Capstone Term 8

Capstone 5 Project S21

D'Noise

Multimodal Speech Capturing Device

Spring 2025

Authors	Student ID
Abel Lee Yang Yeow	1006085
Caitlin Daphne Tan Chiang	1006537
Jone Chong Jin	1006338
Loh Jianyang John	1006360
Leong Wen Jie Lucas	1003418
Ong Jing Ting	1003573
Wang Jun Long Ryan	1005923

Acknowledgements

The team would like to express sincere gratitude to the many individuals whose support and guidance have been instrumental throughout the Capstone project. Their contributions have been invaluable in helping the team progress and achieve its goals to date.

The team is especially thankful to **Professor Tan Mei Chee** and **Professor Zhao Fang**, the team's SUTD Capstone Mentors, for their consistent and constructive feedback during weekly updates and presentations throughout Term 7 and 8. Both mentors also ensured the team was well-supported administratively, verifying the accuracy of deliverables and promptly disseminating essential information to keep the team aligned with project requirements. The team also appreciates the ongoing support of the Capstone Office, whose resources and assistance proved helpful whenever needed.

The team's gratitude extends to their industry partner, KLASS Engineering and Solutions Pte. Ltd. In particular, **Lu Zheng Hao**, the Project Manager, for coordinating the meetings and offering valuable feedback. The team also thanks their industry mentors—**Nicholas Chan, Terence Goh, and Syed Asif**—for their insightful technical advice, which helped them address and resolve several technical challenges encountered in this Capstone project.

The team would also like to acknowledge **Professor Park Jihong**, an SUTD Capstone Mentor, for his insightful feedback during the interim and final reviews. Special thanks go to **Ms. Belinda Seet** from the Center for Writing and Rhetoric (CWR) for her guidance in preparing for the team's presentations and the various documents, including this report. Additionally, the team is grateful to **Professor Teo Tee Hui, Professor Joel Yang**, and **Professor Kwan Wei Lek** for their technical guidance, which has been critical to the development of the team's project.

Each of these individuals has played a meaningful role in the team's journey, and the team is truly thankful for their time, expertise, and commitment.

Executive Summary

This report outlines the development of D'Noise, a multi-modal speech-capturing device designed to reduce the risk of missing out critical information from speech, particularly for security personnel operating in noisy environments such as airports and malls. Recognizing that current commercial solutions rely heavily on cloud-based processing and lack robust noise suppression, D'Noise was conceived as a portable, on-device system capable of audio enhancement with noise suppression, playback, and real-time transcription.

To achieve this, the team established three important design specifications – playback an enhanced speech of the target with noise suppression, on-device computation, and real-time transcription capabilities in a discreet prototype. To achieve these specifications, measurable hardware and software objectives were established. Hardware objectives included capturing speech from within conversational distances (3 meters), maintaining a discreet and lightweight form factor under 1.25kg, ensuring at least three hours of continuous battery life, and maximizing comfort during extended wear. Software objectives targeted playback latency below 200 milliseconds, a perceptible improvement in Signal-to-Noise Ratio (SNR), transcription accuracy with Word Error Rate (WER) below 30%, and full support for on-device processing without reliance on internet connectivity. These metrics guided the technical exploration and prototype development documented in the Methodology section.

The system design involved three core components: audio capture, audio processing, and audio transcription. In the development process, the audio capture module, various microphone configurations and options were weighed to optimize for further speech capturing distances. A shotgun microphone was selected for its directionality and far reach. In audio processing, both static filters and deep learning-based denoisers were tested. RNNoise was selected for its low-latency performance on CPU, enabling real-time speech enhancement while conserving computational resources. For transcription, Whisper was fine-tuned on over 500 hours of Singlish data from the IMDA National Speech Corpus, significantly reducing WER and enabling near real-time text output on a custom handheld device.

Quantitative and qualitative evaluations documented in the Discussion demonstrated that D'Noise successfully fulfilled all objectives. In noisy environments, the system maintained intelligible playback beyond conversation distances – up to 4.5 meters, and 6 meters in quiet conditions. Battery stress tests confirmed a minimum runtime of 4 hours, while temperatures of the satchel housing D'Noise remained below 35.7°C. The system weighed a total of 1.186kg, thus meeting the sub-1.25kg form factor specifications, which is comparable to the lightest laptops in the industry. From a software standpoint, the average playback latency for denoising is just 0.755ms, with significant gains in speech intelligibility that can handle both static and dynamic noises. Whisper, after fine-tuning, achieved an average WER of 15.2% (in $\sim$72dB ambient noise) with an average latency of 0.33ms, achieving near real-time operation in noisy environments. All processes are entirely on-device using the Jetson Orin Nano, making D'Noise independent of cloud services and their recurring costs, which distinguishes this prototype from current solutions.

Ultimately, D'Noise addresses critical gaps of existing commercial solutions by delivering an on-device, low-latency system designed specifically for local operational requirements. D'Noise can suppress dynamic background noise while maintaining real-time transcription performance, establishing a new benchmark for wearable speech-capturing devices in security applications, offering capabilities not currently available in any existing market offerings.

Table of Contents

1 Introduction

In noisy environments—such as airports, shopping malls, and condominiums—security personnel and staff consistently encounter significant challenges in maintaining clear communication (Alnuman & Altaweel, 2020). This is due to adverse conditions (such as unclear, disfluent or environmental noise) that can degrade the intelligibility of the target speech (Mattys et al., 2012). Thereby, increasing the risk of losing critical information during conversations, which would delay prompt and effective responses in dangerous situations (Keller et al., 2017). These conditions underscore the need for a device that can reliably capture and enhance speech in noisy settings.

Problem Statement How might we support security personnel in Singapore by developing a speech-capturing device capable of enhancing, playing back, and transcribing audio in real time within noisy environments, thereby reducing the risk of missing critical information?

Design Specifications To address this problem, the design of the prototype should be able to

- Playback an enhanced speech of the target and suppress noise (improve SNR).
- Offer near real-time transcription with low word error rates and display it.
- Support on-device computation in a discreet and portable prototype.

Objectives	**Measurable Metrics**
Hardware	
(I) Effective Operational Range	Capture speech beyond typical conversational distances (3m radius) in noisy environments.
(II) Discreet and Compact Design	Unobtrusive form factor within 1.25kg.
(III) Extended Battery Life	Continuous operation for at least 3 hours[1].
(IV) Comfort and Wearability	Comfortable throughout 3 hours[1].
Software	
(I) Minimal Latency	Enhanced and denoised audio playback latency within 200ms[2].
(II) Enhanced Audio Playback	Improve noisy speech for clearer speech playback.
(III) Near Real-Time Transcription	Transcription with a Word Error Rate within 30% and latency within 1s.
(IV) On-Device Computation	Perform all processes on-device.

Table 1: Summary of hardware and software objectives with measurable outcomes to fulfill the design specifications.

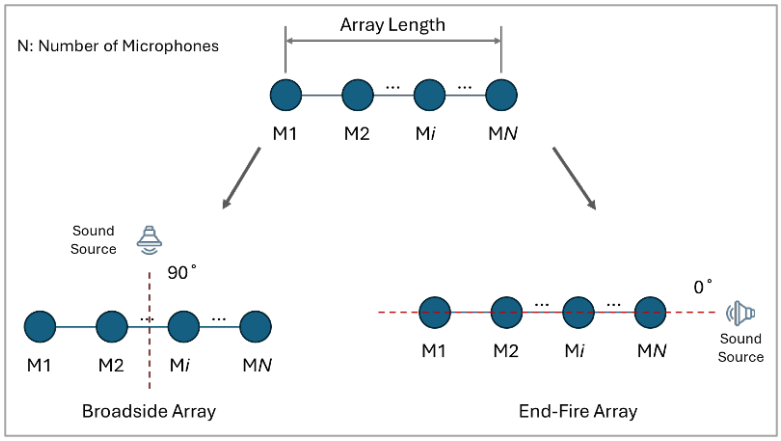
[1]The average number of hours before a security personnel takes a break as specified by the industry mentor.
[2]The acceptable latency range for noise suppression within industry standards (Y. Yang et al., 2023).

2 Background

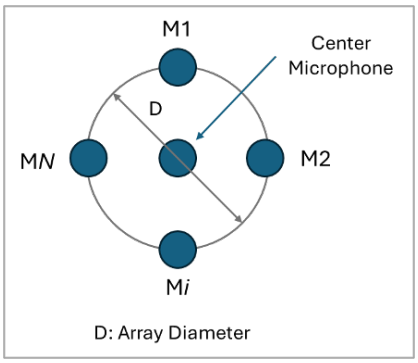
2.1 Hardware Literature Review

Motivated by the hardware objectives (I – IV), as outlined in the introduction (Section 1), the team considered hardware components to enable the software modules whilst keeping the prototype discreet and comfortable. The components include microphone technologies, processing units, and power solutions, and their detailed comparisons are documented in Appendix B.

2.1.1 Microphones



(a) Visualization for the Uniform Linear Array (ULA) microphones



(b) Visualization for the Uniform Circular Array (UCA) microphones

Figure 1: Side-by-side comparison of ULA and UCA microphones.

Directional microphones (e.g., shotgun, electret condenser) offer high sensitivity and noise isolation (Popescu, 2023), but their larger form factors often hinder compact and wearable designs. Alternatively, Micro Electro-Mechanical System (MEMS) microphones are small in size, have low power consumption, and support microphone arrays. These arrays enable advanced signal processing such as beamforming and noise suppression, enhancing speech clarity in noisy settings.

2.1.2 Processing Units

Local processing of speech signals, including tasks involving Large Language Models (LLMs) or Automatic Speech Recognition (ASR), demands significant computational resources (Georgescu et al., 2021). Single-board computers like the NVIDIA Jetson Nano or Orin Nano provide a balance of compactness, energy efficiency, and hardware accelerators, making them well suited for on-device speech processing (NVIDIA, n.d.).

2.1.3 Power Solutions

Wearable devices require batteries that offer high energy density to minimize both size and weight while ensuring adequate runtime. Lithium-ion (Li-Ion) and lithium-polymer (Li-Po)

batteries are commonly employed for this purpose (Manthiram, 2017; Townsend et al., 2020). That said, they necessitate robust battery management systems to maintain safety standards and prolong the overall battery lifespan.

2.2 Software Literature Review

Motivated by the software objectives (I – IV), as outlined in the introduction (Section 1), the team researched various software techniques to process the audio from its initial input to transcribing into natural language (text).

2.2.1 Audio Capture

Direction of Arrival (DOA) Algorithms	
GCC-PHAT (Knapp & Carter, 1976)	Estimates the Time Difference of Arrival (TDOA) between microphone signals by applying a phase transform to normalize magnitudes, then computes cross-correlation. It is computationally efficient but prone to reduced accuracy in reverberant settings.
MUSIC (Stoica & Nehorai, 1989)	Decomposes the signal covariance matrix into signal and noise subspaces. Peaks in the noise subspace indicate the DOA. This method provides high resolution for closely spaced sources but demands significant computational resources and is sensitive to modeling errors.
Beamforming (BF) Algorithms	
MVDR (Guo et al., 2012)	Minimizes power from all non-target directions while preserving a distortionless response for the desired signal. It requires accurate covariance matrix estimation and may become sensitive to rapid noise-level fluctuations.
GSC (Moonen & Proudler, 2000)	Extends MVDR by adding a noise cancellation module to handle sidelobe interference. Its modular design offers flexibility in dynamic environments, yet it can suffer from signal cancellation if the blocking matrix does not perfectly eliminate the target speech.

Table 2: Summary of Literature Review on DOA and Beamforming Algorithms

From the initial audio input stage, the team considered Direction of Arrival (DOA) estimation algorithms, which are critical for identifying the direction of incoming sound sources, especially in multi-source environments. Accurate determination of the target speech angle allows the system to effectively suppress interference originating from other directions. Two promising DOA algorithms explored are Generalized Cross-Correlation with Phase Transform (GCC-PHAT) and Multiple Signal Classification (MUSIC). Following DOA estimation, Beamforming (BF) algorithms were also considered to enhance the target speech signal by selectively filtering signals from the desired direction while attenuating interference from others. The beamforming algorithms assessed include a recent implementation of Minimum Variance Distortionless Response (MVDR) and the Generalized Sidelobe Canceller (GSC).

2.2.2 Noise Suppression (NS)

Static Methods	
Low-Pass Filter	Passes lower-frequency components and suppresses higher frequencies, which reduces high-pitched noise while retaining most of the speech signal.
Wiener Filter	Minimizes the mean squared error between the noisy input and the estimated clean output through the noise and signal spectrum.
Median Filter	Replaces each sample with the median of its neighboring values, which reduces impulse noises.
Spectral Gating (Y. Yang et al., 2023)	Operates in the frequency domain by applying a threshold to the signal's power spectrum, attenuating noise components below the threshold while preserving dominant frequencies.
Deep Learning Methods	
RNNoise (Valin, 2018)	A lightweight RNN-based model designed for real-time applications. It estimates ideal band gains and incorporates pitch filtering for harmonics, requiring only 40 Mflops—making it well-suited for deployment on embedded devices like the Jetson Orin Nano (NVIDIA, n.d.).
Facebook Denoiser (Defossez et al., 2020)	A full-band real-time speech enhancement model that operates directly on waveform input. It leverages convolutional encoder-decoder architectures with skip connections and has demonstrated strong performance on real-world noise.
DeepFilterNet (Schröter et al., 2022)	A compact neural network designed for speech denoising with a focus on low-latency processing. It uses temporal convolutional layers and can run efficiently on both CPU and GPU, making it suitable for resource-constrained real-time systems.
Neural Beamforming (Gu et al., 2023)	Uses neural networks to enhance speech from multi-channel audio from microphone arrays by converting signals into time-frequency representations and estimating spatial filters.
Guided Speech Enhancement (GSE) Network (Y. Yang et al., 2023)	Combines classical beamforming with a deep neural network by taking both the raw input and the beamformed output as inputs, allowing the model to learn and further suppress the residual noise from beamforming.

Table 3: Summary of Noise Suppression Algorithms from Literature

To suppress noise, deep learning and algorithmic approaches were explored as an additional processing stage after initial processing (Section 3.1). Table 3 summarizes the literature review of the methods explored in the design-thinking process. Given that `D'Noise` is meant to be portable, noise suppression methods have to be robust and dynamic. The static methods reviewed specialize in static noises (such as white noise), which made them insufficient to deal with dynamic noises (such as chatter) in realistic environments. Thus, the team considered deep learning methods to circumvent this limitation of the static methods.

2.2.3 Audio Speech Recognition (ASR)

ASR Models	
Kaldi (Povey et al., n.d.)	A widely adopted speech recognition toolkit known for its high accuracy and flexible architecture supporting deep neural networks. It offers comprehensive tools for feature extraction and acoustic modeling. However, it is complex to configure and resource-intensive, making it less suited for real-time or on-device use.
Wav2Vec 2.0 (Baevski et al., 2020)	A self-supervised model developed by Facebook AI Research that learns representations directly from raw audio. It achieves strong results in low-resource conditions and supports fine-tuning for domain-specific applications. Its high inference cost, however, limits its practical use in low-latency or embedded systems.
Whisper (Radford et al., 2022)	OpenAI's robust end-to-end model trained on a diverse corpus for multilingual and noisy speech recognition. It is highly resilient to background noise and accent variability, aligning well with real-world deployment scenarios. The trade-off lies in its model size, which still presents challenges for efficient real-time execution.
Fine-Tuning Methods (Sanchit, 2022)	
Targeted Layer Training	Only decoder and projection layers are updated, while the encoder remains frozen. This reduces training time and mitigates catastrophic forgetting, allowing efficient domain adaptation with limited data.
Whisper Normalizer	A text pre-processing step that standardizes punctuation, removes hesitations, and maps numbers to their spoken forms. This helps align reference transcriptions with Whisper's decoding targets during training.
Timestamp Supervision	Fine-tuning on datasets with precise word-level or segment-level timestamps enables better alignment and segmentation, improving diarization and real-time streaming performance.
Data Sampling and Resampling	Ensures class balance across training steps by over-sampling underrepresented classes or domains. This prevents overfitting to frequent speakers or acoustic conditions in small corpora.

Table 4: Summary of the Literature Review for Speech Recognition Models and Fine-Tuning Methods

After processing the audio, the team explored various open-source ASR solutions to address the real-time transcription requirement (Objective (III)). Existing methods are summarized in Table 4. To further tailor these ASR models for Singaporean security personnel, an important technique to consider would be fine-tuning. Fine-tuning improves transcription accuracy by adapting pre-trained models to task-specific data distributions—particularly important when facing domain shifts such as local English accents (Singlish), background noise, or spontaneous speech. Key techniques used in fine-tuning are also summarized in Table 4.

2.3 Market Research

Product	TranscribeGlass	Pocketalk	Sudio T2
Visualization			
Description	A wearable device that attaches to glasses, providing real-time captions (TranscribeGlass, n.d.).	A handheld two-way voice translator for real-time translation (Pocketalk, n.d.).	True wireless earbuds that utilize beamforming for noise suppression (Sudio, n.d.).
Advantages	Affordable alternative to expensive AR devices; lightweight and comfortable design.	Accurate translations; large intuitive touchscreen; noise-canceling microphones	Comfortable and customized fit; drivers for clear audio; quick charging.
Limitations	Relies on external speech-to-text apps; may not be suitable for all types of hearing loss; potential latency issues in transcription.	Requires internet connection for translations; struggles with colloquial language and regional dialects; relatively high cost.	Active noise cancellation is average; touch controls can be less intuitive than physical buttons; lacks support for advanced audio codecs.

Table 5: Market Research of Competing Products

Existing market products primarily comprise real-time transcription glasses, handheld transcription devices, and directional earbuds. Although each offers distinct advantages, they share common shortcomings, including reliance on cloud-based processing, inadequate noise suppression, and limited adaptability to noisy environments.

Reliance on Cloud-Based Processing: Many devices depend on cloud services for speech recognition and processing, leading to ongoing subscription costs and security expenses for encryption and data protection. Moreover, security officers often operate in locations with inconsistent or unreliable internet access, making cloud-based solutions impractical in the field.

Inadequate Noise Suppression and Environmental Adaptability: Existing products often lack advanced noise suppression to handle environmental noise or cross-talk, impeding clear audio capture. They also fail to adapt to direction-specific speech requirements, such as focusing on a single speaker in a crowded space. This insufficient directional sensitivity compromises the reliability and clarity of recorded audio in demanding operational settings.

3 Methodology

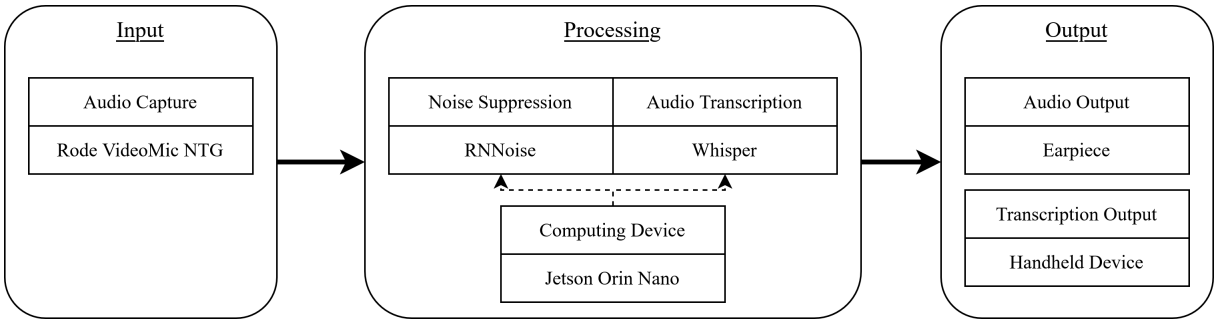


Figure 2: Overview of the workflow of `D'Noise`.

The overall system (Figure 2) is segmented into three main components - audio capture, processing, and transcription. The audio is captured through the input device and processed by the denoising module and transcription model in parallel. The denoised audio is then played back and the transcription displayed. This section details the design thinking and experimental process before the team arrived at their final prototype. The resultant design of the prototype would be reflected in the Discussion (Section 4).

3.1 Audio Capture

The audio capturing component of the prototype is responsible for taking in raw input waveforms with a focus on hardware objectives (I) and (II) - the device's operational range and a discreet and compact design.

3.1.1 Standard Microphones

MEMS microphones are miniature devices that integrate mechanical elements and electronic circuitry on a single silicon chip, offering high reliability, low power consumption, and consistent performance over a broad frequency range (typically 20 Hz–20 kHz). Their compact form factor makes them ideal for array-based beamforming applications, and their operational range in practical environments can extend from a few centimeters to several meters, depending on the array design, ambient noise levels, and signal processing strategies.

3.1.2 Microphone Arrays

MATLAB simulations were conducted to determine the microphone's suitability for the team's prototype. The team considered the beam patterns and Signal-to-Noise Ratios (SNR) in the simulations and their impact on the prototype's operational distance. This also involved implementing a Delay-and-Sum beamforming technique to enhance sound from specific directions while suppressing noise.

Microphone Array Structures	
Uniform Linear Array (ULA) (Figure 1a)	Comparisons of Broadside and End-Fire configurations (Figure 12) showed that End-Fire provides a single dominant main lobe and fewer side lobes, resulting in superior directional focus. Increasing the number of microphones within a fixed array length further sharpened the main lobe, thereby improving noise suppression. Optimizing array length also helped reduce redundant noise (Appendix A, Figure 13 and 14).
Uniform Circular Array (UCA) (Figure 1b)	Simulations varying microphone count and diameter (Appendix A, Figure 15 and 16) indicated that an increased diameter significantly improves the beam pattern, creating a more focused cardioid shape. This makes UCAs well-suited for noisy environments and demonstrates robustness even with fewer microphones, offering flexibility in design and addressing market constraints.

Table 6: Summary of Observations on Microphone Array Structures

Operational Range: Simulations (Appendix A, Figure 17) were performed by plotting SNR against distance under typical noise levels (approximately 90 dBA). ULA performance varied with microphone count and array length, showing shorter operational distances for smaller arrays. By contrast, UCAs delivered a superior range as the number of microphones increased, making it a strong candidate for the microphone in `D'Noise`.

3.1.3 Shotgun Microphones

Shotgun microphones, also known as interference tube microphones, are highly directional devices designed to capture sound from a narrow forward-facing region while minimizing off-axis noise. Their long, slotted tube shape uses phase cancellation to attenuate sound arriving from the sides, making them well-suited for applications that require precise audio capture over distances, which is a very important design specification in this `D'Noise`. While shotgun microphones excel at isolating the desired source in environments with significant ambient noise, they may still exhibit reduced performance when sound sources arrive from multiple directions or when there is substantial room reverberation. In light of these tradeoffs, an extensive test with the other components of `D'Noise` is detailed in the Discussion (Section 4.1) to more concretely ascertain the viability of Shotgun Microphones.

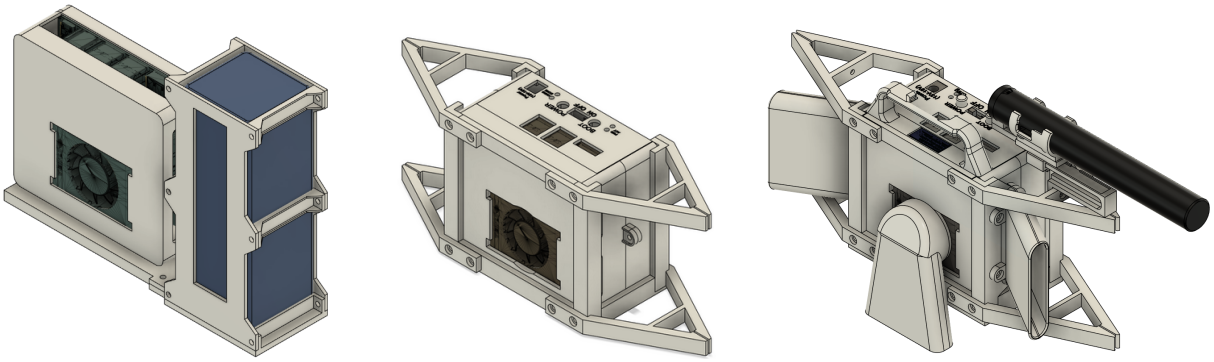
3.2 Audio Processing

3.2.1 Single Board Computers

Computer: In order to handle the computational demands of the denoising and transcription modules, various single-board computers (SBCs) were evaluated based on computational performance, power efficiency, size, and cost (Appendix B, Figure 19). The Jetson Orin Nano was the strongest contender for its performance-to-power ratio to enable low-latency speech processing and transcription while keeping power consumption moderate. Utilizing this edge

device for the prototype's processing requirements would **eliminate the dependence on cloud services** and maintain operational capabilities without an internet connection – an advantage over other market products that is quintessential for security guards who may not have stable internet connections.

Power System: To fuel the Jetson Orin Nano, a 4-series Li-Ion battery pack (14.8 V, 7000 mAh) was selected and regulated down to 12 V. This configuration runs the device for approximately six hours (half a typical 12-hour security shift in Singapore) to balance between operational requirements and manageable weight. Li-Ion batteries were preferred over LiPo (Appendix B, Figure 21) due to their robustness and safety under mechanical stress (Detailed calculations in Appendix B, Figure 20).



(a) First Iteration (b) Second Iteration (c) Third Iteration (Final Design)

Figure 3: Design iterations of the computing hardware for audio processing and transcription.

Design Iterations: The development of housing the computing hardware (SBC and the power system) underwent three major iterations, with each version addressing critical limitations identified during testing and user feedback. Figure 3 above illustrates this progression.

Iteration 1 (Figure 3a) prioritized enabling local speech denoising and transcription using the Jetson Orin Nano due to its favorable performance-to-power ratio and compact form factor (176 g). The Jetson was initially powered by a 7500 mAh 4S Li-Po battery (590 g), regulated to 12 V using a buck converter. While this configuration achieved 6 hours of continuous operation—half of a typical security guard's shift—it exhibited critical drawbacks. Battery swapping was cumbersome, the Li-Po cell posed mechanical safety concerns, and the 766 g system (excluding bag and frame) resulted in poor weight distribution, limiting long-term wearability.

Iteration 2 (Figure 3b) addressed the power-safety and modularity concerns of Iteration 1. The Li-Po battery was replaced with three 21700 Li-Ion cells, housed in steel casings for improved durability and intuitive hot-swapping. The team integrated the Waveshare Uninterruptible Power Supply (UPS) for Jetson Orin Nano, which allowed seamless switching between AC and battery power and ensured overcharge/discharge protection. A new internal frame was designed to accommodate the increased thickness of the Jetson–UPS–battery assembly. This reduced the weight of the device and improved the modularity of the design. Yet, there were thermal concerns for this device as it would be stored in a satchel bag (Appendix E, Figure 34) for discreetness.

Iteration 3 (Figure 3c) naturally focused on the thermal concerns from iteration 2 and further improved user experience. Passive thermal management was implemented by creating air intake and exhaust channels within the satchel zippers and seams. The hardware module

was redesigned to be removable as a single slide-out unit to enable easy battery replacement and debugging. A front-facing interface panel was introduced with labeled ports and buttons, and an OLED display was added to show live power metrics. This iteration formed the final satchel-based prototype, satisfying desirable traits like portability, low-latency operation, and discreetness in the design specifications for Singaporean Security Personnel.

3.2.2 Noise Suppression and Speech Enhancement

With the computing hardware from the previous section, the team explored possible denoising methods to fulfill the software objectives (Table 1) to denoise and playback enhanced speech.

Method	**ΔSNR (dB)**	**PESQ**	**ΔSTOI**	**Latency (ms)**
Static Methods				
Wiener Filter	+0.18	1.129	+0.001	0.301
Median Filter	+0.20	1.115	-0.005	0.096
Low Pass Filter	+0.12	1.167	+0.000	0.486
Spectral Gating (Sainburg et al., 2020)	+1.26	1.095	+0.018	44.104
Deep Learning Models				
RNNoise (CPU) (Valin, 2018)	+0.90	**1.224**	-0.114	**0.755**
Facebook Denoiser (CPU) (Defossez et al., 2020)	**+1.42**	1.044	-0.172	155.062
Facebook Denoiser (GPU)	**+1.42**	1.044	-0.172	65.387
DeepFilterNet (CPU) (Schröter et al., 2022)	+1.41	1.075	-0.033	76.457
DeepFilterNet (GPU)	+1.41	1.075	-0.033	52.843

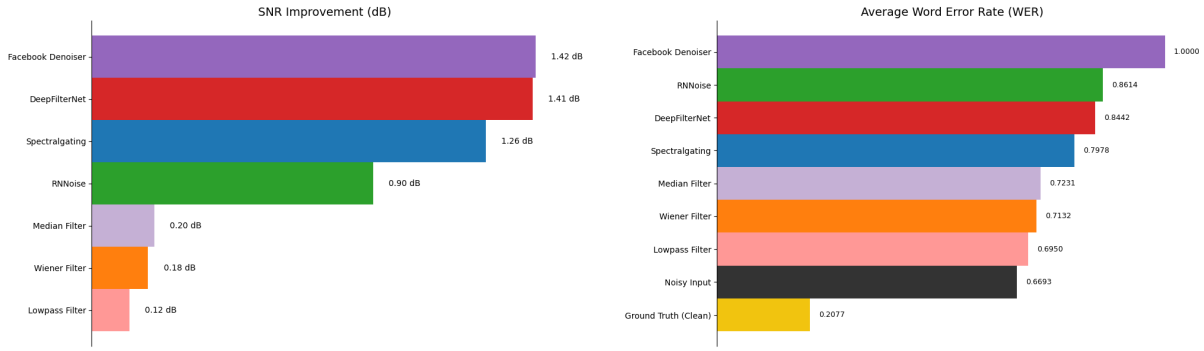
Table 7: Comparison of speech denoising methods using change in SNR, PESQ, ΔSTOI, and average latency per 10 ms chunk. Each metric is the average computed over the test set.

As a baseline, the team compared various static (filter-based) algorithms. The results are summarized in Table 7, where the deep learning models significantly outperformed the static methods in SNR and PESQ. As the test set consists of conversational local English from IMDA's National Speech Corpus (IMDA, 2019), it posed significant challenges for the filter-based methods to adapt to a dynamic range of noises. In view of this limitation, deep learning methods were considered for a more robust and effective solution to suppress noise.

Deep Learning Methods: To overcome the limitations of static methods, the team evaluated several deep learning models capable of adapting to dynamic noise conditions. Among the models tested—*RNNoise*, *Facebook Denoiser*, and *DeepFilterNet*—trade-offs were observed

between signal quality, perceptual clarity, and processing latency. *Facebook Denoiser* offered the highest noise suppression but incurred significant latency, while *RNNoise* provided the best balance of perceptual quality and real-time performance. For this reason and its low computational demands, *RNNoise* was selected for the prototype. Table 7 details the in-depth comparisons.

Deep Learning with Microphone Arrays: Given the promising theoretical effectiveness of microphone arrays (Appendix A), considering models that leverage on these microphones was an intuitive step towards a cohesive prototype design. *Neural Beamforming* and *GSE Network* are two models that utilized the microphone arrays. However, the use of unique hardware also meant a lack of realistic data to train these models. Hence, the team initiated an extensive data collection effort (Appendix D.1) with the microphone arrays to overcome this challenge. The training process is detailed in Appendix D.



(a) Average SNR improvements of different denoising methods.

(b) Average WER of ASR outputs for different denoising methods (Whisper-small).

Figure 4: Comparison of denoising performance based on (left) Word Error Rate (WER) and (right) Signal-to-Noise Ratio (SNR). Full sized figures in Appendix D.3.2.

Limitations: Despite the efforts in integrating deep learning models that utilize microphone arrays into the system, the team found that beamforming-based approaches **did not yield noticeably better results** compared to models using single-channel inputs. Additionally, the team discovered that SNR improvements as a metric were not informative, as a higher value does not translate into better intelligibility. For instance, the Facebook Denoiser had the highest SNR improvement, but the output was garbled noise. In an attempt to address this, the team considered WER (Figure 27) by an ASR model as an alternative metric for comparison. Surprisingly, the WER for the denoising methods were higher than the noisy inputs. The team reasoned that this result lies in issues depending on ASR models for evaluation (Appendix D.3.2). Finally, to reconcile with these results, the team conducted human evaluation (Section 4.1) to investigate the denoising capabilities (considering distance and intelligibility of the outputs) of the various models. This resulted in RNNoise being chosen for its high perceptual quality with low latency and hardware requirements – only requiring a CPU for its computation, which frees up GPU resources for the transcriptions.

3.3 Audio Transcription

After processing the audio (Section 3.2), the resultant enhanced speech is passed to an Automatic Speech Recognition (ASR) model for transcription. It is then displayed on a commu-

nication device to achieve near real-time transcription (Software Objective IV). In this section, the team documents in detail their design thinking process towards the final prototype.

3.3.1 Display Design

Throughout the iterative design process, the team continuously revisited the project requirements to ensure alignment with user needs and system objectives. Regular consultations with industry mentors also provided valuable insights into the optimal form factor for practical deployment. Ultimately, a handheld device—shown in Figure 7 and conceptually similar to the PocketTalk identified in the team's market research (Table 5)—was selected. This form factor was determined to be the most viable, effectively meeting all defined functional and non-functional requirements.

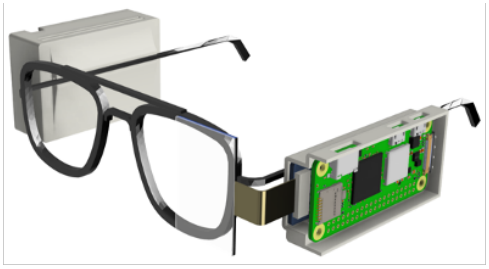


Figure 5: Iteration 1 of the Display Device

Iteration 1: The team designed a glasses-based prototype (Figure 5) as inspired by the *Transcribe-Glass* in the team's market research (Table 5). This design was intended to make it convenient for the users to read the transcribed text without the need to alternate their gaze between a separate hand-held screen and the target speaker. In the initial testing of this design, the glasses felt uncomfortable and heavy with a bulky mounting structure that impeded peripheral vision. Further, this iteration was **not discreet**, which was a key requirement from the industry partners.

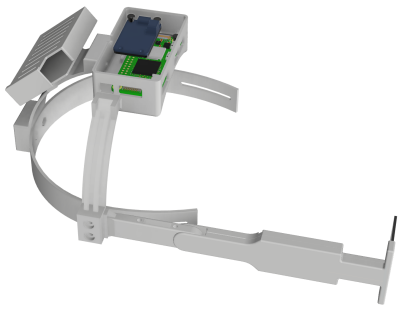


Figure 6: Iteration 2 of the Display Device

Iteration 2: Following feedback and initial ergonomic testing of Iteration 1, the team proceeded to develop a headset-based design aimed at improving weight distribution for enhanced comfort. To maintain optimal microphone performance, the frame was enclosed within an acoustically transparent cap. However, to position the transparent OLED display at a readable distance for the user, the screen had to extend outward from the cap. This significantly compromised the discretion of the device, which conflicted with one of the key non-functional requirements. As such, the headset design, while more ergonomic, was ultimately deemed unsuitable due to its high visual profile.

Figure 7: Iteration 3 of the Display Device

Iteration 3: The team arrived at their current design - a handheld device. It features a compact form factor, is lightweight and discreet, and small enough to fit in the user's palm or a pants pocket. The transcribed text appears on an LCD screen and has a battery life of approximately 4.5 hours. Charging the device is straightforward, thanks to a standard USB-C charging port, allowing easy implementation into existing setups. It is also user-friendly, with zero learning curve required. Features such as the simple on/off switch, battery percentage display, contrast adjustment, and built-in troubleshooting ports further enhance its usability.

Limitations of Handheld Device: While the handheld device successfully fulfilled all defined project requirements, several limitations remain in terms of usability and discretion. As the device must be held in the user's hand, it requires the user to look away from their surroundings to view the transcription, potentially disrupting situational awareness during critical interactions. In addition, the current design is still out of place when compared to a mobile phone. Furthermore, the existing screen size limits the amount of information displayed at once. Increasing the display size could improve readability and allow users to view more transcribed content without scrolling. A potential improved version of this product based on these limitations and the user validation survey (Section 4.4) is detailed in this report's Future Work (Section 4.5).

3.3.2 Audio Speech Recognition (ASR) Models

To transcribe audio to text, the team evaluated three lightweight, offline-capable ASR models—Kaldi, wav2vec 2.0, and Whisper. Table 8 compares the Normalized Word Error Rate (WER) across five different datasets, demonstrating that **Whisper significantly outperformed the other models**. Based on these results, Whisper was selected for fine-tuning to better support the target users—security personnel in Singapore.

Dataset	Kaldi	wav2vec 2.0	Whisper
Conversational AI	64.2	36.3	**19.9**
Phone Call	69.9	31.0	**16.6**
Meeting	44.0	27.4	**13.9**
Earnings Call	65.8	28.1	**9.7**
Video	47.6	23.3	**8.9**

Table 8: Normalized WER (%) across different ASR models and datasets (Seagraves, 2022). Lower WER indicates better performance.

To contextualize `D'Noise` for local security personnel, the team fine-tuned the `whisper-base` model on a noise-augmented Singlish database derived from the IMDA National Speech Corpus (IMDA, 2019). Crowd noise recordings averaging 72 dB were added to the dataset at varying Signal-to-Noise Ratios (SNRs) to improve robustness. The resulting

model achieved a Word Error Rate (WER) of **15.28%** on 40 minutes of noisy Singlish test audio—representing a **53.58% improvement** over `whisper-base.en` and a **94.0% improvement** over the multilingual `whisper-base` model (Table 9). Details of the initial fine-tuning process are documented in Appendix D.4.

Model Variants	Fine-tuned Whisper-base	Whisper-base.en	Whisper-base
Normalized WER (%)	**15.28%**	32.91%	254.61%

Table 9: Comparison of Normalized WER across variants of Whisper after fine-tuning. Lower WER indicates better transcription performance.

Real-time Transcription: The team used a Sliding Window approach to perform real-time transcription to continuously refine the transcription by using audio buffers (context windows). To find the best balance between performance and latency, a simulation on a 14-minute continuous test audio consisting of crowd noise at varying SNR ratios was conducted. For the best balance between latency and performance (Table 10), a 15-second context window was selected for real-time transcription.

Context Window Size (s)	5	10	15	20	25	30	Pauses
WER (%)	23.5	21.19	19.48	20.74	**19.44**	19.99	24
Latency	**0.14**	0.23	0.33	0.43	0.53	0.62	–

Table 10: Word-error rate (WER) and latency for different context-window sizes. Lower values are better; the best result for each metric is shown in bold.

3.3.3 Limitations of Denoising for ASR

While denoising methods are typically used to enhance speech quality and intelligibility, the team's experiments revealed an important limitation when such processed outputs are used for automatic speech recognition (ASR). Specifically, the team observed that using denoised outputs as inputs to ASR models led to a degradation in performance, quantified by their Word Error Rate (WER). Using the Whisper-small model as the team's ASR system, the team found that denoised audio inputs often resulted in higher WER compared to their corresponding noisy versions (Apppendix D.3.2, Figure 27). In some cases, even advanced deep learning-based denoisers such as the Facebook Denoiser or DeepFilterNet produced transcriptions that were less accurate than those derived from the unprocessed noisy input.

The team reasoned that this performance drop stems from artifacts or distortions introduced during denoising or the upsampling process, which shift the input distribution away from what the ASR model was trained on. To address this limitation, the team downsampled the denoised audio instead of upsampling to fit the high-quality audio of the shotgun microphone. As a result, there were significant gains in the WER as demonstrated in Figure 9. Additionally, the team conducted further experimentation and analysis on the capabilities of denoising to ascertain the validity of the denoising component. The details of this evaluation are documented in the Discussion (Section 4.1).

4 Discussion

This section details the quantitative and qualitative results of the overall prototype as well as the potential limitations for each component – Audio Capture, Audio Processing, Audio Transcription. From the observations and results, the team shows that it satisfies the hardware and software objectives (outlined in Section 1) and provides evidence that `D'Noise` is effective in addressing the problem.

4.1 Operational Distance

The team conducted experiments with human evaluation (Appendix F) on the denoising capabilities of `D'Noise`. The experiments **serve as an indicator** for the intelligibility of the playback audio in realistic environments. In the setup (Figure 8), three evaluators (receiver), at a fixed position, would hear the audio playback from `D'Noise` in capturing the Target's speech. After which, the receiver would indicate the interpretability of the captured speech based on their ability to transcribe the target speech. In the setup, only one source of interference was played in each session at different locations, I1, I2, and I3. The results of this experiment are detailed in Table 11 below.

This evaluation was conducted in an indoor corridor and at an open space near the SUTD canteen with average ambient sounds of 57dB and 72dB, respectively. These ambient volumes are naturally occurring sounds and are close to industry standards (60dB, 70dB, 80dB) (McCuaig, 2024) for noise isolation.

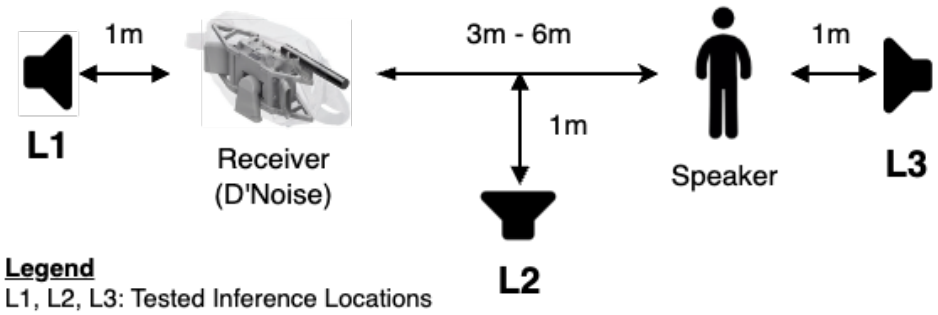


Figure 8: Setup for human evaluation to measure audibility to distance

Discreetness: Achieving similar performance at other interference positions is an important insight as it enables the security personnel to be even more discreet, without being in the line of sight of the target speaker. This directly fulfills one of the main design specification[1] (Section 1) for this prototype.

Distance: Achieving an approximate interpretable maximum distance of 5.5m and 4m for quiet and noisy environments, respectively, directly fulfills another design specification[2] of `D'Noise`. This range goes beyond conversational distances and the original target of a 3m-4m radius.

[1]Hardware Objective (II): A discreet design that is unobtrusive.

[2]Hardware and Software Objective (I): Enhanced playback within 200ms latency and within conversational distances (3m-4m).

Market Comparisons: `D'Noise` outperforms current market offerings by delivering clear audio capture up to ≈ 5.5 m in quiet environments and ≈ 4 m in noisy conditions—significantly beyond the typical 2–3 m range reported for comparable devices[3] like the Sudio T2 earbuds (Sudio, n.d.). Moreover, a key feature of `D'Noise` is its ability to **suppress dynamic noise**, filling a critical gap in existing Active Noise Cancellation (ANC) implementations that specialize in filtering static noises (McCuaig, 2024; Morantz, 2024). This makes `D'Noise` superior in both the operational distance and noise suppression capabilities over existing market products.

	3 m	3.5 m	4 m	4.5 m	5 m	5.5 m	6 m
I1 @ 57 dB	✓✓✓	✓✓✓	✓✓✓	✓✓✓	**✓✓×**	✓××	×××
I2 @ 57 dB	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓	**✓✓×**
I3 @ 57 dB	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓	**✓✓×**	✓××
I1 @ 72 dB	✓✓✓	**✓✓×**	✓××	✓××	×××	×××	×××
I2 @ 72 dB	✓✓✓	✓✓✓	✓✓×	**✓✓×**	×××	×××	×××
I3 @ 72 dB	✓✓✓	✓✓✓	**✓✓×**	✓××	×××	×××	×××

Table 11: Human evaluation of `D'Noise` interpretability at different distances and two average ambient sound levels (57 dB, 72 dB). Tick (✓) indicates that the evaluator was able to hear and transcribe the information, while cross (×) indicates they were not. *Boxed* results indicate the furthest interpretable distance by majority voting.

Next, the team considered the transcription capabilities of the fully integrated `D'Noise` in various locations with the results detailed in Figure 9. The setup of this test is similar to Figure 8, with a focus on transcription instead of the intelligibility of the speech. With a WER of at most 15.2% within conversational distance in noisy environments, `D'Noise` has successfully fulfilled Software objective (III) as detailed in Table 1.

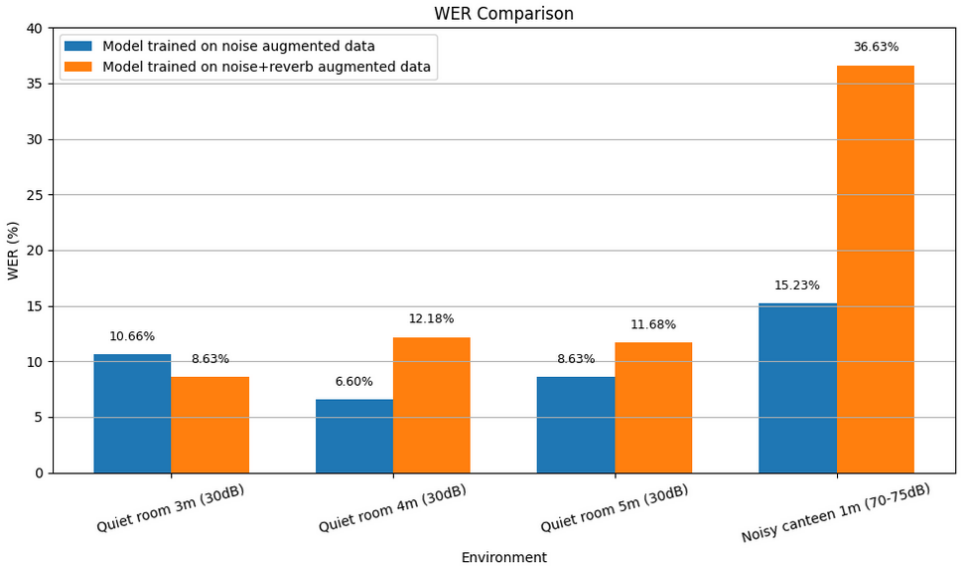


Figure 9: Transcription Word Error Rates (WER) in different settings and distances

[3] See Market Research in Section 2.3.

4.2 Operational Lifespan

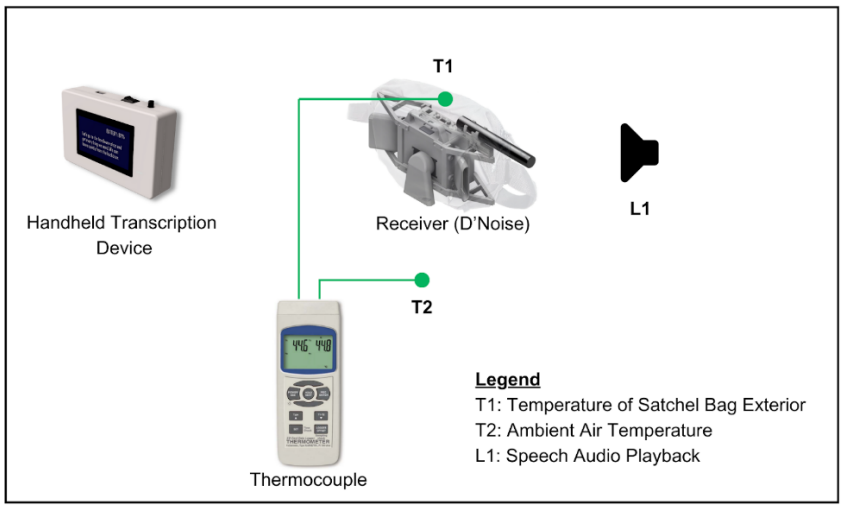


Figure 10: Stress testing setup to measure the battery lifespan and temperature of the device

To further validate the operational capabilities of `D'Noise` in helping security staff, the team conducted a stress test on the device. The motivation behind this test is to serve as an **indicator** for the prototype's ability to maintain its function continuously throughout the average number of hours before a break (3 hours). Figure 37 details the results of the stress test on `D'Noise`.

Jetson Nano Battery: The Jetson Orin Nano is the computational workhorse of `D'Noise` that all other processes, from denoising to transcription, are dependent on. Thus, a thorough test on the worst-case scenario for the Jetson Orin Nano is necessary to gauge how long `D'Noise` could last. The tests were set up to continuously play a recording of Singaporean local English from various sources, which would invoke the denoising and transcription features. The device was tested exactly from 90% battery and was allowed to drain fully (until `D'Noise` was forced to shut down). From this testing, the team observed that the device lasted around **4 hours and 27 minutes**, beyond the necessary shift. Given that the device was utilizing all features throughout the test, this duration serves as an indicative **lower bound** of the operational lifespan of `D'Noise`. Thereby, fulfilling the hardware requirement to last an entire security personnel's average number of hours before a break (3 hours).

Raspberry Pi Battery: The Raspberry Pi (RPi) is a supplementary computational device that connects to the handheld display, which visualizes the transcriptions from Whisper. The testing of this device is also necessary to assess the operational capacity of the transcription feature. From Appendix F Figure 38, the team also observed that the RPi could last **4 hours and 30 minutes** before shutting off, which matches the Jetson Orin Nano's battery lifespan.

Temperature and Comfort: A crucial metric that the team explored was the temperature of the satchel that houses `D'Noise`. The warmth of the bag would directly affect the comfort in using this device, which affects the viability of `D'Noise` for security personnel, especially in the warm weather of Singapore. From Appendix F Figure 39, the temperature of the satchel was not noticeably warmer than the ambient environment, reaching a **maximum difference of** $5^{\circ}C$**, capped at** $35.7^{\circ}C$. This validates the `D'Noise` in being a comfortable device, suitable for a prolonged operation if the situation necessitates as such.

4.3 Sustainability

The team developed and designed D'Noise while taking into account the 17 Sustainable Development Goals (SDG) outlined by the United Nations, where applicable. To demonstrate these considerations, the design of the device employs a modular layout by mounting the battery pack, shotgun microphone, and Jetson Orin Nano on separate brackets secured with standard screws. This approach extends service life and reduces e-waste by letting technicians replace any component independently, aligning with resource efficiency and infrastructure resilience (SDG 9) (United Nations, 2015d).

During development, the team used a low-infill 3D-printing setup with recycled PLA to reduce filament mass and energy consumption. Additionally, the battery relies on user-replaceable 21700 Li-Ion cells, so the entire unit is not discarded when the power source degrades. The team also ordered components from the same source at the same time as far as possible, to minimize the carbon footprint during the transportation of these components. These practices directly support responsible consumption and production (SDG 12) (United Nations, 2015b).

Lastly, providing a short disassembly guide and labeling all parts with polymer codes and e-waste symbols in the technical documentation is a future improvement the team could make. This labeling route helps users recycle worn components while retaining functional modules, contributing to sustainable cities (SDG 11) (United Nations, 2015a) and safeguarding health by minimizing hazardous disposal (SDG 3) (United Nations, 2015c).

4.4 Security Staff Validation Survey

The team conducted a user survey approved by SUTD's Institutional Review Board to further validate D'Noise from the intended users (security personnel). This survey interviewed two SUTD security guards to gather operationally relevant feedback. The test was conducted in indoor conditions and feedback was collected through interviews and structured response forms (Appendix G). Their evaluations are summarized in Table 12.

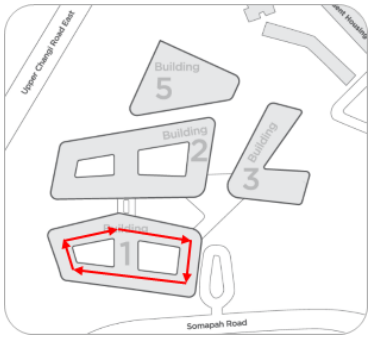


Figure 11: Security Staff Validation test route.

Validation Procedure:
The user validation was conducted in SUTD's Building 1, Level 6, along the designated patrol route shown in Figure 11. Security personnel were instructed to walk the route while interacting with the system under realistic noise and movement conditions. Evaluation criteria included system responsiveness (latency), transcription clarity, weight comfort, and usability of the interface components. Participants provided structured feedback after each patrol session.

Overall, this validation survey indicated the satisfactory performance of D'Noise for Singaporean security personnel as it addresses the primary design specifications of playing back an enhanced speech with real-time transcription and on-device computation whilst being discreet and portable. Thereby, placing D'Noise in a unique position that no existing market product could fill.

Form Factor	Participants positively evaluated the comfort and form factor of the device. It was described as comfortable and lightweight, preferred as a sling bag rather than attached at the hip, which was perceived as unstable. This successfully met Hardware Objective (II) concerning comfort and wearability, even with its non-standard format. Also, concerns were raised regarding discretion due to some visible wiring. This helped the team design the final prototype iteration for the Jetson Orin nano (Section 3.2, Figure 3c).
Performance	Security personnel commended clear speech playback up to 3 meters in environments with approximately 55 dB ambient noise, though performance notably declined beyond this range or with increased background noise. These findings match experimental results detailed in Section 4.1, thus meeting Hardware and Software Objective (I). Device integration into routine workflows was straightforward, with battery swapping and charging understood without formal training. However, tech-averse participants felt hesitant handling modular components, suggesting future iterations incorporate more intuitive physical cues or sealed connectors.
Ease of Use	Feedback on the user interface was mixed; participants found the OLED display bulky and showed a strong preference for mobile phone integration. While functionality targets were technically achieved, **transitioning to a mobile interface is recommended** to align better with market preferences and enhance user satisfaction.

Table 12: Summary of the User Validation Survey and Interview

Potential Improvements: In the interview, the team also asked the security guards for suggestions to further improve `D'Noise`. In summary, the participants suggested integrating features such as recording and playback capabilities for post-event reviews, adopting mobile interfaces for usability, strengthening data security through locking mechanisms, and adding weather-proofing or ruggedness for outdoor use. In light of this, the team proposed potential future directions to further improve the prototype in Section 4.5.

4.5 Future Work

Mobile Integration: As recommended by a security personnel during the user validation survey (Section 4.4) and informed by the limitations in Section 3.3.1, a potential direction for the display device would be to integrate the transcription into a mobile phone as illustrated in Appendix E.4, Figure 35. This iteration looks toward a design that could blend in more seamlessly into real-world scenarios, reducing unnecessary attention.

Audio Recording: As advised by the security personnel, implementing the ability to record and playback enhanced audio would greatly aid them in post-incident reviews. Thereby, further reducing the risk of losing out critical information that could jeopardize public safety.

Data Security: Another advantage of integrating the transcription into a mobile phone would be the use of privacy mechanisms such as two-factor authentication or biometric authentication, such as FaceID in iPhones. This would further secure the transcribed and the suggested audio recording data.

5 Conclusion

D'Noise presents a compelling entry in reducing the risk of missing out critical information by filling in gaps of existing market products – suppressing static and dynamic noise, enhancing and transcribing speech in adversarial settings. Throughout the development of D'Noise, the team customized the solution to help security personnel in Singapore and has effectively addressed the problem statement and design specifications (Table 13) outlined in the Introduction. Future directions (Section 4.5) can further optimize D'Noise for discreetness and its operational capabilities. As a result, this project successfully addressed a critical problem for security personnel and provided a foundation to further aid them in ensuring safety in Singapore

Objectives	D'Noise **Features**
Hardware	
(I) Effective Operational Range Capture speech within conversational distances (3m radius) in noisy environments.	✓An operational range up to 4.5m in noisy environments and 6m in quiet environments (Section 4.1).
(II) Discreet and Compact Design Unobtrusive form factor within 1.25 kg.	✓A total weight of $0.936kg$ (main body) + $0.25kg$ (handheld device) = $1.186kg$
(III) Extended Battery Life Continuous operation for at least 3 hours[4].	✓A lower bound battery life of 4 hours in extreme conditions (Section 4.2).
(IV) Comfort and Wearability Comfortable to wear throughout 3 hours[4].	✓A comfortable satchel bag that only reaches a maximum temperature of $35.7°C$ (Section 4.2).
Software	
(I) Minimal Latency Enhanced and denoised audio playback latency within 200 ms[5].	✓An average latency of 0.755ms for noise suppression and enhanced audio playback (Section 3.2, Table 7).
(II) Enhanced Audio Playback Improve baseline signal-to-noise ratio (SNR) for clearer speech playback.	✓Human evaluations (Section 4.1, Table 11) indicated more intelligible speech compared to the noisy input.
(III) Near Real-Time Transcription Transcription with a WER within 30% and a latency within 1s.	✓Average WER of 15.2% (Section 4.1) with an average latency of 0.33ms (Section 3.3.2, Table 10).
(IV) On-Device Computation Perform all processes on-device.	✓All processing modules are performed on the device without cloud services.

Table 13: Specified project objectives and the outcomes achieved by D'Noise.

[4]The average number of hours before a security personnel takes a break as specified by the industry mentor.
[5]The acceptable latency range for noise suppression within industry standards (**yang˙guided˙2023.**)

References

Alnuman, N., & Altaweel, M. Z. (2020). Investigation of the acoustical environment in a shopping mall and its correlation to the acoustic comfort of the workers. *Applied Sciences*, *10*(3), 1170.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020, October). Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations [arXiv:2006.11477 [cs]]. https://doi.org/10.48550/arXiv.2006.11477

Defossez, A., Synnaeve, G., & Adi, Y. (2020). Real time speech enhancement in the waveform domain. *Interspeech*.

Georgescu, A.-L., Pappalardo, A., Cucu, H., & Blott, M. (2021). Performance vs. hardware requirements in state-of-the-art automatic speech recognition [Number: 1 Publisher: SpringerOpen]. *EURASIP Journal on Audio, Speech, and Music Processing*, *2021*(1), 1–30. https://doi.org/10.1186/s13636-021-00217-4

Ghosh, S., Kumar, S., Seth, A., Chiniya, P., Tyagi, U., Duraiswami, R., & Manocha, D. (2024). Lipger: Visually-conditioned generative error correction for robust automatic speech recognition. *arXiv preprint arXiv:2406.04432*.

Gu, R., Zhang, S.-X., & Yu, D. (2023, February). 3D Neural Beamforming for Multi-channel Speech Separation Against Location Uncertainty [arXiv:2302.13462 [cs]]. https://doi.org/10.48550/arXiv.2302.13462

Guo, X., Wan, Q., Zhang, Y., & Liang, J. (2012). Teaching notes of MVDR in digital signal processing (DSP). *Proceedings of IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE) 2012*, H3A–5–H3A–8. https://doi.org/10.1109/TALE.2012.6360341

Hu, Y., Chen, C., Yang, C.-H. H., Li, R., Zhang, C., Chen, P.-Y., & Chng, E. (2024). Large language models are efficient learners of noise-robust speech recognition. *arXiv preprint arXiv:2401.10446*.

IMDA. (2019). National speech corpus [Retrieved 26 March, 2025]. https://www.imda.gov.sg/about-imda/emerging-technologies-and-research/artificial-intelligence/national-speech-corpus

Keller, M. D., Ziriax, J. M., Barns, W., Sheffield, B., Brungart, D., Thomas, T., Jaeger, B., & Yankaskas, K. (2017). Performance in noise: Impact of reduced speech intelligibility on sailor performance in a navy command and control environment. *Hearing Research*, *349*, 55–66.

Knapp, C., & Carter, G. (1976). The generalized correlation method for estimation of time delay [Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing]. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *24*(4), 320–327. https://doi.org/10.1109/TASSP.1976.1162830

Manthiram, A. (2017). An outlook on lithium ion battery technology. *ACS Central Science*, *3*(10), 1063–1069. https://doi.org/10.1021/acscentsci.7b00288

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive processes*, *27*(7-8), 953–978.

McCuaig, V. (2024). Our headphone tests: Noise isolation [Accessed: 2025-04-04]. https://www.rtings.com/headphones/tests/isolation/noise-isolation-cancellation-passive-active

Moonen, M., & Proudler, I. (2000). MVDR beamforming and generalized sidelobe cancellation based on inverse updating with residual extraction [Conference Name: IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing]. *IEEE Transac-*

tions on Circuits and Systems II: Analog and Digital Signal Processing, *47*(4), 352–358. https://doi.org/10.1109/82.839671

Morantz, A. (2024). Testing the capabilities of anc headphones: Bridging the gap to earmuffs [Accessed: 2025-04-04]. https://www.rtings.com/headphones/learn/research/noise-isolation

NVIDIA. (n.d.). Jetson orin nano developer kit [Accessed: 2025-03-21].

Pocketalk. (n.d.). Pocketalk - two-way voice translator [Accessed: 2025-03-21]. https://www.pocketalk.com/

Popescu, I. (2023). *Types of microphones: Dynamic, condenser, and ribbon* [Accessed: 2025-04-04]. AVIXA. https://www.avixa.org/pro-av-trends/articles/types-of-microphones

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlıcek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (n.d.). The Kaldi Speech Recognition Toolkit.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022, December). Robust Speech Recognition via Large-Scale Weak Supervision [arXiv:2212.04356 [eess]]. https://doi.org/10.48550/arXiv.2212.04356

Sainburg, T., Thielk, M., & Gentner, T. Q. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires [Publisher: Public Library of Science]. *PLOS Computational Biology*, *16*(10), e1008228. https://doi.org/10.1371/journal.pcbi.1008228

Sanchit, G. (2022). Fine-tune whisper for multilingual asr with transformers [Accessed: 2025-04-04].

Schröter, H., Escalante-B., A. N., Rosenkranz, T., & Maier, A. (2022). DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering. *ICASSP 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Seagraves, A. (2022, December). Benchmarking top open source speech recognition models: Whisper, facebook wav2vec2, and kaldi [Accessed: 2025-04-01]. https://deepgram.com/learn/benchmarking-top-open-source-speech-models

Stoica, P., & Nehorai, A. (1989). MUSIC, maximum likelihood, and Cramer-Rao bound [Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing]. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(5), 720–741. https://doi.org/10.1109/29.17564

Sudio. (n.d.). Sudio t2 - true wireless earbuds with anc [Accessed: 2025-03-21]. https://www.sudio.com/sg/t2-black

Townsend, A., Jiya, I. N., Martinson, C., Bessarabov, D., & Gouws, R. (2020). A comprehensive review of energy sources for unmanned aerial vehicles, their shortfalls and opportunities for improvements. *Heliyon*, *6*(11), e05285. https://doi.org/10.1016/j.heliyon.2020.e05285

TranscribeGlass. (n.d.). Transcribeglass - real-time captions for the deaf and hard-of-hearing [Accessed: 2025-03-21]. https://www.transcribeglass.com/

United Nations. (2015a). Sustainable development goal 11: Make cities and human settlements inclusive, safe, resilient and sustainable [Accessed: 2025-03-21].

United Nations. (2015b). Sustainable development goal 12: Ensure sustainable consumption and production patterns [Accessed: 2025-03-21].

United Nations. (2015c). Sustainable development goal 3: Ensure healthy lives and promote well-being for all at all ages [Accessed: 2025-03-21].

United Nations. (2015d). Sustainable development goal 9: Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation [Accessed: 2025-03-21].

Valin, J.-M. (2018). A hybrid dsp/deep learning approach to real-time full-band speech enhancement. *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, 1–5. https://doi.org/10.1109/MMSP.2018.8547084

Yang, C.-H. H., Park, T., Gong, Y., Li, Y., Chen, Z., Lin, Y.-T., Chen, C., Hu, Y., Dhawan, K., Żelasko, P., et al. (2024). Large language model based generative error correction: A challenge and baselines for speech recognition, speaker tagging, and emotion recognition. *2024 IEEE Spoken Language Technology Workshop (SLT)*, 371–378.

Yang, Y., Shih, S.-F., Erdogan, H., Lin, J. M., Lee, C., Li, Y., Sung, G., & Grundmann, M. (2023, March 13). Guided speech enhancement network. https://doi.org/10.48550/arXiv.2303.07486

A Simulations

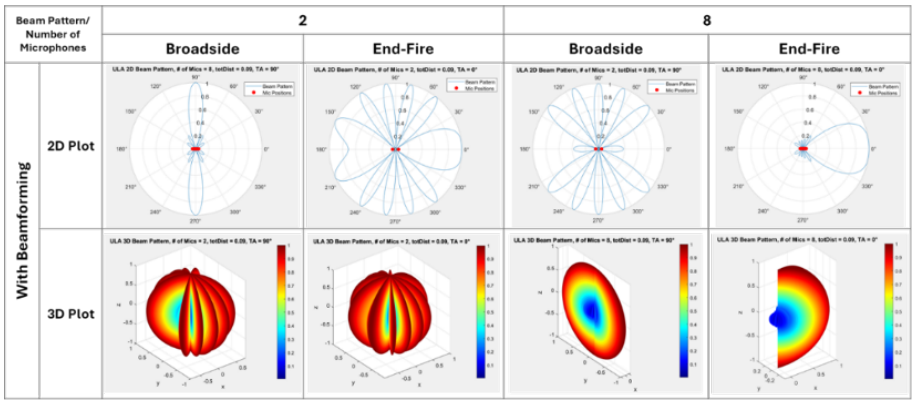


Figure 12: Broadside vs End-Fire Uniform Linear Array Simulation

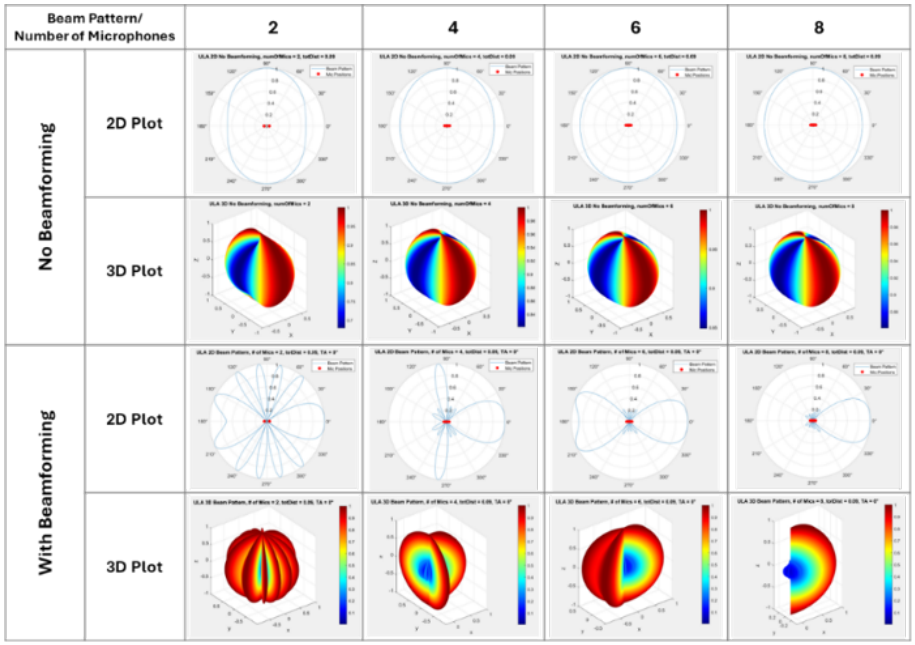


Figure 13: ULA 2D and 3D Beam Pattern Comparisons at 9cm Total Length with Varying Number of Microphones

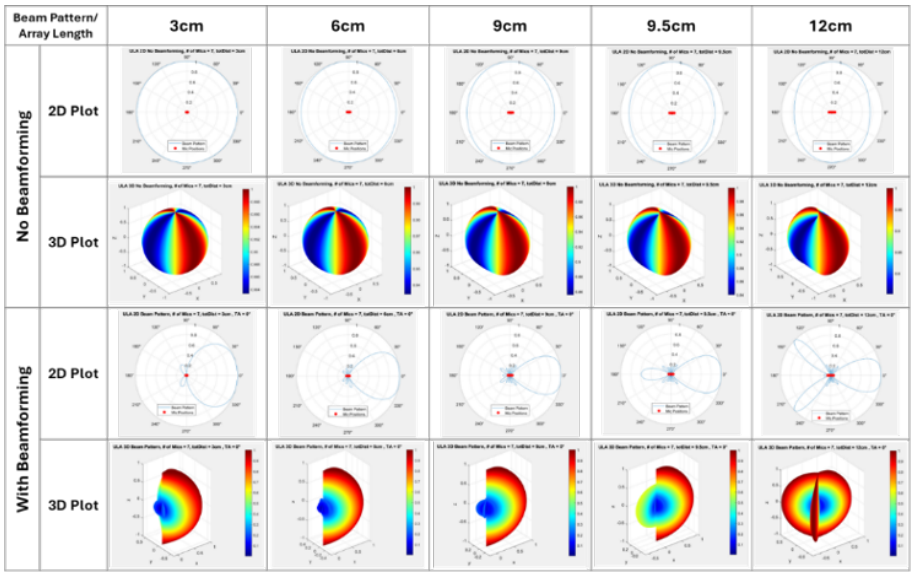


Figure 14: ULA 2D and 3D Beam Pattern Comparisons with 7 Microphones and Varying Array Length

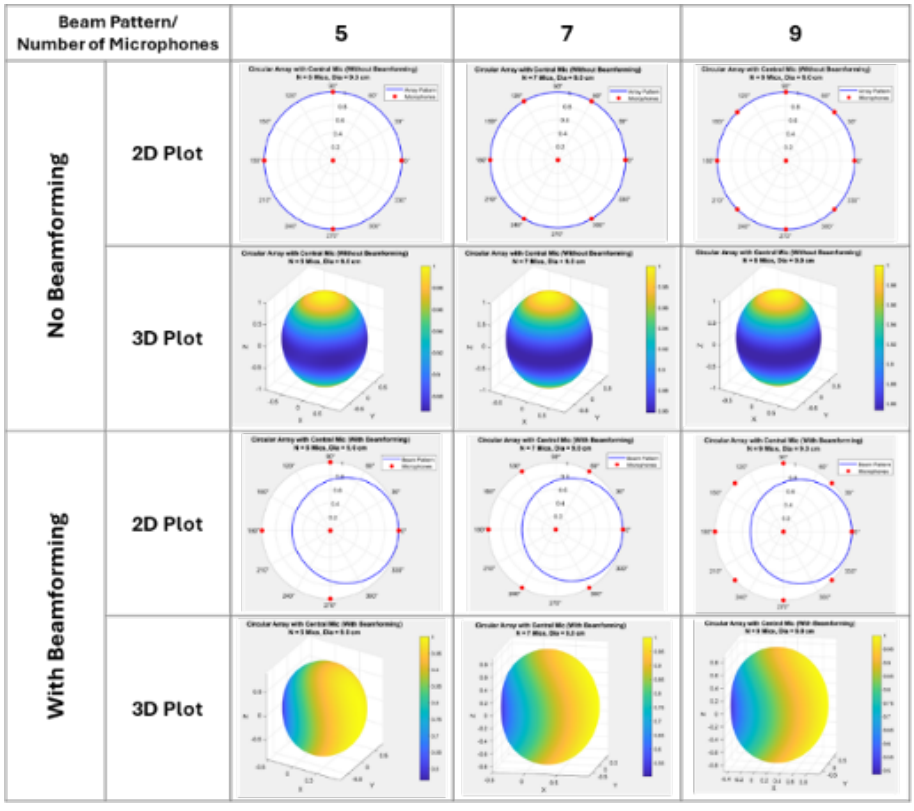


Figure 15: UCA 2D and 3D Beam Pattern Comparison at 9cm Diameter and Varying Number of Microphones with a Center Microphone

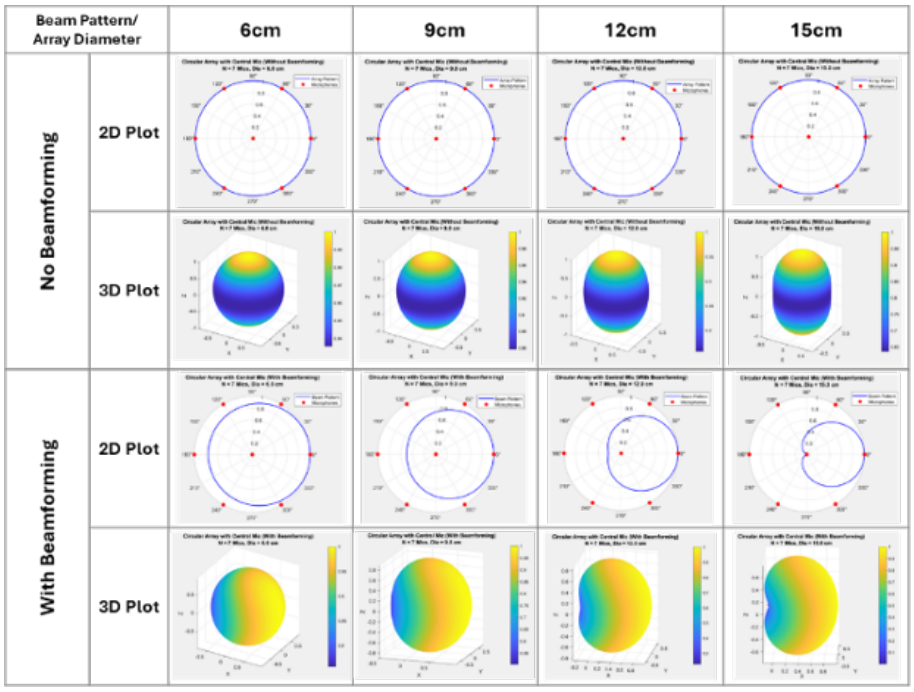


Figure 16: UCA 2D and 3D Beam Pattern Comparison at 7 Microphones (with Center Microphone) and Varying Diameter

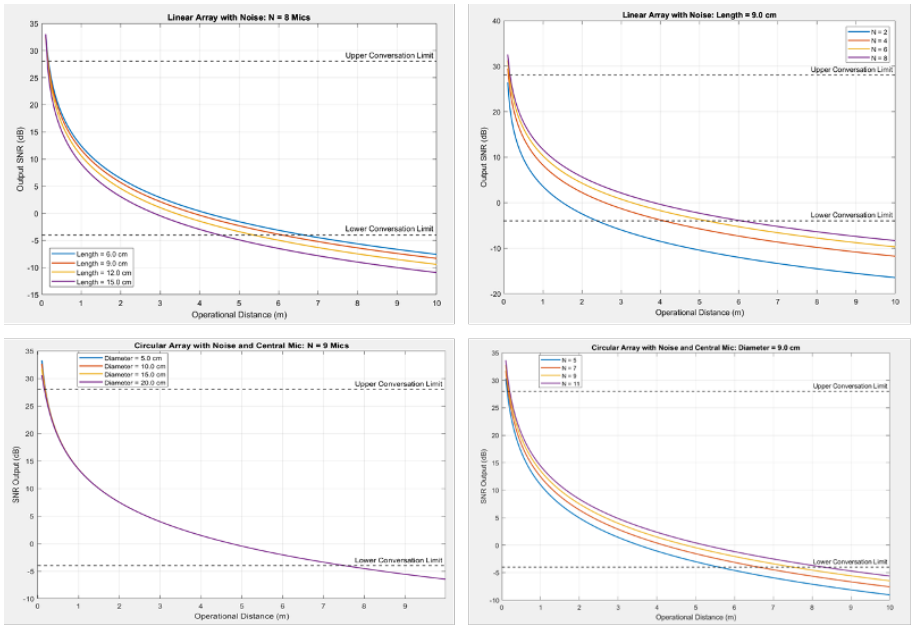


Figure 17: Microphone Array Configurations vs Operational Range

B Hardware Considerations

Parameters	UMA-8 v2 USB Microphone Array	ReSpeaker Mic V2	ReSpeaker 4-Mic Linear Array Kit	ReSpeaker 2-Mic Pi HAT V1.0
Image				
DSP/Processor	XMOS XVF3000 - Multicore USB audio processor with embedded DSP	XMOS XVF3000 - Multicore USB audio processor with embedded DSP	2 x X-Power AC108 ADC 1 x X-Power AC101 DAC	WM8960 Audio Codec
Microphones	7 x Knowles SPH1668LM4H	4 x ST MP34DT01-M digital MEMS microphone	4 x MSM321A3729H9BP MEMS microphones	2 x analogue MEMS microphones
Size	90mm (Diameter), 20mm (Height)	70mm (Diameter), 13.3mm (Height)	65mm x 30mm x 17mm (Pi HAT), 100mm x 20mm x 7mm (Mic Array)	65mm x 30mm x 15mm
Beamforming Capability and Configurability	• 360° beamforming, echo cancellation, and noise suppression • Access to individual raw microphone signals and DSP single output	• 360° beamforming, Direction of Arrival, Voice Activity Detection • Access to individual raw microphone signals and DSP single output	• No Beamforming stated • Access to individual raw microphone signals	• No Beamforming stated, Voice Activity Detection, Direction of Arrival and Key Word Spotting
Effective Range	3 metres (Noisy), 5 metres (Quiet)	3 metres (Noisy), 5 metres (Quiet)	No Specifications	3 metres (claimed)
Interface Type	USB	USB, 3.5mm Jack	Raspberry Pi 40-pin headers, I2C	Raspberry Pi 40-pin headers, I2C, 3.5mm jack
Ease of Integration with SBCs	Easy setup with USB port SBCs	Easy setup with USB port SBCs	Compatible with Raspberry Pi ONLY	Compatible with Raspberry Pi, Jetson Nano(GPIO reconfig)
Cost	SGD179.84 /unit	SGD128.57/unit	~ USD47/unit	~USD12.9/unit

Figure 18: A Comparison Between Microphone Arrays

The team conducted a comprehensive market analysis of MEMS microphone arrays, evaluating both linear and circular configurations. Products like the vicDIVA Beamforming Evaluation Kit and the VocalFusion 4-MIC XVF3500 Development Kit include integrated DSP systems that add significant bulk, limiting their suitability for compact designs. Meanwhile, arrays such as the ReSpeaker 4-Mic Linear Array Kit and 2-Mic Pi HAT, designed primarily for Raspberry Pi platforms, rely on GPIO pin interfaces, making integration with non-Raspberry Pi SBCs more complex.

Based on MATLAB simulations (Appendix A), circular microphone arrays were selected as the optimal configuration, narrowing the options to the UMA-8 v2 USB Microphone Array and the ReSpeaker Mic V2. Both were evaluated for effective range, beamforming performance, noise suppression capabilities, and ease of integration with SBCs. While the UMA-8 excelled in noise suppression, the ReSpeaker Mic V2 outperformed in Direction of Arrival (DOA) and beamforming. Both offered a similar effective range of 3–5 meters. Ultimately, the ReSpeaker Mic V2 was chosen for its ability to stream both raw and processed audio outputs, meeting the project's requirements. at the same time, which is essential for the denoising approach chosen for this project.

Parameters	Jetson Orin Nano	Jetson Nano	Raspberry Pi 5 + Hailo AI Accelerator	Google Coral Dev Board	BeagleBone AI-64
Image					
CPU/GPU /Processor	6-core ARM Cortex-A78AE + 1024-core GPU	Quad-core ARM Cortex-A57 + 128-core GPU	BCM2712 Quad-core Arm Cortex A76 processor @ 2.4GHz + Hailo 8l AI	Quad-core Cortex-A53 + Google Edge TPU	Quad-core Cortex-A72 + dual DSP cores
Neural Processing (AI)	High (40 TOPS , CUDA, TensorRT support)	Moderate (472 GFLOPS (FP32) ≈ 3.78 to 7.55 TOPS (INT8), CUDA)	Moderate (13 TOPS. NPU)	Low (4 TOPS, Edge TPU)	Moderate (8 TOPS, Matrix Multiply Accelerator)
Power Consumption	Moderate (~10-15W)	Low (~5-10W)	High (16.5-20W)	Low (~3W)	Moderate (~10-12W)
Physical Dimensions	103mm x 90mm x 35mm	100mm x 80mm x 35mm	85mm x 56mm x 17mm (Pi 5) 65mm x 56.5mm (Hailo + M2. Hat)	88mm x 60mm	86.4mm x 53.3mm
Cost	SGD663.81	SGD265	SGD191.97	SGD180.4 - 289.9	SGD252

Figure 19: A Comparison Between Single Board Computers

The Jetson Orin Nano was selected as the SBC for its excellent performance-to-power trade-off, offering 40 Trillions of Operations Per Second (TOPS) of computational power to ensure low-latency speech processing and transcription while maintaining moderate power consumption (10-15 W). This balance allows for extended operation using a smaller battery, making it well-suited for integration into a wearable device. Other SBCs, such as the Raspberry Pi 5 with an AI accelerator and the BeagleBone AI-64, were evaluated based on factors like computational performance, power efficiency, size, and cost. While some lower-power devices were considered, they lacked sufficient processing capabilities to meet real-time requirements. Additionally, the team explored offloading audio processing and transcription to the cloud, which could allow for a smaller, lighter device. However, local processing on the Jetson Orin Nano was ultimately chosen for its ability to reduce overall costs, simplify system architecture, and eliminate the dependency on a stable network connection.

Parameters	Lithium Polymer (LiPo)	Lithium Ion (Li-Ion)
Image		
Power Density	High - excellent for compact designs	High - Slightly lower than LiPo
Discharge Rate	High – 20C or more	Moderate – 2C to 5C
Safety/Fire Hazard	Plastic casing more prone to physical damage which can result in fire	Rigid, metal cylindrical cells more resistant to mechanical stress
Form Factor	Customizable shapes, ideal for thin or compact designs	Fixed cylindrical shape is less flexible for integration
Weight	Slightly lighter due to plastic casing	Heavier due to metal casing

Figure 20: A Comparison Between Lithium-Based Batteries

For a wearable device, a Lithium-based battery is desirable due to its high energy density relative to other battery types (Li et al., 2024). Although Lithium Polymer (LiPo) batteries offer slightly higher energy density and versatile form factors for integration into wearables, Lithium Ion (Li-Ion) batteries were ultimately selected for this application because of their superior resistance to mechanical stress. Furthermore, the Jetson Orin Nano's maximum current draw of 1.25 A at 12 V does not require LiPo-grade high discharge rates, making Li-Ion batteries more than sufficient.

Battery Sizing: Standard security officer shifts in Singapore last for 12 hours (Today Online, 2018). To reduce the frequency of battery swaps to a single swap while keeping overall weight manageable, the device was designed to operate independently for 6 hours. Assuming the Jetson Orin Nano consumes 12 W on average and the buck converter exhibits an efficiency of 90%, the power draw from the battery is calculated as:

$$\textbf{Power Draw from Battery (W)} = \frac{\text{Power Consumed by Jetson Orin (W)}}{\text{Buck Converter Efficiency}} = \frac{12}{0.9} \approx 13.33\,\text{W}.$$

The required battery capacity to run for 6 hours at 14.8 V is:

$$\text{Capacity (mAh)} = \frac{13.33\,\text{W} \times 6\,\text{h}}{14.8\,\text{V}} \times 1000 \approx 5392\,\text{mAh}.$$

Adding a safety margin of 30% leads to:

$$5392\,\text{mAh} \times 1.3 \approx 7009.6\,\text{mAh}.$$

Consequently, the battery chosen for this application is a 4-Series 2-Parallel Li-Ion pack with a capacity of 7000 mAh, providing sufficient runtime while maintaining a compact form factor.

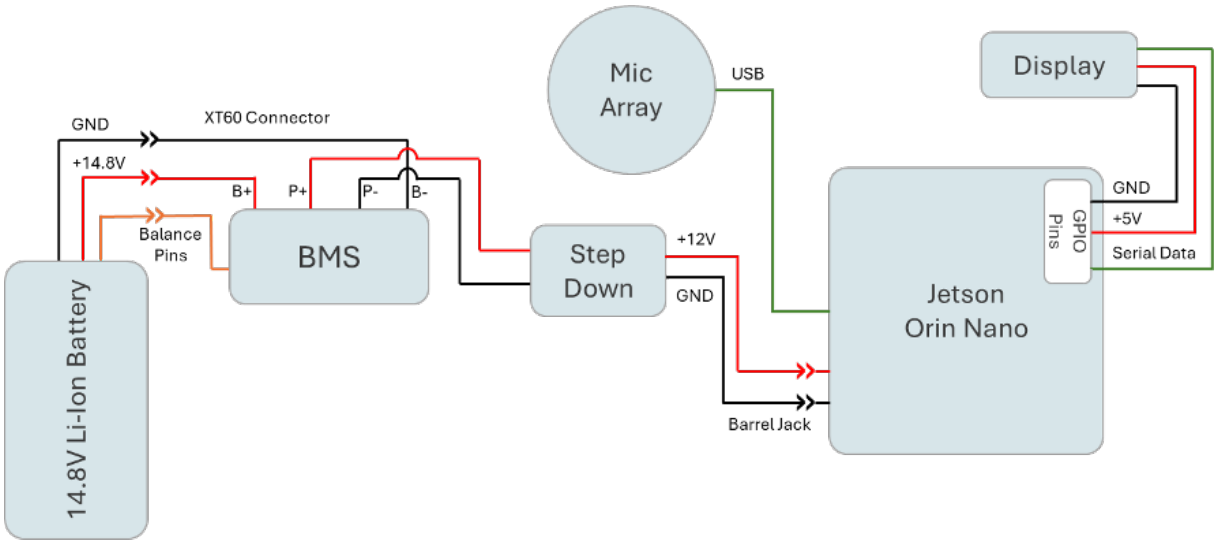


Figure 21: Connection Diagram for LiPo Battery Power Systems

The operation of the Jetson Orin Nano via a Li-Ion battery poses some safety issues, which are addressed with the following components.

Battery Management System (BMS): A Battery Management System (BMS) prevents the battery from over-discharging, which causes permanent damage and a potential fire or explosion hazard. Additionally, it protects the rest of the electronics from excessively high current.

Step Down (Buck) Converter: While the selected Lithium-Polymer battery has a nominal voltage of 14.8V, its voltage will vary from 16.8V (full charge) to 12.0V (cutoff voltage) as it discharges during use. A step-down converter can take in this varying voltage and provide a steady voltage of 12.0V to the Jetson Orin Nano.

C Software Considerations

C.1 Generative Error Correction (GER)

To improve the performance of the Automatic Speech Recognition (ASR) models, incorporating Generative Error Correction (GER) (C.-H. H. Yang et al., 2024) modules is a possible direction to improve transcription accuracy, especially in extremely noisy scenarios where traditional denoising methods fall short. GER refines ASR outputs by leveraging the linguistic capabilities of generative models to correct transcription errors. Initial research considered two GER approaches—RobustGER and LipGER—to help improve ASR outputs.

RobustGER (Hu et al., 2024) improves transcription accuracy by performing language-space denoising. ASR models are modified to generate N-best plausible transcriptions, and language-space noise embeddings are calculated from the inconsistencies among these hypotheses. Greater discrepancies represent noisier conditions. Sentence-level embeddings summarize overall noise across hypotheses, while token-level embeddings address fine-grained variations. These embeddings, derived using Sentence Transformers (SBERT) and alignment algorithms, are input to a Large Language Model, which is instructed to ”subtract” the calculated noise embeddings to refine the transcriptions based on the N-best outputs. LipGER framework enhances transcription accuracy by integrating audio inputs with visual embeddings derived from lip movements (Ghosh et al., 2024). This multimodal approach is particularly effective in noisy environments where audio signals alone may be inadequate, making it a promising solution for challenging acoustic conditions. However, this method introduces another modality to the system which would have cascading implications on latency, computational demands, and power draw.

For all evaluations and experiments, the **Word Error Rate** is calculated as such:

$$\text{WER} = \frac{\text{number of substitutions} + \text{number of insertions} + \text{number of deletions}}{\text{length of reference transcription}}$$

D Deep Learning Models

D.1 Dataset

D.1.1 National Speech Corpus (IMDA, 2019)

The dataset used in this capstone project includes speech recordings drawn primarily from the IMDA National Speech Corpus (NSC), which forms the foundation for training and evaluating the many of the deep learning models we used. The IMDA NSC is a curated speech dataset developed by Singapore's Infocomm Media Development Authority to support research and development in local speech technology. It consists of high-quality audio recordings of Singaporean English, reflecting a variety of accents and speaking styles that are representative of the local population.

The corpus includes recordings from multiple speakers across different demographics from Singapore, and the content spans a range of topics including everyday conversations, scripted phrases, and read passages. All recordings were captured in acoustically controlled environments using high-fidelity microphones to ensure minimal background noise, making the dataset particularly suitable as a clean speech reference in speech enhancement tasks.

The IMDA NSC was selected not only for its linguistic relevance but also for its high signal-to-noise ratio, which allows for precise simulation of corrupted speech when overlaid with synthetic noise. This makes it well-suited for use in training models to perform supervised speech denoising. The segmented clean speech data was stored in the directory `prepared_data/clean_speech` and used as ground truth in all denoising experiments and training of the GSE Network, DeepFilterNet and also in the finetuning of Whisper.

D.1.2 Data Collection with Microphone Arrays

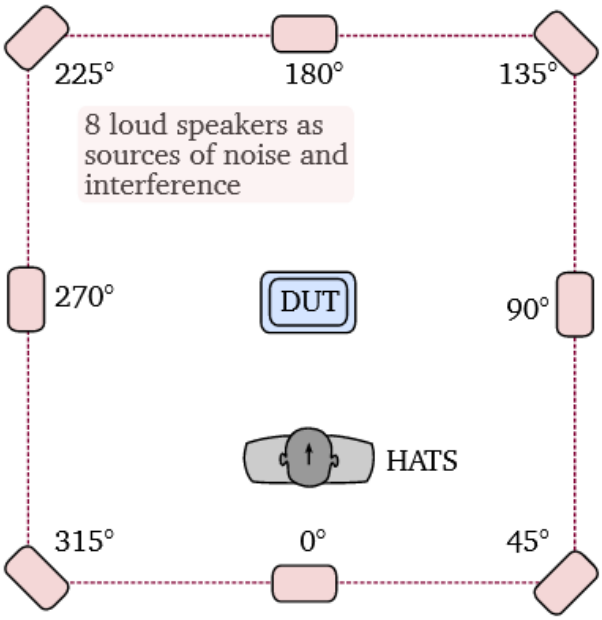


Figure 22: The Head and Torso Simulator from the Guided Speech Enhancement Network (Y. Yang et al., 2023)

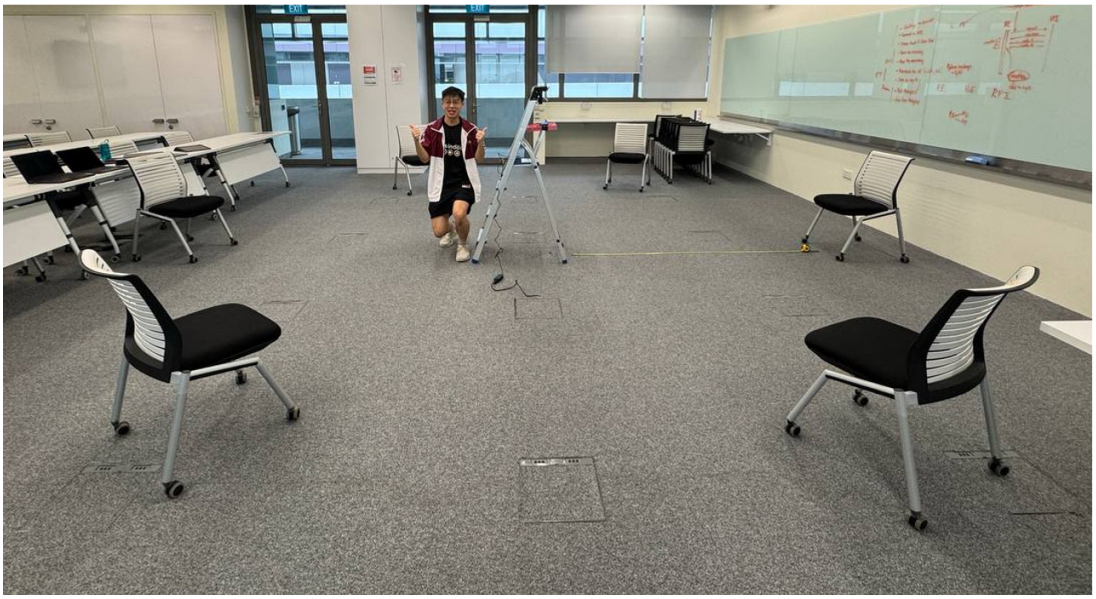


Figure 23: The Head and Torso Simulator setup in SUTD for the data collection initiative.

The team also initiated an in depth data collection effort to train the deep learning model that uses the microphone arrays as mentioned in Section 3.2.2. This data collection effort follows the Head and Torso Simulator setup (Figure 22) with a similar distribution of the data collection parameters from the authors of the Guided Speech Enhancement Network (Y. Yang et al., 2023). Instead of 8 loud speakers for the interference, the team used 5 instead with an equal angular spread from the torso at 3 metres from the recording device (Figure 23). In total, the team collected around **20 hours of data** using the specialized hardware – ReSpeaker Microphone V2 – to train and validate the GSENetwork and the DeepFilterNet.

D.2 Guided Speech Enhancement Network (Y. Yang et al., 2023)

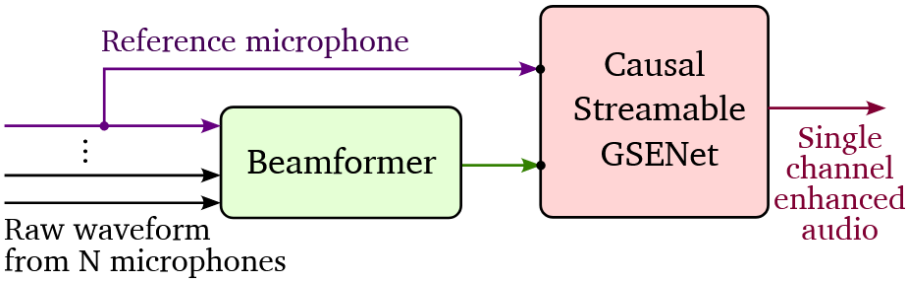


Figure 24: System diagram for the Guided Speech Enhancement Network. GSENetwork takes in both the beamformer outputs and the raw inputs.

The Guided Speech Enhancement Network (GSE Network) combines a beamformer with a deep learning model to enhance speech by reducing noise and interfering signals. It takes in both the raw microphone audio and beamformer output as inputs to improve spatial filtering, denoising, and dereverberation 24. Using a causal U-Net architecture 24, the GSE Network is able to operate in real-time to reduce noise and enhance the target speech.

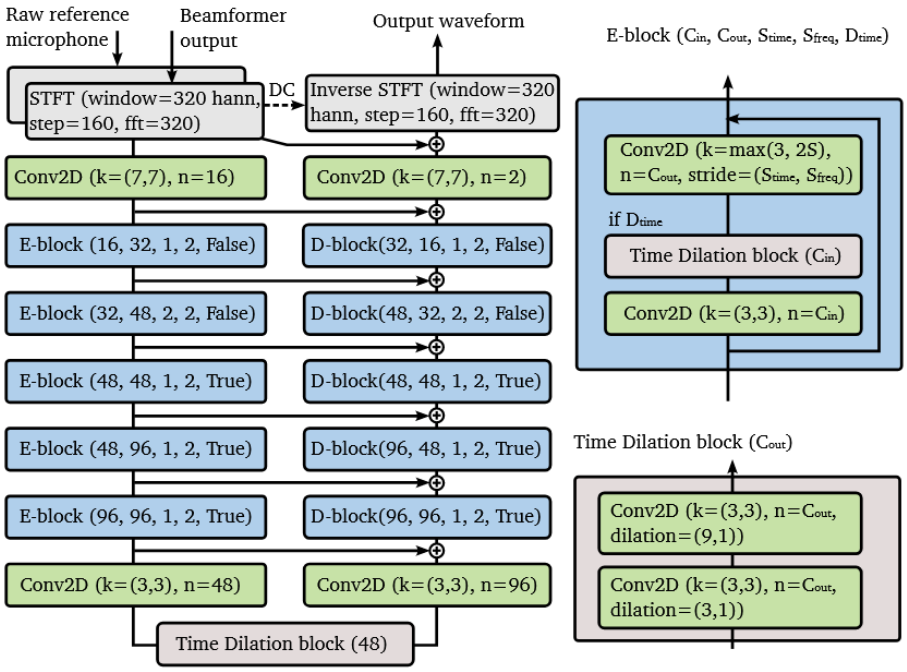


Figure 25: Architecture of GSENetwork as defined in the paper.

D.2.1 Training Preamble

The training script was optimized to utilize Apple Metal Performance Shaders (MPS) when available, defaulting to CPU otherwise. Dependencies required for MPS-accelerated PyTorch were specified in `gsenet_requirements.txt`. The training was executed on an M2 Max MacBook Pro, completing in approximately 4 hours. The process begins by setting a random seed of 13 for reproducibility across PyTorch, NumPy, and Python's random module. If CUDA is available, its seed is also initialized. The training configuration employs a batch size of 32 and a learning rate of 0.001. The detailed training implementation can be found in this GitHub Repository.

D.2.2 GSENetwork Training Procedure

The training procedure for the Guided Speech Enhancement Network (GSENet) was implemented to support both local and cloud-based training, specifically leveraging Google Cloud Platform (GCP) for scalable compute capabilities.

Local training was first performed on a personal M2 Max MacBook Pro. The training script `train.py` was executed with arguments specifying the dataset path, batch size, number of epochs, and the directory for saving model checkpoints. The model was trained over 20 epochs with checkpoint files saved periodically to disk. This allowed fast iteration during the early phases of model development.

For more scalable training on larger datasets, the workflow was migrated to **Google Cloud**. A Google Cloud Storage (GCS) bucket was first created to host the self-collected training data. The local dataset was uploaded to this bucket using the `gsutil` utility with multi-threaded transfer enabled via the `-m` flag to accelerate the upload process. A Google Compute Engine instance was then provisioned with GPU support to handle the training workload. After initializing the

instance and connecting via SSH, the training repository was cloned and all dependencies were installed. The dataset was accessed either by downloading it from the GCS bucket or by mounting the bucket using `gcsfuse` for direct access without local duplication.

The training script was then executed on the virtual machine, using similar arguments as before but updated with the correct paths reflecting either local or mounted GCS data directories. Checkpoints were stored locally on the VM, with the option to upload them back to the GCS bucket for persistence and portability. Upon completion of training, the model's performance was evaluated using the held-out test dataset. The best model weights were saved to `gsenet_final_model.pth`, resulting in a model file approximately seven megabytes in size. This file encapsulates the learned weights and biases, and is suitable for downstream deployment or fine-tuning.

D.3 Multi-Channel RNNoise

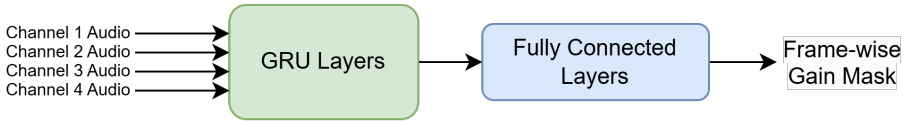


Figure 26: Architecture of Multi-Channel RNNoise.

The Multi-Channel RNNoise architecture builds upon the original single-channel RNNoise model by extending it to process four audio channels. This modification was inspired by beamforming algorithms and technologies, which make use of multi-channel microphone arrays to enable directional audio capture. The network aims to leverage spatial cues—particularly phase differences between microphones—for improved noise suppression and speech enhancement. Instead of predicting clean waveforms directly, the network estimates a frame-wise gain mask in the spectral domain, which is then applied to the noisy signal. This approach is efficient for low-latency applications (as in our use case, which requires real-time audio playback), as it operates in the frequency domain and avoids complex signal reconstruction. The architecture combines principles from traditional speech enhancement with recurrent neural networks, leveraging multi-channel input to better filter target speech from background noise. By doing so, we aim to improve the model's robustness to challenging acoustic conditions such as noisy or multi-speaker environments.

D.3.1 Multi-Channel RNNoise Training and Results

The model architecture began with a simple extension of RNNoise—consisting of a few GRU layers followed by fully connected layers—adapted to process all four channels of audio input from a microphone array. After computing the gain mask, this was applied to the spectral representation of the input signal. As development progressed, we experimented with different architectural variants, including increased depth, skip connections, and other modifications. We also evaluated the effect of different loss functions: mean squared error (MSE) loss, and short-time Fourier transform (STFT) loss. The latter computes the error in the time-frequency domain, offering better perceptual alignment with audio signals and sensitivity to both magnitude and phase differences. Lastly, we also tried adding the beamformed output from the microphone

array as an additional input (much like in the GSE network in the section before) to the neural network.

Despite these variations in architecture and loss functions, our multi-channel models unfortunately did not outperform the baseline single-channel RNNoise. The results suggest that more advanced or complex spatial modeling or architectural changes may be necessary to fully leverage multi-channel input for enhanced performance in directional audio capture.

D.3.2 Limitations in Evaluating Noise Suppression

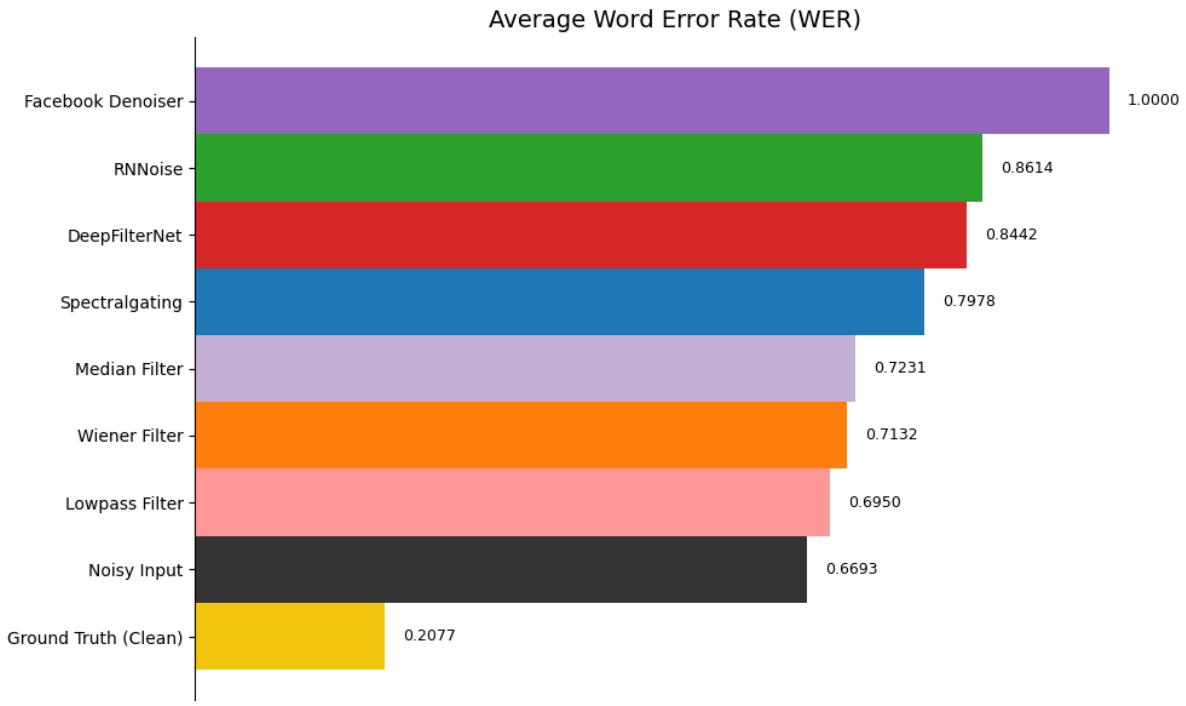


Figure 27: Average WER of ASR outputs for different denoising methods (Whisper-small)

During the evaluation for the denoising deep learning models, a major challenge that the team encountered was the selection of a quantifiable metric that can effectively describe how well each method has performed. Initially, the team considered SNR improvements to quantify the performance. However, this metric was not representative for real time use cases as the Facebook Denoiser outputs garbled intelligible noise but obtained the highest SNR improvement. To address this, we considered the use of average WER by routing the denoised outputs to a Whisper model. The intention is to convert the denoising capabilities into a quantifiable metric - WER - to evaluate the models. Unexpectedly, the WER for the denoising methods was higher than the noisy input.

While further investigations could be conducted to address this, the team focused on exploring other methods to compare the different models instead due to time and resource constraints. Thus, human evaluations in Section 4.2 was conducted to better determine which denoising method produces the highest quality speech – with comparisons between the noisy input and the latency involved.

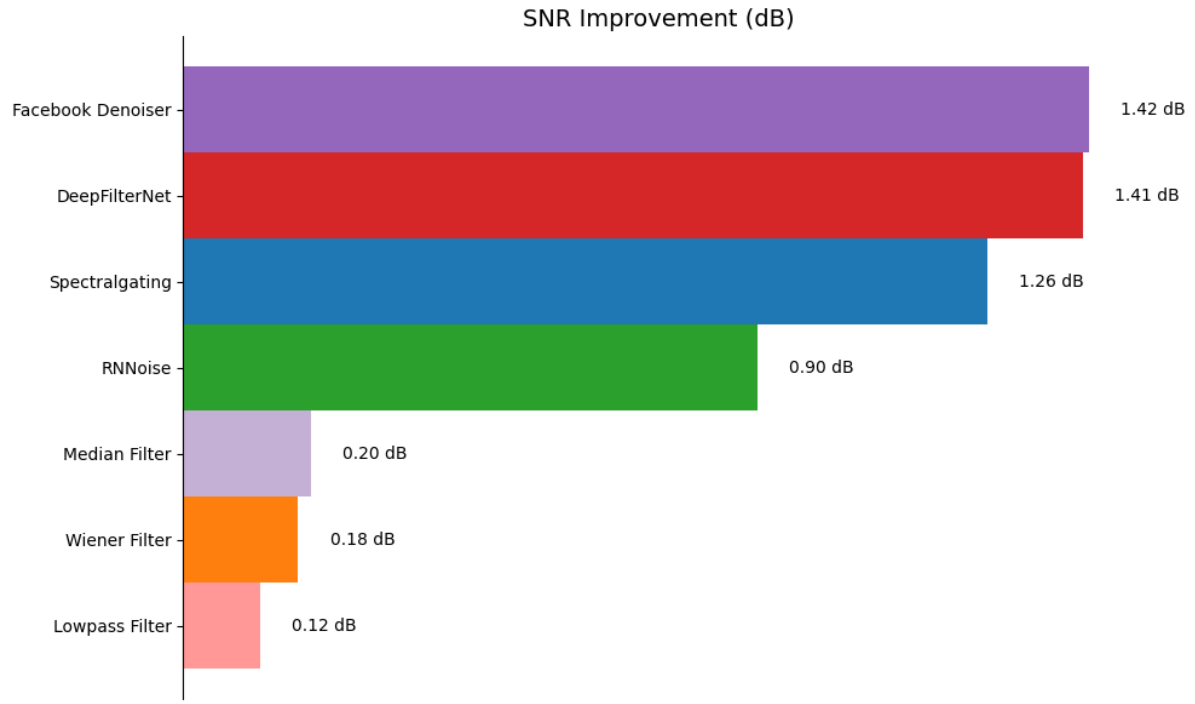


Figure 28: Average SNR improvement for different denoising methods

Insight: This led us to the conclusion that our initial assumption – the denoised outputs would be easier for the whisper model to transcribe – is not necessarily true. As such, we concluded that the best way to measure a model's ability to suppress noise is through human evaluations as detailed in Section 4.1.

D.4 Whisper Finetuning

Open-source ASR models such as Kaldi, Wav2Vec 2.0, and Whisper were evaluated for performance, scalability, and community support. The Whisper model was ultimately selected for the project due to its strong multilingual pretraining, robustness to background noise, and active ecosystem for low-resource language adaptation. Among its variants—tiny, base, small, and medium—the base model was primarily used for deployment and fine-tuning, offering a balance between computational feasibility and transcription performance in noisy environments such as airports and shopping malls. Compared to Kaldi and Wav2Vec 2.0, Whisper demonstrated superior transcription accuracy, with relative improvements of 45.18% and 69%, respectively.

To tailor the prototype toward the target users—security personnel operating in Singapore—the Whisper model was fine-tuned to transcribe Singlish, a colloquial form of English commonly spoken in the region. Singlish presents unique challenges for ASR systems due to its informal structure, localized vocabulary, and non-standard grammar, which are not well-represented in most pretraining corpora. To address this, the conversational portion of the National Speech Corpus (NSC), curated by the Infocomm Media Development Authority (IMDA), was used for fine-tuning (Appendix D.1). This dataset was selected for its close linguistic alignment with the target domain and was hosted using HuggingFace Datasets for scalable access during training.

The fine-tuning process involved adapting the Whisper model's parameters to capture the nuances of Singlish speech while retaining the benefits of multilingual pretraining. The model was trained to minimize transcription error using the NSC conversational data, with training and evaluation loss curves provided in Figure 29 and Figure 30. The fine-tuned model achieved a 66% improvement in normalized Word Error Rate (WER) compared to the Whisper English-only baseline, and a 93% improvement relative to the unadapted Whisper model.

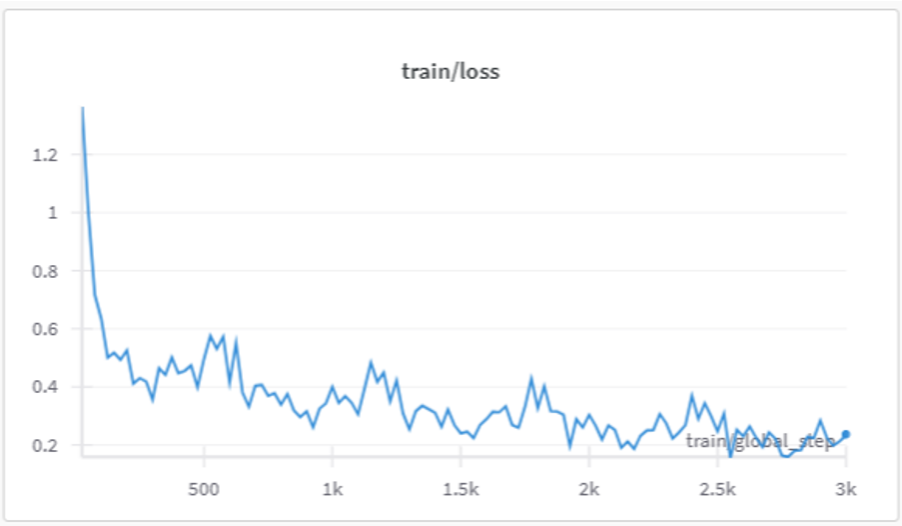


Figure 29: Training Loss Curve during Finetuning

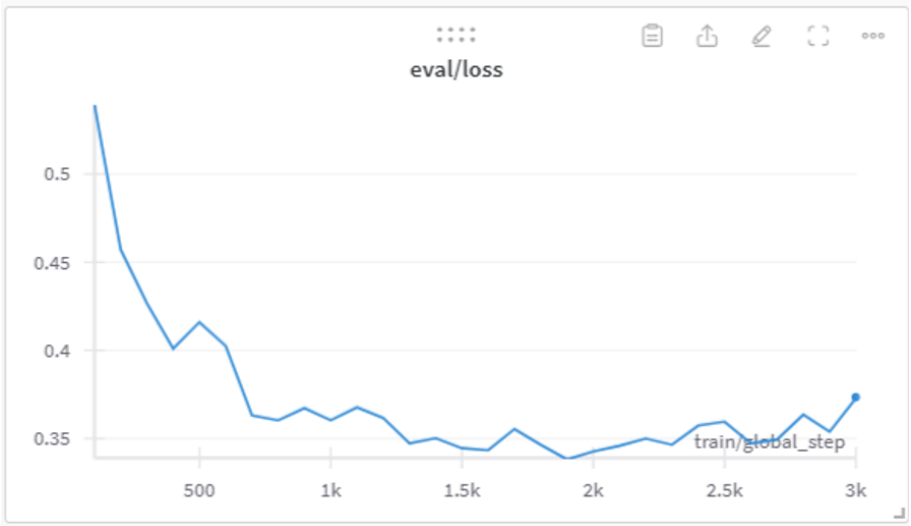


Figure 30: Evaluationn Loss Curve during FineTuning

In addition to performance gains, training was carefully configured to mitigate overfitting and improve generalization to real-world speech. This was done by tuning the frequency at which the model's parameters were updated, effectively addressing the generalization gap by increasing data variation across training iterations. Through this process, Whisper was effectively adapted to meet the linguistic demands of the deployment environment, enhancing its suitability for real-time transcription tasks in noisy, multilingual, and informal communication scenarios encountered by front-line security staff.

The team also initially customized Whisper by team fine-tuning the model specifically for Singlish, a colloquial English variant unique to Singapore, characterized by distinct grammar, vocabulary, and pronunciation patterns. For this initial finetuning process, the dataset was **not augmented** with additional noise and reverberations unlike the process documented in Section 3.3 A conversational subset of the IMDA National Speech Corpus (IMDA, 2019) comprising 530 hours of training and 132 hours of testing data was used. We achieved a **66% improvement** over the English-only version of the Whisper model and a **93% improvement** over the multilingual base model (Table 14).

Model Variants	**Fine-tuned Whisper-base**	Whisper-base.en	Whisper-base
Normalized WER (%)	**12.87%**	37.77%	192.86%

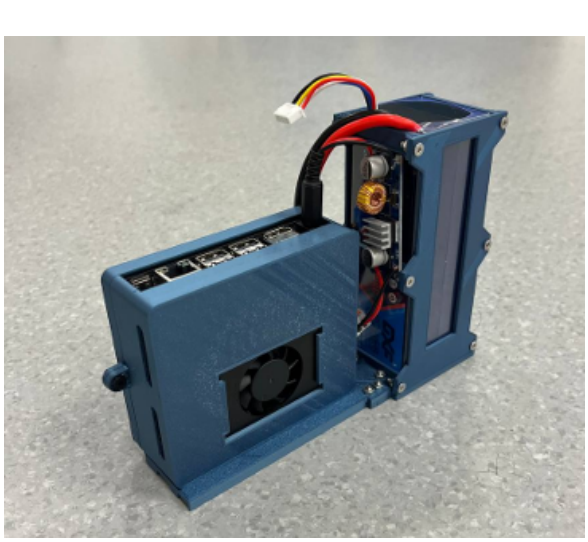
Table 14: Comparison of Normalized Word Error Rate (WER) across Whisper model variants after fine-tuning on Singlish. Lower WER indicates better transcription performance.

E Overall Design Iterations

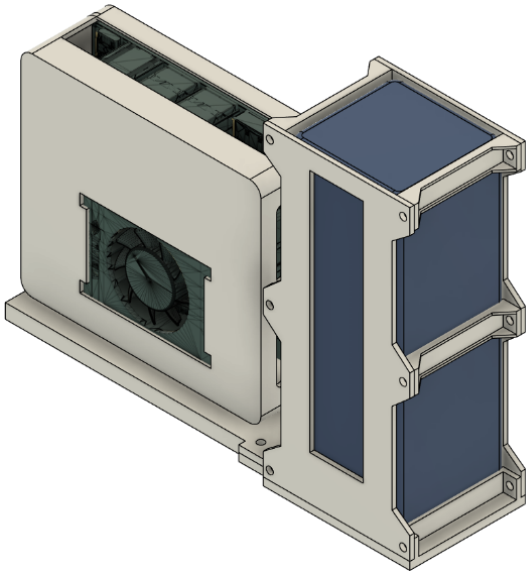
E.1 Main Body First Iteration

This prototype consisted of a Raspberry Pi Zero 2W powered by a 3.7V, 2000 mAh Li-Po battery through a PowerBoost 1000 Basic module, and incorporated a transparent OLED display. The ReSpeaker microphone array was originally placed at the hat area to capture audio input.

Consequently, the system underwent a design evolution, replacing the ReSpeaker array with a single or shotgun microphone to free space on the hat, and relocating the Raspberry Pi, booster, and battery from the glasses to the hat area. This revised layout used adjustable head straps to achieve greater comfort and incorporated a rotating arm (extending approximately 10 cm from the hat) to position the transparent OLED directly in front of the user's eye.



(a) First iteration of housing the Jetson Nano and the Battery



(b) 3D Render of the first iteration

Figure 31: Side-by-side comparison of hardware housing iterations.

As discussed in Section 3.2, there were various limitations to the first iteration which helped inform subsequent designs. Namely, the team further improved on the weight distribution, heat, and discreetness of the device.

E.2 Main Body Second Iteration

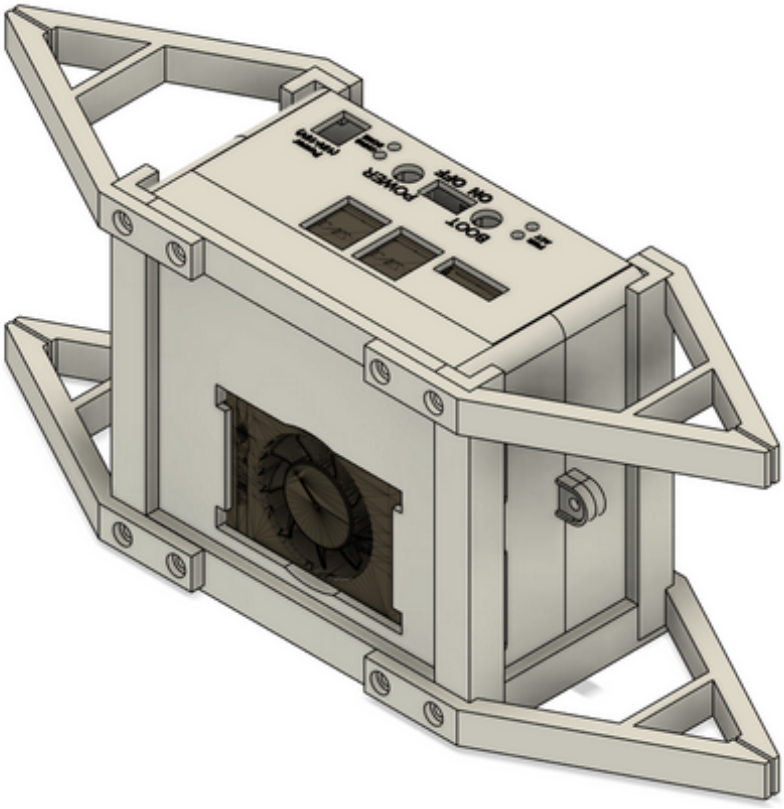


Figure 32: Second iteration of the main body

To address the limitations identified in Iteration 1, the team transitioned to a modular and safer power architecture. The 4S Li-Po battery was replaced with three 4800 mAh 21700 Li-Ion cells, each enclosed in a protective steel casing. This change significantly improved both the physical robustness of the system and the ease of battery replacement. The cylindrical batteries could be swapped intuitively, much like conventional AA batteries, thereby enhancing field usability.

For power management, the Waveshare Uninterruptible Power Supply (UPS) for the Jetson Orin Nano was adopted. This module was selected over a custom Battery Management System (BMS) due to its integrated feature set, which includes:

- Hot-swapping between AC and battery power,
- Overcharge and over-discharge protection,
- Internal charging via standard AC wall plug.

The UPS mounts directly beneath the Jetson Orin Nano, maintaining the same horizontal footprint while increasing the vertical stack height. As a result, a new internal frame was designed to accommodate the combined Jetson, UPS, and battery compartments efficiently.

With the improved modular power system in place, the operational runtime requirement was revised to 3 hours. This decision aligned with the availability of natural breaks during a typical shift, such as meal times, during which batteries could be replaced. The updated configuration

not only ensured adequate runtime but also contributed to reducing the overall system weight, enhancing user comfort during prolonged wear.

E.3 Main Body Third Iteration

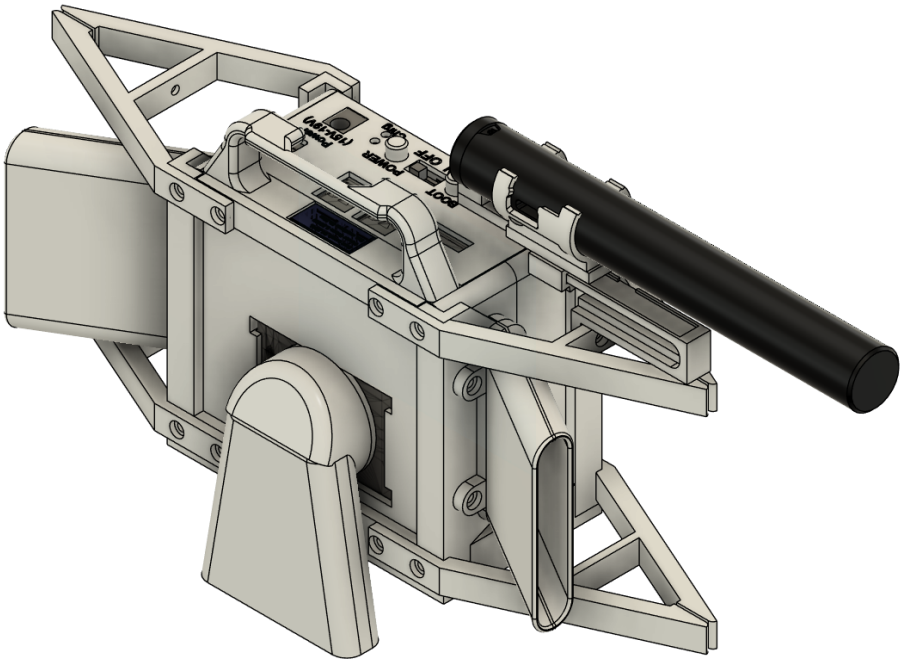


Figure 33: Third iteration of the main body

In real-world testing conditions, thermal issues emerged as a critical bottleneck. When the system was fully enclosed in the satchel bag, the Jetson Orin Nano's CPU and GPU temperatures exceeded 70°C during continuous denoising and ASR workloads. After 30 minutes of operation, thermal throttling was observed, confirming the need for an effective heat mitigation strategy.

To manage the thermal load, the team evaluated two approaches:

- **Conduction-based solutions**, such as the use of heat pipes or copper sheets;
- **Convection-based solutions**, involving the creation of airflow intake and exhaust pathways.

To preserve the discreet profile of the wearable device, the team implemented the convection-based strategy. Discreet air channels were integrated into the seams and zippers of the satchel, enabling passive airflow without altering the visual appearance of the bag. This significantly improved thermal stability during sustained operation. User experience was also a focus in this iteration. Several design enhancements were introduced to simplify interaction and maintenance:

- A **slide-out module** was designed to house the Jetson and UPS, enabling straightforward access for battery replacement or debugging.
- A **front-facing interface panel** was added, featuring labeled USB and DisplayPort connectors, along with physical control buttons to streamline user interaction.

- A **0.91-inch OLED display** (Waveshare) was integrated to display real-time power metrics, including voltage, current, and battery status.

Microphone Integration: To enhance speech capture at distance, the team integrated the RODE VideoMic NTG—a highly directional supercardioid shotgun microphone—into the satchel. A custom mounting solution was developed to preserve its acoustic performance while remaining physically unobtrusive. Two mounting configurations were evaluated:

- **Fully Concealed:** The microphone was embedded entirely within the satchel. While discreet, this setup limited effective capture to approximately 2–3 meters in moderately noisy conditions (55 dB).

- **Partially Exposed:** The microphone's front-facing pickup surface was left visible outside the bag. This significantly improved capture range—up to 6 meters—but at the expense of visual discreetness.

To balance these competing constraints, the team implemented a **variable mounting system**:

- The microphone could be extended outward or rotated toward the speaker for enhanced directional capture;

- It could also be fully retracted into the satchel when discreetness was prioritized.

This flexible integration allowed the system to adapt dynamically to environmental and situational demands, ensuring consistent performance while maintaining operational invisibility when required.

Figure 34: The Computing Hardware Stored in a Satchel Bag

The main body of `D'Noise` is housed in a comfortable satchel bag for portability and discreetness (Figure 34). A satchel bag is used as it is a common bag that would not look out of place on the security staff.

E.4 Future Work

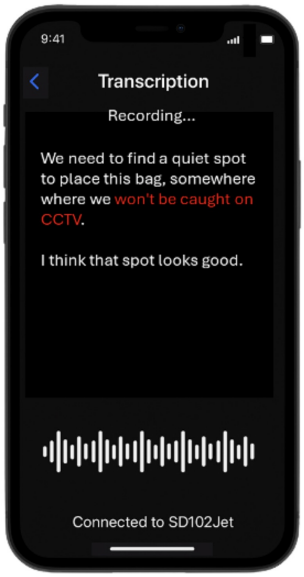


Figure 35: Mock up of an improved handheld transcription device

As mentioned in Section 4.5, the team would consider integrating `D'Noise` with mobile phones for a more intuitive and seamless design. This can enable more features like 2FA or biometric authentication for further security, and help the prototype look more natural and discreet from a third person's perspective.

F Experiments

F.1 Denoising Human Evaluation

Figure 36: Human evaluation experiment set up for the quiet environment at an average ambient sound of 57db.

F.2 D'Noise Stress Test

To validate *D'Noise*'s suitability for deployment by security personnel, a controlled stress test was conducted to evaluate its sustained performance over a 3-hour period. This duration was chosen to reflect a realistic operational segment within a typical 6- to 12-hour shift, during which personnel are expected to take intermittent breaks. The prototype was designed with modular battery replacement in mind, allowing for straightforward hot-swapping after approximately 3 hours of use. Figure **??** presents the results of the system-level endurance test.

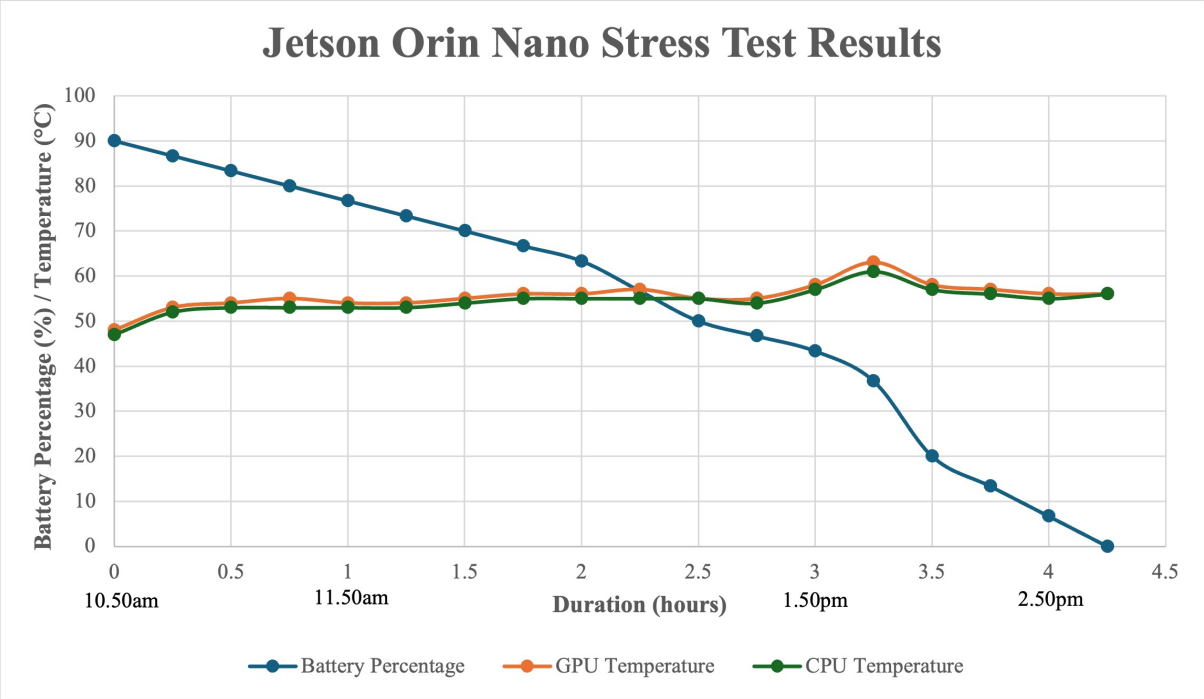


Figure 37: A plot of the battery percentage and temperature against the number of hours the Jetson Orin Nano could operate in the stress test.

Stress Testing Methodology: The stress test was conducted outdoors under shelter at SUTD Building 1, Level 6, on a moderately sunny day. Ambient environmental noise primarily originated from nearby road traffic. The complete system—including the handheld display unit and the satchel-based Jetson Orin Nano processing module—was placed on a table. A Bluetooth speaker was positioned nearby to simulate continuous speech playback. Two thermocouple probes were used to capture thermal behavior:

- **User-facing probe:** Placed between the inner lining of the satchel and the table surface, simulating skin contact during actual wear.
- **Ambient probe:** Left exposed to measure environmental air temperature.

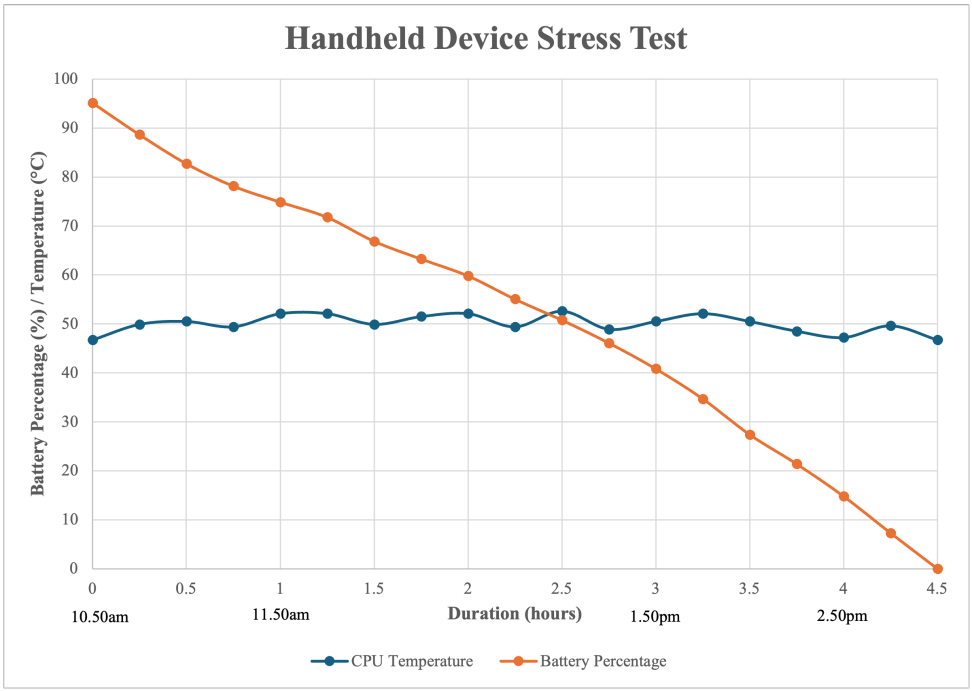


Figure 38: The results of the stress test for the handhel device, plotting the battery percentage and temperature against time

Test Objectives and Evaluation Metrics: The following parameters were monitored throughout the test to assess the system's operational reliability:

- **Battery Life (Processing Unit):**
  - Continuous inference on the Jetson Orin Nano running both RNNoise (denoising) and Whisper (transcription) models.
  - Time to depletion under sustained workload.
- **Thermal Performance (Jetson Orin Nano):**
  - CPU and GPU temperatures were recorded to verify that they remained below thermal throttling thresholds, ensuring stable processing throughput.
- **Battery Life (Handheld Display Device):**
  - The Raspberry Pi Zero 2 W powered the handheld transcription interface, and its battery discharge curve was logged concurrently.
- **Satchel Surface Temperature:**
  - User-facing surface temperature was monitored to assess potential thermal discomfort over time.

Data Logging Procedure

Processing Unit (Jetson Orin Nano): A Bash script interfaced with the Waveshare UPS to log CPU and GPU temperatures, and battery voltage at 15-minute intervals. These metrics were used to track thermal stability and battery depletion under realistic use.

Handheld Device (Raspberry Pi Zero 2 W): A separate script accessed battery data from the onboard fuel gauge and logged it at the same interval, ensuring consistent sampling across system components.

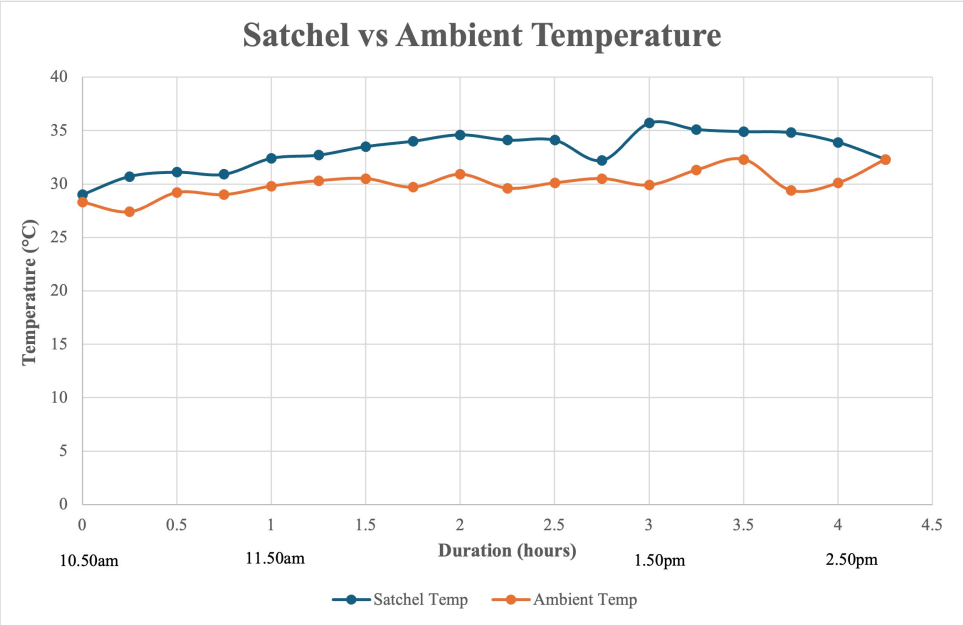


Figure 39: Temperature of the satchel against the ambient temperature

Thermal Measurements: Thermocouple data for both ambient and satchel-contact temperatures were recorded at 15-minute intervals to allow comparison of internal and external heating trends.

G User Validation Questionnaire

This appendix details the responses of the two security guards for the questionnaire and the and the interview conducted after the tests detailed as in Section 4.4.

Section 1: Comfort and Usability

- **How comfortable was the device to wear?**
  - Very Uncomfortable
  - Uncomfortable
  - Neutral
  - Comfortable
  - Very Comfortable

(Participant Responses): ”Very Comfortable”, ”Comfortable”

- **Did the headgear and satchel bag feel secure while walking or moving?**
  - Not Secure
  - Slightly Secure
  - Moderately Secure
  - Very Secure

(Participant Responses): ”Moderately Secure”, ”Very Secure”

- **Were there any parts of the device that caused discomfort or irritation?**
  - Yes (Please describe: _____)
  - No

(Participant Responses): ”No”, ”No”

Section 2: Design Feedback

- **Were there any difficulties in putting on or removing the device?**
  - Yes (Please describe: _____)
  - No

(Participant Responses): ”No”, ”No”

- **How would you rate the weight of the device?**
  - Too Heavy
  - Slightly Heavy
  - Just Right
  - Too Light

(Participant Responses): ”Just Right”, ”Just Right”

- **Were there any features or design elements that you particularly liked?**

(Participant Responses): ”Yes”, “Noise cancelling justto focus on the person.”

- **Do you have any suggestions for improving the design?**

(Participant Response): “Make with no wires so that it won’t look suspicious.”

(Participant Response): “2FA for the design if it was stolen/lost or a more wet weather design”

Section 3: Workflow Integration

- **Did the device interfere with your ability to perform routine tasks?**
  - Yes (Please describe: *”while suspicious looks like person”*)
  - No

(Participant Responses): ”Yes”, “Yes”

- **How well do you think this device could fit into your daily workflow?**
  - Not Well
  - Slightly Well
  - Moderately Well
  - Very Well

(Participant Responses): ”Very Well”, “Slightly Well”

- **Do you think this device would improve communication in noisy environments?**
  - Yes
  - No
  - Not Sure

(Participant Responses): ”Yes”, “Yes”

Section 4: Additional Feedback

- **Is there anything else you’d like to share about your experience with the device?**

(Participant Responses): ”No”, “No”