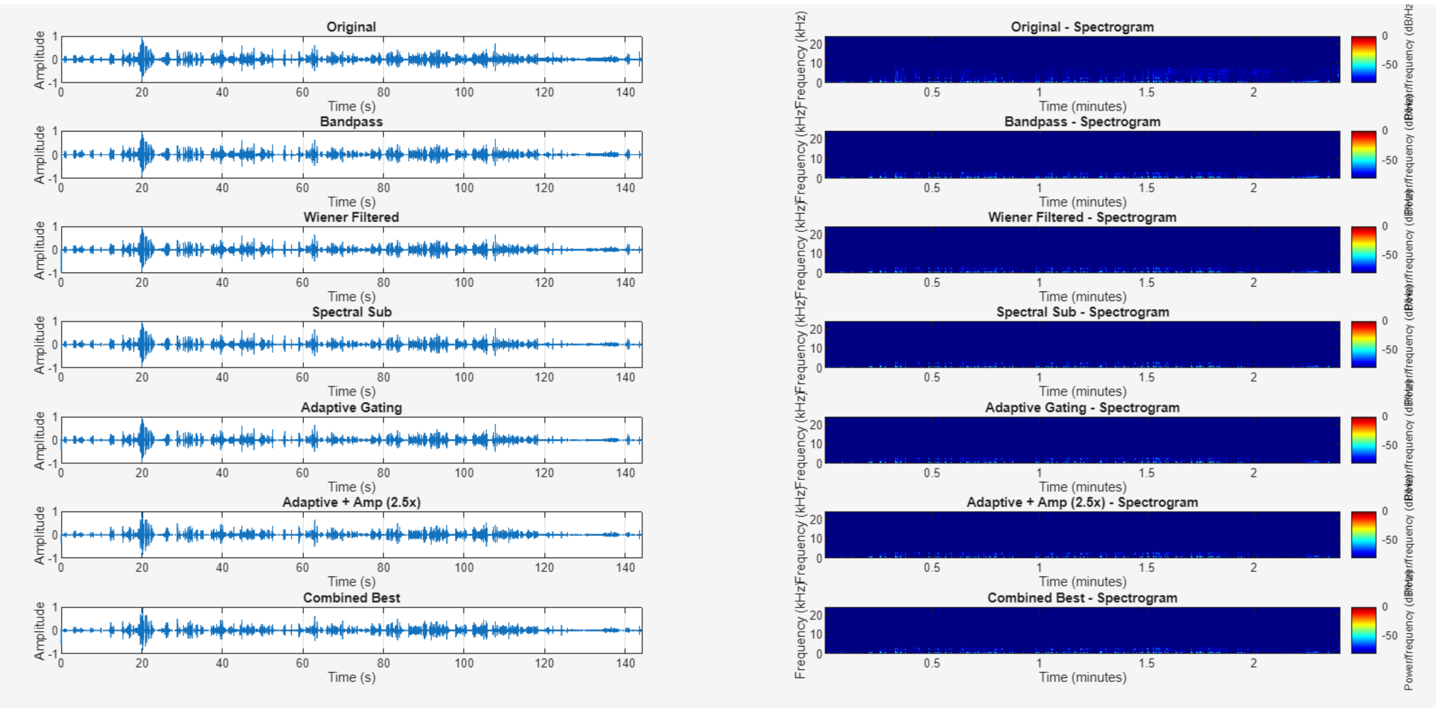


Figure A.3: The waveforms and the spectrogram of all the audio enhancement stages before adaptive amplification

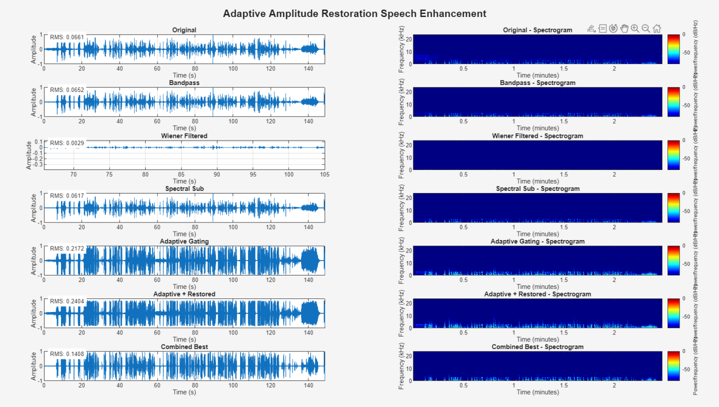


Figure A.4: Waveforms and the spectrograms after adaptive amplification

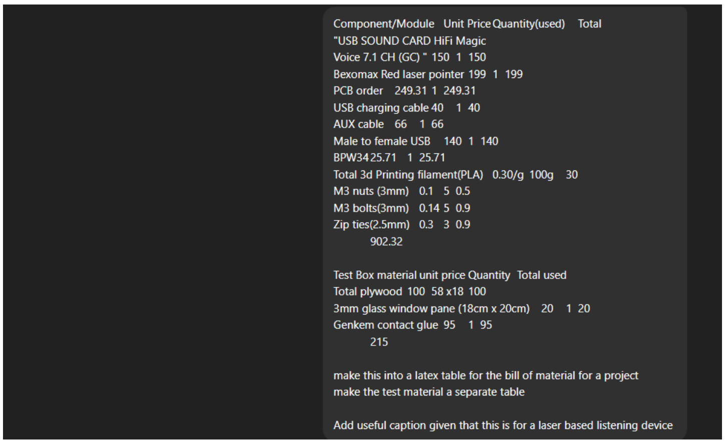
A.7 SNR calculation code

```
%% Method 1: SNR with Clean Reference (SEGMENTED APPROACH)
if has_reference
    fprintf('--- Method 1: SNR Using Clean Reference (Segmented) ---\n');

    for i = 1:n_signals
        sig = signals{i};

        % Match lengths
        min_len = min(length(sig), length(clean_ref));
        sig_trim = sig(1:min_len);
        ref_trim = clean_ref(1:min_len);

        % Voice Activity Detection on reference
        frame_size = round(0.025 * fs);
        hop_size = round(0.010 * fs);
        num_frames = floor((length(ref_trim) - frame_size) / hop_size) + 1;

        voice_mask = zeros(length(ref_trim), 1);

        for j = 1:num_frames
            start_idx = (j-1) * hop_size + 1;
            end_idx = min(start_idx + frame_size - 1, length(ref_trim));

            frame = ref_trim(start_idx:end_idx);
            energy = sum(frame.^2) / length(frame);
            zcr = sum(abs(diff(sign(frame)))) / (2 * length(frame));

            % Voice detection
            if energy > 0.001 && zcr > 0.03 && zcr < 0.25
                voice_mask(start_idx:end_idx) = 1;
            end
        end

        % Identify speech and silence regions
        speech_indices = voice_mask > 0.5;
        silence_indices = voice_mask <= 0.5;

        if sum(speech_indices) > 0 && sum(silence_indices) > 0
            % Gain compensation using speech regions only
            speech_sig = sig_trim(speech_indices);
            speech_ref = ref_trim(speech_indices);

            sig_rms_speech = rms(speech_sig);
            ref_rms_speech = rms(speech_ref);

            if sig_rms_speech > eps && ref_rms_speech > eps
                gain_compensation = ref_rms_speech / sig_rms_speech;
                sig_normalized = sig_trim * gain_compensation;

                % Calculate signal power from speech regions in reference
                signal_pwr = sum(ref_trim(speech_indices).^2);

                % Calculate noise power from silence regions in processed signal
                % (where there should be no speech, only noise)
                noise_pwr = sum(sig_normalized(silence_indices).^2);

                if noise_pwr > eps
                    snr_with_reference(i) = 10 * log10(signal_pwr / noise_pwr);
                else
                    snr_with_reference(i) = Inf;
                end

                fprintf(’  %s: %.2f dB (Speech: %.1f%%, Gain: %.2fx)\n’, ...
                    stage_names{i}, snr_with_reference(i), ...
                    sum(speech_indices)/length(speech_indices)*100, ...
                    gain_compensation);
            else
                snr_with_reference(i) = NaN;
                fprintf(’  %s: N/A (insufficient speech signal)\n’, ...
                    stage_names{i});
            end
        else
            snr_with_reference(i) = NaN;
            fprintf(’  %s: N/A (cannot separate speech/silence)\n’, ...
                stage_names{i});
        end
    end
    fprintf(’\n’);
else
    fprintf(’--- Method 1: Skipped (no reference signal) ---\n\n’);
    snr_with_reference(:) = NaN;
end
```

Listing A.2: SNR Calculation with Clean Reference - Segmented Approach. This method separates speech and silence regions using Voice Activity Detection (VAD) on the reference signal. Signal power is calculated from speech regions in the clean reference, while noise power is measured from silence regions in the processed signal (where only residual noise should exist). The processed signal is gain-compensated to match the reference RMS level before comparison, ensuring that amplitude differences don’t affect the SNR measurement. This approach provides a true measure of residual noise after processing, unlike simple difference-based methods that conflate noise with signal distortion.

A.8 AI prompts and responses

Model -chatgpt



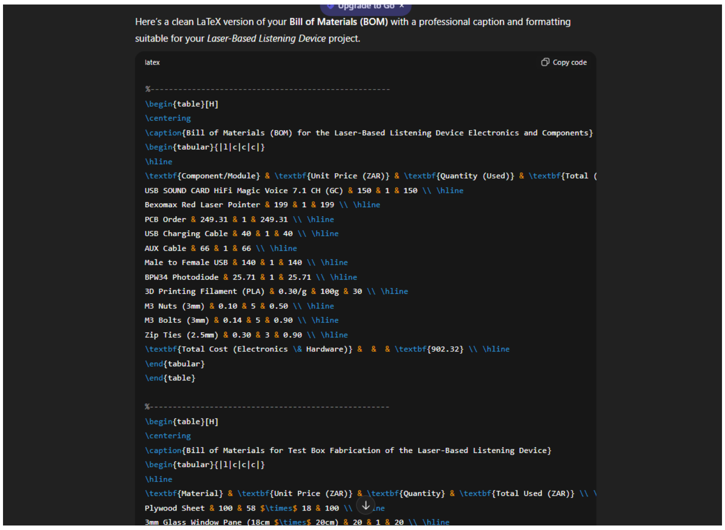
Component/Module Unit Price Quantity(used) Total
"USB SOUND CARD HiFi Magic
Voice 7.1 CH (GC) " 150 1 150
Bexomax Red laser pointer 199 1 199
PCB order 249.31 1 249.31
USB charging cable 40 1 40
AUX cable 66 1 66
Male to female USB 140 1 140
BPW34 25.71 1 25.71
Total 3d Printing filament(PLA) 0.30/g 100g 30
M3 nuts (3mm) 0.1 5 0.5
M3 bolts(3mm) 0.14 5 0.9
Zip ties(2.5mm) 0.3 3 0.9
902.32

Test Box material unit price Quantity Total used
Total plywood 100 58 x18 100
3mm glass window pane (18cm x 20cm) 20 1 20
Genkem contact glue 95 1 95
215

make this into a latex table for the bill of material for a project
make the test material a separate table

Add useful caption given that this is for a laser based listening device

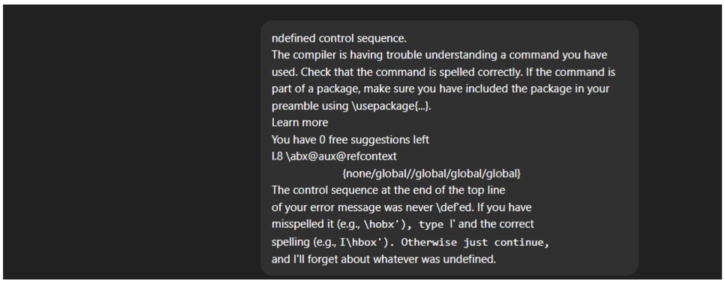
Figure A.5: creating a table



Here's a clean LaTeX version of your **Bill of Materials (BOM)** with a professional caption and formatting suitable for your *Laser-Based Listening Device* project.

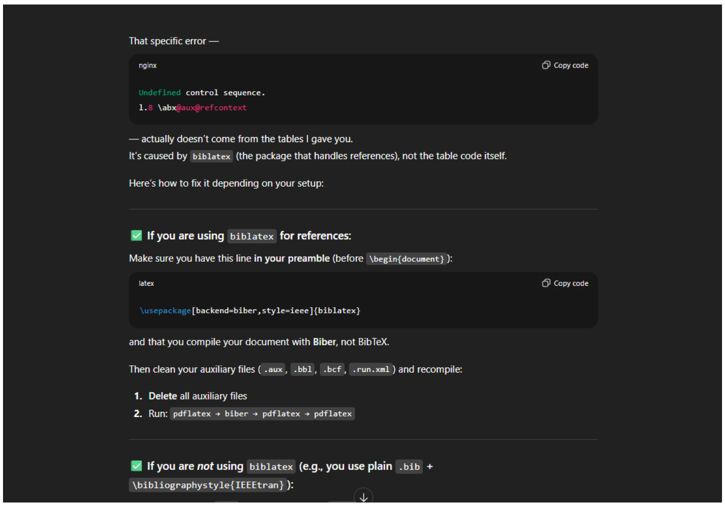
```latex
%--------------------------------------------------
\begin{table}[H]
\centering
\caption{Bill of Materials (BOM) for the Laser-Based Listening Device Electronics and Components}
\begin{tabular}{|l|c|c|c|}
\hline
\textbf{Component/Module} & \textbf{Unit Price (ZAR)} & \textbf{Quantity (Used)} & \textbf{Total (
USB SOUND CARD HiFi Magic Voice 7.1 CH (GC) & 150 & 1 & 150 \\ \hline
Bexomax Red Laser Pointer & 199 & 1 & 199 \\ \hline
PCB Order & 249.31 & 1 & 249.31 \\ \hline
USB Charging Cable & 40 & 1 & 40 \\ \hline
AUX Cable & 66 & 1 & 66 \\ \hline
Male to Female USB & 140 & 1 & 140 \\ \hline
BPW34 Photodiode & 25.71 & 1 & 25.71 \\ \hline
3D Printing Filament (PLA) & 0.30/g & 100g & 30 \\ \hline
M3 Nuts (3mm) & 0.10 & 5 & 0.50 \\ \hline
M3 Bolts (3mm) & 0.14 & 5 & 0.90 \\ \hline
Zip Ties (2.5mm) & 0.30 & 3 & 0.90 \\ \hline
\textbf{Total Cost (Electronics \& Hardware)} &  &  & \textbf{902.32} \\ \hline
\end{tabular}
\end{table}

%--------------------------------------------------
\begin{table}[H]
\centering
\caption{Bill of Materials for Test Box Fabrication of the Laser-Based Listening Device}
\begin{tabular}{|l|c|c|c|}
\hline
\textbf{Material} & \textbf{Unit Price (ZAR)} & \textbf{Quantity} & \textbf{Total Used (ZAR)} \\ \
Plywood Sheet & 100 & 58 $\times$ 18 & 100 \\ \hline
3mm Glass Window Pane (18cm $\times$ 20cm) & 20 & 1 & 20 \\ \hline
```

Figure A.6: Promp1



ndefined control sequence.
The compiler is having trouble understanding a command you have used. Check that the command is spelled correctly. If the command is part of a package, make sure you have included the package in your preamble using \usepackage{...}.
Learn more
You have 0 free suggestions left
l.8 \abx@aux@refcontext
{none/global//global/global}
The control sequence at the end of the top line
of your error message was never \def'ed. If you have
misspelled it (e.g., `\hobx'), type I' and the correct
spelling (e.g., I\hbox'). Otherwise just continue,
and I'll forget about whatever was undefined.

Figure A.7: Promp1



That specific error —

```nginx
Undefined control sequence.
l.8 \abx@aux@refcontext
```

— actually doesn't come from the tables I gave you.
It's caused by `biblatex` (the package that handles references), not the table code itself.

Here's how to fix it depending on your setup:

✅ **If you are using `biblatex` for references:**

Make sure you have this line **in your preamble** (before `\begin{document}`):

```latex
\usepackage[backend=biber,style=ieee]{biblatex}
```

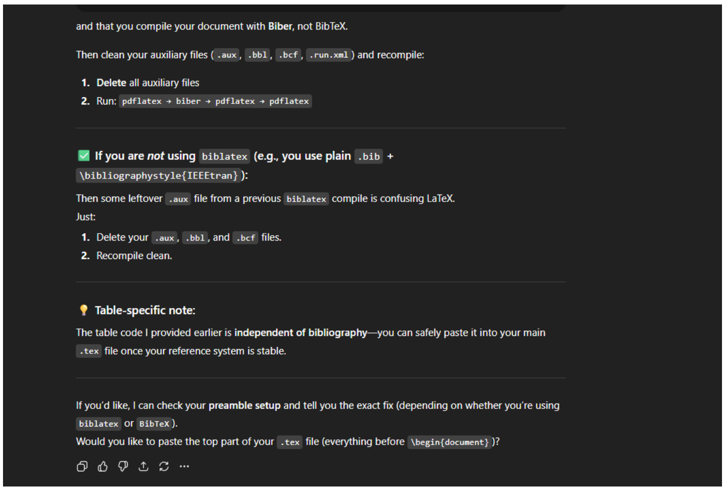
and that you compile your document with **Biber**, not BibTeX.

Then clean your auxiliary files (`.aux`, `.bbl`, `.bcf`, `.run.xml`) and recompile:

1. **Delete** all auxiliary files
2. Run: `pdflatex → biber → pdflatex → pdflatex`

✅ **If you are *not* using `biblatex` (e.g., you use plain `.bib` + `\bibliographystyle{IEEEtran}`):**

Figure A.8: Promp3



and that you compile your document with **Biber**, not BibTeX.

Then clean your auxiliary files (`.aux`, `.bbl`, `.bcf`, `.run.xml`) and recompile:

1. **Delete** all auxiliary files
2. Run: `pdflatex → biber → pdflatex → pdflatex`

✅ **If you are *not* using `biblatex` (e.g., you use plain `.bib` + `\bibliographystyle{IEEEtran}`):**

Then some leftover `.aux` file from a previous `biblatex` compile is confusing LaTeX.
Just:

1. Delete your `.aux`, `.bbl`, and `.bcf` files.
2. Recompile clean.

💡 **Table-specific note:**

The table code I provided earlier is **independent of bibliography**—you can safely paste it into your main `.tex` file once your reference system is stable.

If you'd like, I can check your **preamble setup** and tell you the exact fix (depending on whether you're using `biblatex` or `BibTeX`).
Would you like to paste the top part of your `.tex` file (everything before `\begin{document}`)?

Figure A.9: Prompt2 response