Research Project in Mechatronics Engineering

Final Research Report

Capturing speech audibility in classrooms through machine learning techniques

Hugo Doughty

Project Report ME035-2026

Co-worker:

Supervisor:

12 January 2026

ENGINEERING
DEPARTMENT OF MECHANICAL
AND MECHATRONICS ENGINEERING

ME035-2026

CAPTURING SPEECH AUDIBILITY IN CLASSROOMS THROUGH MACHINE LEARNING TECHNIQUES

Hugo Doughty

ABSTRACT

Speech audibility plays a critical role in children's learning, yet modern New Zealand classrooms introduce acoustic challenges, known as Flexible Learning Environments (FLEs) [1–4]. These open-plan rooms encourage collaboration and co-teaching but significantly increase background noise. This reduces the signal-to-noise ratio (SNR) below the levels required by New Zealand guidelines [5] and, more importantly, hinders children's ability to learn. Children need an SNR above a minimum threshold; if that is not met, previous research has shown they cannot hear the teacher clearly and struggle to learn effectively [6] [7] [5]. This highlights the need for a reliable way to estimate SNR in classrooms.

Previous research at the University of Auckland (UoA) identified consistently poor SNRs in FLEs, prompting the development of a custom Raspberry Pi–based device to estimate classroom SNR in real-time while maintaining privacy for teachers and students [8] [3] [4].

This project compares a baseline voice-activity-detection (VAD) method and a neural-network approach for estimating classroom SNR [8] [4]. A newly implemented Neural Network (NN) process has been designed to recognise and detect the teacher's voice from a mix of noise in the classroom [9] [10]. Both methods were tested in $10\,\mathrm{dB}$, $0\,\mathrm{dB}$, and $-10\,\mathrm{dB}$ SNR conditions using a controlled acoustic laboratory and calibrated speakers to simulate classroom environments. The results showed that, for all conditions, the NN estimated SNRs closer to the verified ground-truth values than the baseline. To quantify this behaviour, the NN achieved similarity coefficients with the ground truth ranging from 0.3 to 0.7, while the baseline produced similarities between 0.2 and 0.25 (out of 1). The system was tested offline against verified ground-truth SNR values to ensure data reliability.

This validated device provides an accurate and privacy-compliant means of estimating classroom SNR in real-time, addressing the acoustic challenges of modern New Zealand learning environments.

Table of Contents

Acknowledgements

I would like to express my sincere gratitude to the following individuals for their continuous support and contributions throughout this project, without which it would not have been achievable.

- , my project partner, for his consistent effort, valuable insights, and collaboration. It has been a pleasure working alongside him.
- , our project supervisor, for her continuous guidance, constructive feedback, and encouragement. Her expertise has been invaluable to the success of this project.
- , for his technical assistance and commitment, providing the necessary insight and tools to implement and test the Neural Network method.
- , for his support with the technical components and for providing the essential equipment required for testing.

The collective contributions of these individuals have been instrumental to this project, and I am deeply grateful for their time, support, and involvement.

Glossary of Terms

Audibility	How easily a listener can hear and understand a sound or speech.
Signal-to-noise Ratio	The difference in loudness between the desired sound (signal) and background noise.
Ground Truth	Verified reference data used to check the accuracy of other results.
Flexible Learning Environment	An open and adaptable classroom designed for shared teaching and group learning.
Signal Separation	The process of isolating individual sounds from a mixed audio recording.
Voice Activity Detection	A method for detecting when speech is present in an audio signal.
Time Difference of Arrival	A signal-processing technique that determines the location or direction of a sound source, by measuring the differences in the time it takes for the sound to reach each microphone in an array.
Delay-and-sum Beamforming	A technique that improves signal clarity by aligning audio from multiple microphones. Sounds from a specific direction will increase in clarity, while noise from others will cancel out.
ANOVA	A statistical test used to analyse the differences in group means.
Masking	The process of hiding a signal.
Energetic Masking	The masking of a weaker sound signal due to an overlapping louder sound.
Linguistic	Relating to language or the study of language
Second Language	The language learned after one's first (native) language
deciBels	A unit or quantifying sound and its loudness, (even across frequencies in this report)
Auditory	Relating to hearing or the sense of hearing
Reverberation Time	The time for sound to decay by 60 dB

Abbreviations

MoE	Ministry of Education
FLE	Flexible Learning Environment
SNR	Signal-to-Noise Ratio
VAD	Voice Activity Detection
NN	Neural Network
TDOA	Time Difference of Arrival
ANOVA	Analysis of Variance
dB	deciBels

1. Introduction

A successful and efficient learning experience in schools heavily depends on a child's ability to perceive and understand what is being taught verbally [3] [11]. To achieve this, high speech audibility and clarity are required, obtained through minimal background noise. However, classrooms exhibit large amounts of noise, especially with young children, as they are more restless and energetic than older students [1] [8]. Noise in a closed space can drown out the original signal attempting to be transmitted — in this case, the teacher's voice [8] [11] [12]. Children have a harder time naturally filtering out unwanted noise when listening, as they have had little experience for their brains to learn how to do so [1] [2] [11] [13]. For children exhibiting hearing difficulties [11] or for whom English is a second language, the challenge is even greater.

The Ministry of Education (MoE) has been transitioning all New Zealand classrooms to FLEs, also known as Modern Learning Environments. These open-plan rooms are designed to hold more than one class and several teachers at a time [14]. However, there is concern that these new classrooms do not meet the international guidelines for acoustic performance. A child's ability to learn greatly depends on the acoustic properties of their learning environment; thus, a crucial concern arises [4] [8].

In 2023, UoA developed a device to assess key acoustic properties of classrooms, including factors that influence how easily students can hear speech [8]. The device estimates the SNR in real-time while maintaining privacy constraints. However, while the current method can estimate SNR, its accuracy under controlled conditions and across different estimation methods has not been thoroughly evaluated [3] [4].

Speech transmission describes how well a room can accentuate conversation, which can be measured in several ways. Acoustic performance splits into audibility and intelligibility, depending on whether the sound reaches the listeners and whether the listener can distinguish words in the sound. This study focuses on audibility, which is best represented by the SNR, quantifying how strong the desired signal (the teacher's voice) is compared to the background noise [6] [12].

Building on this foundation, this project evaluates the performance of the UoA's SNR device by comparing two signal-estimation methods: the existing baseline algorithm, which utilises VAD and signal separation, and a newly implemented NN approach [4] [8]. Both were designed to detect teacher voices in noisy rooms. These methods were tested under controlled acoustic conditions to assess their usability for real-time classroom monitoring [9] [10].

2. Literature Review

This review highlights gaps in current research, forming the basis for the project's scope and objectives by identifying previously implemented projects and what has and has not yet been achieved. This section also critically evaluates the effectiveness of previous studies and identifies how their limitations directly motivate the current investigation.

2.1 Flexible Learning Environments

New Zealand classrooms have shifted in design in recent years, moving away from single-teacher rooms to FLEs. Driven by the MoE, a nationwide plan aims to modernise learning

spaces, mandating that all new and refurbished classrooms adopt the new open-planned FLE model [1] [2] [14]. The open-plan layout containing moveable sections and shared teaching spaces characterises the FLEs. It encourages adaptability and supports collaborative work between students and teachers. These factors improve flexibility, allowing teachers to co-deliver lessons across interconnected rooms.

To achieve this open-planned environment, the internal walls are sliding doors to move and connect the classrooms easily [1] [2]. The furniture is unsymmetrical, and the timber used for the walls is light. This results in an environment that may let unwanted external noise travel through or accentuate the internal noise further. Either way, the room layout severely hinders the desired acoustic properties. Wilson [1], and Dodd and Legget [2] both highlighted that while this design supports collaboration, it weakens sound isolation between teaching zones. These findings align with observations by Twinn [4], who measured consistently poor SNR in these spaces.

This reduced acoustic control directly affects the SNR in a classroom. As speech and background noise mix across these shared spaces, the teacher's voice often becomes less able to be perceived by the students, lowering audibility.

Compared to single-cell classrooms, which are enclosed and acoustically contained, FLEs heavily rely on the teaching coordination to maintain the students' focus. In traditional classrooms, sound levels and clarity were easier to manage and control due to the physical isolation of classes [15]. In contrast, the open nature of FLEs introduced exposure to higher background noise and reduced audibility, as shown by Wilson [1], and Dodd and Legget [2], who both noted that removing internal partitions decreases audibility.

As discussed earlier, the FLE features encourage collaboration and flexibility; however, the spaces introduce new acoustic challenges that can directly interfere with communication and education. It is crucial to understand how the noise behaves in these environments and how it affects teaching and learning [3] [4] [8]. The following sections explore the nature of noise in learning environments

Accurately quantifying the SNR in these classrooms allows researchers to identify if there are any areas where these FLEs acoustically succeed or fail, forming the foundation of this project's objective to assess methods of measuring SNRs in these learning environments.

2.2 Audibility

Understanding classroom acoustics first requires defining how sound is perceived and quantified. This section introduces the concept of audibility, which describes a listener's ability to hear a desired signal, and explains the process of measuring it with the SNR.

2.2.1 What is Audibility

Audibility is the quantity to which listeners can hear a desired audio signal, but it is not necessarily understood. Factors such as Energetic Masking and general background noise can drown out desired signals [6] [13]. This creates a situation where the listener can hear the sound produced, but may not necessarily understand what is said exactly. In a classroom environment, the teacher's voice must be able to carry to the children, especially young children, as they have difficulty processing words alone [4] [11].

It may seem like the content that the children are learning and hearing is the sole focus, but it is not. Children are at school to learn, but they are simultaneously developing a more

hidden skill: their ability to listen and digest vocal signals, breaking them down. It is a skill that comes naturally but takes time, and in the early stages of development, is essential to master [10, 13]. Audibility is vital as, without it, the learning process is interrupted. Audibility's importance justifies SNR's usefulness through its ability to measure it. The following section describes what SNR is and how it can represent audibility. [3] [8] [16].

2.2.2 Signal-to-Noise Ratio

The SNR represents the sound level ratio between the target speech signal and the background noise, usually expressed in unweighted decibels (dB(Z), or dB). In a classroom context, it quantifies how well a listener can perceive the teacher's voice over the background sounds [12] [15] [16]. These include student chatter, ventilation, or external noise. A higher SNR indicates a clearer listening environment; a low or negative SNR reflects undesirable listening conditions, with much louder noise. An SNR of +15 dB means that the teacher is 15 dB higher in level than the background noise; this scenario is typically considered an optimal condition [6]. In contrast, an SNR of 0 dB indicates that the speech and noise are equally loud, resulting in poor audibility.

SNR, being the main quantifiable characteristic of a classroom's acoustic learning capabilities, means it is helpful to have an actual means of gathering it [3] [8] [11]. There is limited previous research on any statistics, data, or even accurate methods for this SNR [3] [4] [8]. Existing research is primarily carried off the back of predetermined and simulated learning environments, providing little insight into real-time behaviour [8] [16]. Therefore, it is of great interest to develop and test an accurate way to create such a device to obtain the SNR of any classroom [8].

Previous studies have identified SNR as one of the most crucial acoustic properties to quantify audibility. It is directly related to audibility as it provides a measurable value of how identifiable the desired speech signal is from the surrounding noise [4] [11] [15] [16]. While audibility describes a listener's ability to detect and hear sound, SNR quantifies it.

SNR is not uniform across the classroom; students closer to the teacher typically experience higher SNR than those seated at a distance or near noise sources [4] [11] [15], making position an essential consideration for the use of SNR.

2.3 Effects on Teaching and learning

While audibility and SNR describe how clearly the speech can be heard, it is equally as important to understand the impact of poor acoustic conditions, expanding on how SNR can be spatially affected. This section examines how noise exposure affects both students and teachers, and how they both can affect the audibility of a classroom.

2.3.1 Noise Exposure on Students

In classrooms, specifically FLEs, background noise directly affects the audibility of the teacher's speech [4] [11] [12] [15]. Student chatter, ventilation, or external sources introduce noise and decrease the SNR, making it difficult for the students to detect the speech signal from the teacher [11] [15]. SNR can also vary across the classroom, depending on the student's relative location to the teacher and noise sources [4] [15].

Research indicates that SNR in FLEs can often fall below the levels for clear audibility [4]. Reports show SNRs as low as 3 to 6 dB during active learning periods, compared to 12 to 15 dB in the traditional single-celled classroom [4]. The lower SNRs indicate the

conditions where the background noise partially masks the speech signal, limiting students' ability to hear the teacher [11] [12] [15].

2.3.2 Noise Exposure on Teachers

Teachers play a crucial role in maintaining a high SNR in classrooms; their speech serves as the sound signal for the students [1] [11] [12] [15]. Background noise reduces SNR for the students and challenges the teacher in maintaining audibility [12] [15]. It is often natural for teachers to attempt to compensate by raising their voice or projecting their speech louder [1] [12]. However, even with an increased sound signal, excessive noise may still drown it out, especially if the noise increases in response [12] [15].

The impact of noise exposure is especially evident for FLE classrooms, where the open-planned layout requires the teacher to move around, changing the source position and thus the SNR [4]. Some students may hear the teacher clearly if they are closer, while others may experience reduced audibility [4] [15] [11]. The teacher's positional variance, and how noise contributes to it, highlights the importance of measuring SNR across the space [8] [4].

Maintaining adequate SNR is therefore crucial for teaching and learning [15] [11] [4]. The following section outlines the SNR thresholds required for children to perceive speech reliably in a classroom environment.

2.4 Children Learning Requirements

Children are at the dawn of their learning journey; they must have proper learning environments to rely on. It is the foundation of everything they will base on later in life, cognitively aware of it or not; their brains are learning how to receive human-generated sentences [11]. Their ability to perceive and comprehend speech in noisy environments is substantially different from that of adults, due to their still-developing auditory systems. Unlike adults who use linguistic cues to compensate for degraded speech signals, children are less able to fill in missing segments (due to masking). Therefore, this requires listening conditions with a higher SNR to achieve the same level of understanding [7] [17].

Research consistently proves that adults can understand speech signals adequately with SNRs from 0 to 6 dB, whereas children typically require much higher levels, research indicates that children require an SNR approximately +15 dB for clear speech perception [6], underscoring the sensitivity of young listeners to noise. The difference highlights the significance of a high SNR environment for classrooms. The classroom should be acoustically sound for every child and possible situation - not only should a child with basic hearing capabilities be able to listen to their teacher while sitting quietly on a mat [1], a student with hearng difficulties sitting amongst a noisy table of their peers must also be aple to comprehend the signal. Previous papers show that children with moderate hearing loss struggle even more with differentiating signals from noise, typically requiring higher SNRs to achieve the same levels of speech perception [11]. Their hearing loss dilutes their linguistic input. As stated by Meijer (2025), children with hearing loss or those learning a second language need SNRs of 20 dB to process speech efficiently [11]. These elevated requirements represent the importance of maintaining favorable acoustic conditions in learning environments.

Classroom acoustic standards, such as AS/NZS 2107:2016 and ANSI S12.60, recommend maintaining SNRs of at least 15 dB for adequate audibility for children [5] [18]. These standards reinforce Meijer's claim on children's SNR requirement. Increased background noise and reverberation times cause FLEs to fall below this standard. When SNRs are sub-

optimal, students experience reduced speech recognition and lower academic engagement.

Therefore, children, especially those with hearing loss or learning a second language, require a higher SNR than adults, ranging from +15 dB to +20 dB. This range highlights the importance of proper acoustics standards and the study of FLEs.

2.5 Device

In 2023, UoA developed a device to measure and obtain SNR values [8] —it has since been used in multiple reports [3] [4]. The device contains two Boya BY-WM4 PRO wireless microphones, with their receivers connected to an AudioInjector Octo shield, allowing them to be synchronised and received by an individual system. The Raspberry Pi handles this audio input, and the software that handles the SNR calculation and audio security is stored [3] [4] [8].

For the use of the device in public settings, the device must meet privacy considerations, especially surrounding storing audio [3] [8]. The audio recorded in a classroom environment can not be stored or accessed; therefore, the device must calculate the SNR in real-time without offline processing [3] [8]. Offline processing refers to the storing of audio directly from the real-time classroom,

Microphone 1 is attached to the class teacher and only detects if the teacher is talking, and Microphone 2 is connected to a student. Signal separation only needs microphone 2's stream, which the next section discusses, effectively meaning the device is single-streamed [3] [8].

Figure 1 shows that the device setup consists of a Raspberry Pi, two Bluetooth microphones and receivers, and the software used to calculate the SNR (stored on the Pi) [3] [8].

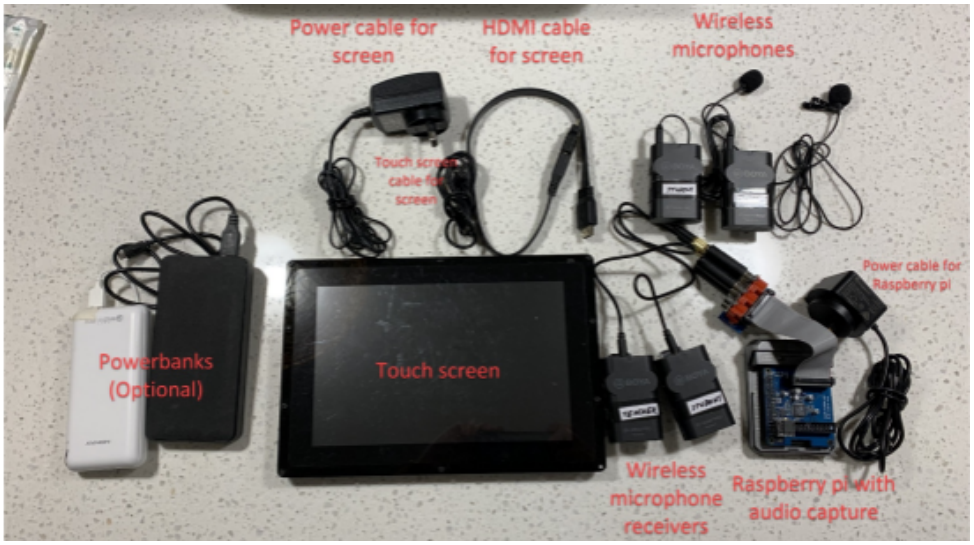


Figure 1 Device and all components used to measure, calculate and display the SNR estimations, sourced from the user manual provided by the University of Auckland [19]

The software in the device constantly takes the two microphone streams as input, treating them as teacher and noise [3] [8]. Obtaining the SNR for them is mathematically easy; it is a ratio of signal and noise, done by measuring their respective magnitudes and dividing them [8], as seen in Equation 1. However, the ratio assumes that the signal (teacher) and noise (students) sources are independent and do not interact. Independence is not the case in a classroom environment, where the teacher is talking amidst a group of noisy students,

and therefore, both measurements will pick up the other [3]. The designed device must separate the two signals to calculate the SNR successfully [4] [8].

$$\mathrm{SNR}_{\mathrm{dB}} = 10 \log_{10} \left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}} \right). \tag{1}$$

A threshold determines if the teacher is significantly louder than the students by finding the difference between the teacher's and the students' magnitudes [8]. This process is VAD and explored further in the methodology [3] [8].

The Baseline code is the current method of calculating the SNR value, which uses VAD and signal separation [8]. A potential and new method has been addressed through analysis of the Baseline code [4]. A machine learning technique to detect and estimate the teacher's signal power in the noisy classroom could accurately calculate an active FLE's SNR [3] [4]. This report explores both methods and their accuracy, which are explained further in the methodology.

2.6 Previous Cases of Signal-to-Noise Ratio Testing

Recent research at the UoA has begun addressing the lack of real-world data on New Zealand classrooms' SNR. Two dissertations, by Benjamin Twinn (2024) [4] and Kate Westbrooke (2024) [3], formed part of a more exhaustive research investigating speech audibility in FLEs, using the exact custom-built Raspberry Pi System designed to capture real-time SNRs.

Twinn focused on identifying the factors determining SNR in a single FLE in Auckland, containing four teachers and 88 students. The study used quantitative sound level data and qualitative interviews with the classroom teachers [4].

In Twins's experiment, the two student microphones collected readings from the student's position near the center and back are referred to as student 1, and 2 respectively of the classroom during 10 school days. The results found that while the noise levels met the MoE guidelines for group work in occupied classrooms, the SNRs recorded were consistently below the +15 dB requirement for clear speech and learning for students. With median SNRs from the two student microphone positions resulted in -1.34 dB and -0.87 dB, which are both well below the threshold [4]. The recorded noise was around 61-62 dB, which is within the MoE guidelines, showing that SNR values remained poor even with acceptable noise levels. The position of the students in the classroom also affects the perceived SNR, where the SNR decreases further near the back of the class [4].

Westbrooke expanded this work by examining how subject, teacher, and space influenced the SNR in FLEs using the same Raspberry Pi setup [3]. She compared open and closed teaching spaces (FLE vs an enclosed breakout room) to observe the room's effect on SNR. The results obtained both highlighted and confirmed those from Twinn's dissertation; the average SNR varied from -1.7 to -0.6 dB in the open teaching spaces, and the closed teaching spaces had an increase at +2.08 dB, which is an improvement but still well below the +15 dB threshold [3].

Both Twinn and Westbrooke's studies confirm that SNR is well below classroom and basic learning requirements for children in open teaching spaces. Even the breakout rooms, expected to have improved activity, perform poorly and are still under the threshold. Both dissertations provided strong evidence that modern New Zealand FLEs do not create acoustically adequate environments for speech audibility despite meeting noise-level guidelines [3] [4].

Their results and outcomes emphasise the importance of proper testing of the SNR in New Zealand classrooms and ensuring the validity of the measured data. The version of software used on the device in Twinn and Westbrooke's studies uses the Baseline code. All SNR values produced are relevant to VAD and signal separation processes. To ensure the validity of the process and thus the data produced, the NN method should be explored and tested.

2.7 Neural Network Applications in Speech Enhancement

Deep learning methods have shown strong performance in speech enhancement tasks [9] [10]. NNs, especially Deep Neural Networks (DNNs), can separate desired speech from noise. Unlike conventional algorithms that use energy thresholds or statistical assumptions, NN can identify human speech patterns directly from raw and unprocessed audio.

NNs are trained on large datasets that contain mixtures of speech and background noise. The NN learns the unique spectral and temporal components that distinguish human speech from noise by providing pairs of noisy mixtures and their respective clean speech signals. This process allows the NN to suppress any unwanted noise while maintaining the integrity of the speech structure [9] [10].

Kinoshita (2017) proposed an NN-based spectrum estimation that combines a DNN into a Weight Prediction Error (WPE) algorithm instead of the WPE alone [9]. Through testing, the DNN-WPE method outperformed the WPE algorithm in nearly every test condition, especially in real-time estimations [9].

The NN, therefore, could have enormous potential for implementing the SNR device employed in this project and Twinn and Westbrooke's dissertations. The previous studies prove that NN can increase accuracy, especially in real-time, therefore proving relevant to the project [9]. Applying this method could provide a novel and presently unimplemented solution for estimating the SNR in New Zealand FLEs.

2.8 Literature Review Conclusion

The studies reviewed highlight the impact that background noise and the environment have on the audibility of a classroom. They emphasise the importance of audibility, its correlation with a child's ability to listen and learn, and how an SNR value can quantify it. The studies show a gap in research, where few reliable previous studies have designed a device for real-time processing [3] [4] [8]. The current method implemented by UoA needs testing and further implementation [3] [4] [8]. In summary, although recent studies have provided quantitative evidence of inadequate SNR levels in New Zealand FLEs, they also highlighted limitations in the measuring process [3] [4]. Current devices rely on a baseline thresholding algorithm that may struggle to capture SNR accurately in noisy environments [3] [4]. Alternative estimation methods can be employed to improve upon this and ensure confidence in SNR data, including a NN approach [9] [10].

3. Research Scope and Objectives

This project aims to evaluate the performance of the SNR device in a controlled environment, exploring and testing two options, Baseline and NN, for quantifying SNR. The device and selected software must be robust to ensure the values obtained are true and can successfully estimate the audibility of a noise-varying learning environment. Consistent performance is essential, especially in FLE spaces, where noise levels and locations continuously change throughout the day.

After validating the device's performance, the primary objective will be to use it in FLE classrooms in New Zealand, with participants' consent, recording the SNR during teaching hours. The device may provide information on the FLE rooms to evaluate their acoustic properties against childrens' learning requirements.

4. Methodology

The following methodology outlines how the device was tested and introduces the methods and testing metrics used for quantifying SNR.

4.1 Device Used

The device currently uses only two of the three microphones. Microphone 1 is attached to the teacher and detects if they are talking, while Microphone 2 is positioned near the students to capture the classroom audio. The methods of real-time SNR calculation only need Microphone 2's stream, which the next section discusses, effectively meaning the device is single-streamed.

However, the calculation process must separate the teacher from the noise audio. The SNR assumes that the target audio (teacher) and noise (students) sources are independent and do not interact. Independence is not the case in a classroom environment, where the teacher is talking amidst a group of noisy students. Therefore, both measurements will pick up the other. The designed device must separate the two signals to calculate the SNR successfully.

The methods of obtaining the SNR use a threshold to determine if the teacher is significantly louder than the students by using VAD, which finds the difference between the teachers' and the students' magnitudes. The two methods of calculating SNR are a baseline algorithm and a NN.

4.1.1 Software Implementation

To recap, the device uses two microphones that receive the teacher and student audio, which leak into the other due to a lack of an independent environment. The microphones connect to a Raspberry Pi through an Octo-Capture, where the processing and calculation of SNR occurs; the code/method to estimate SNR is within the Raspberry Pi.

The device uses a baseline version, which the following section discusses. However, a new method containing a NN could provide a new means of obtaining SNR.

The teacher's microphone does not accurately represent the audio; the purpose is to continuously measure if the teacher is talking, which the following paragraph explains. The goal is to calculate the SNR experienced by students; therefore, the microphone positioned at the students should measure both the teacher's audio power and the noise. An issue arises where the student microphone is recording both the teacher and the noise simultaneously and needs to be separated.

The software must separate the teacher and noise audio signals, bringing VAD into play. VAD measures the difference in power at the teacher and student microphones and assesses it against a threshold. VAD detects teacher-active (TA) regions and teacher-inactive (TI) regions.

VAD is where the similarity of both methods stops; they both use it, but the difference in their calculation process is in the way of obtaining the actual signal power of the teacher.

The current issue is that the teacher and noise-only signals are required to calculate the SNR. However, as the speaker and noise sources are in the same room, they are not independent. The noise audio contains the teacher's remnants, and vice versa.

The noise-only signal can be easily found with VAD in the TI regions; however, the teacher signal can not, as the threshold in TA regions does not account for background noise, resulting in an audio mixture containing both teacher and noise. The following section highlights each method, Baseline and NN, their process of obtaining the teacher signal, and how they calculate SNR.

4.1.2 Method 1: Baseline

The Baseline method uses two audio queues for mixture and noise, respectively. A mixture is a combination of the teacher and the noise audio. Using the data received at the student microphone, if VAD detects a TA region, the mixture queue stores the audio period data. In contrast, in a TI region, the data is stored in the noise-only queue, effectively obtaining a mixture and noise signal. When both queues are full, there is enough data to calculate an SNR; therefore, signal separation occurs. Signal separation subtracts the noise-only component from the total mixture to estimate the teacher's signal power. With an estimate of the average teacher signal and an average from the noise queue, an SNR can be calculated by dividing the two.

Mathematically, the Baseline method has the potential to estimate the SNR accurately; however, the use of a queue-based system leaves it vulnerable to losing temporal representation and precision. The Baseline only produces an SNR when both queues are full, meaning that if the threshold is not high enough, causing the method to assume a constant TA state, data will only be stored in the mixture queue, leaving the noise queue empty. The queue mismatch causes two major issues: the primary issue is that SNR will not be calculated until enough noise has been stored, meaning fewer SNR values will be obtained, and the audio used in the calculation process is old data. Additionally, when a queue is filled for too long (when the other queue is not filled), old audio chunks will be removed to make space for new data, permanently deleting data and losing representation for audio.

The issue of data integrity encourages the development of a new method using a NN to find the teacher's power.

4.1.3 Method 2: Neural Network

Another method of obtaining SNR still uses VAD, but instead of subtracting mixture and noise powers, the device uses a NN, designed by Dr Felix Yan, from Victoria University of Wellington. It requires a five-second recording of the desired teacher's voice, with the permission of the respective teacher.

The software uses the snippet to extract the respective power in the mixture of audio, effectively estimating how much of the recorded audio is the teacher, and the rest is background ambient noise. The NN process eliminates the signal separation and the averaging of audio chunks and produces an SNR using the two estimated powers for the teacher and the noise.

The NN approach avoids averaging and queue delays, providing a more immediate estimate of the teacher's signal from the mixture; however, being computationally expensive, it highlights the risk that it may challenge the Raspberry Pi's real-time processing capabilities.

Both methods of SNR calculation have potential and need a fair and controlled test environment to assess their accuracy. The following sections address this.

4.2 Testing Environment and Setup

Testing the device and the embedded software requires a controlled acoustic environment. The UoA's acoustic building contains an underground lab with a listening room designed for this purpose, shown in Figure 2. The room is cut off from external noise and equipped to mimic the internals of a standard household, including furniture, carpeted floors, and even curtains in place of windows. While not identical to an FLE, the room is certainly proficient in checking the results of the SNR device.

Figure 2 University of Auckland's acoustic lab, mimicking a household room and containing a 16-speaker array.

The listening room features a 16-speaker array arranged in a spherical formation, with each speaker pointed inwards to allow surround sound simulation. Every speaker is connected to a MOTU 16A amplifier, allowing independent playback of up to 16 audio sources. Each speaker has an individual physical gain scale for the speaker array, requiring calibration to ensure consistent output levels across the array.

Calibration and later testing, was performed using the audio software tool REAPER combined with the MOTU 16A. An external decibel reader was placed in the middle of the array while a noise file sequentially played out of each speaker. The digital gain for each channel was adjusted until the reader indicated a LEQ(Z) of 65 dB, (LEQ(Z) represents the average unweighted sound energy). For the teacher speaker (Speaker 1 at 0°) the file *"ElongatedVowel"* was used to represent a sustained vowel from the teacher. The background noise was the *"Australian Babble"* file, a recording of overlapping Australian news articles.

4.3 Experimental Design

Following Calibration, testing could commence using the *"Northwind"* as the teacher's audio signal and *"Australian Babble"* as the noise source. To evaluate the device's robustness, the middle ring of eight speakers was used, with Speaker 1 (0°) assigned to the teacher and Speakers 2–5 representing noise at 45°, 90°, 135°, and 180°, respectively. And the remaining speakers were ignored due to symmetry.

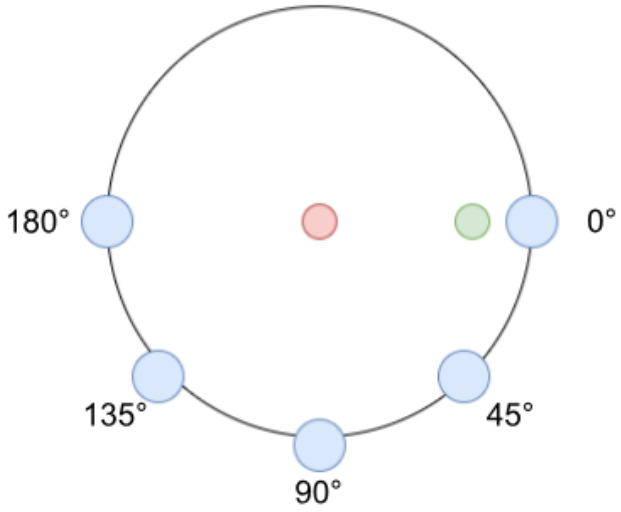


Figure 3 Diagram highlighting the position of speakers used. The large blue circles are speakers, where the speaker at 0°is the teacher. The red and green smaller circles are the student and teacher microphone positions respectively.

The overall SNR of an environment may heavily impact the device's performance; therefore, for each angle, the estimated SNR of the environment was set to $10\,\mathrm{dB}$, $0\,\mathrm{dB}$, and $-10\,\mathrm{dB}$. Adjusting the noise sound level achieved this, while the teacher signal remained unchanged. The range allows for a good understanding of the device's performance and ability to pick up SNRs even in negative environments, which will test how well the device works in noisy or quiet environments.

Each case ran 30 seconds of each file simultaneously, and the device calculated the SNR in real-time. The varying environments tested the device's capability to work with different sound sources and will thus quantify how robust the device and its contained software are at identifying the environmental SNR.

Testing both methods is required, which the next section addresses.

4.4 Data Collection - Real-Time Testing

The two methods of obtaining SNR:
1) Baseline - VAD, signal separation
2) NN - VAD, voice recognition

For each environment, the experiment included the testing and comparison of the Baseline and NN code.

Both code versions produced different and messy SNR data during implementation, sometimes not producing data. Observation showed that the microphones had issues maintaining a connection between the transmitter and receiver. Thus, it was difficult to highlight the area of the problem.

Therefore, the issue caused the refinement of the scope and objective of the project. The new objective was to remove the real-time aspect by recording the audio, running an offline version of the code, and directly allowing the code access to the mixture and noise audio files, eliminating potential errors from the Bluetooth microphone disconnection or any overlooked issues regarding the device.

This method ignores the privacy constraints that are justified only by strict testing. The two versions of code need to be updated and reconfigured so as not to expect real-time input, but two fixed wave files.

4.5 Transition to Offline Testing

The Baseline and NN versions of the code need two audio files: the teacher and student recordings of target and noise signals playing, which is called the mixture audio file, and a noise-only recording. The results are two two-streamed audio files (teacher and student) of the environment with the target audio and background noise playing in one, and only noise in the latter.

These recordings followed the same process as the real-time testing, for every angle, and every SNR environment ($10\,\mathrm{dB}$, $0\,\mathrm{dB}$, and $-10\,\mathrm{dB}$). Resulting in 12 mixture and noise only recordings.

4.5.1 Audio Alignment

However, having offline audio presents a new issue. SNR calculation directly divides the estimated teacher sound level by noise; if any shift or misalignment exists, every SNR value will be off. In some instances, the noise may be considered louder than the mixture, which is technically impossible, as the mixture combines the teacher and the noise. Still, due to the misalignment, this error may occur repeatedly. The resulting signal separation will cause the predicted teacher sound level to be negative, which is mathematically impossible when calculating SNR with a log. An error will occur, causing no SNR value. The misalignment encourages a solution.

The audio file recording process contained a bug that recorded a massive sound spike just before the audio played. Therefore, the bug prevents alignment by searching for the first loud noise in both files. A solution still existed even with this issue, requiring an understanding of the audio file's content. The mixture audio contained the entirety of the noise-only file, but mixed with the teacher audio. It could be aligned if there were a method of finding the noise audio file within the mixture recording. The technique is where cross-correlation comes into play.

4.5.2 Cross Correlation

The study of how an audio shift occurred highlighted the need for the location of the relative file shift. Initially, the assumption was that a set amount would shift the entire file, but observation presented a new error. Due to the bluetooth microphones briefly disconnecting, the recorded audio would be shifted halfway through the recording on some occasions, meaning it was a local shift, not global. Two errors were occurring, a global

shift and a local shift, both possible in the same file. The local error also accentuates the need for reliable microphones.

The global shift was easy to identify and essential to account for first, as it affects the entire file. Applying cross-correlation between the mixture and noise-only audio signals resulted in an array of the "alignment magnitudes". The largest magnitude index, representing the best phase shift for optimal alignment, could be used to align the mixture audio.

The local shifts were different; the audio beforehand was only affected by the global shift (now assumed to be accounted for). The audio afterwards was the focus; a blip in the data inserted a section of space, pushing back the remaining audio.

Each set-lengthened section required a local cross-correlation, producing a relative alignment magnitude. Due to the previous global alignment, if the section is good, then no shift is needed. A blip causes a jump in the required alignment.

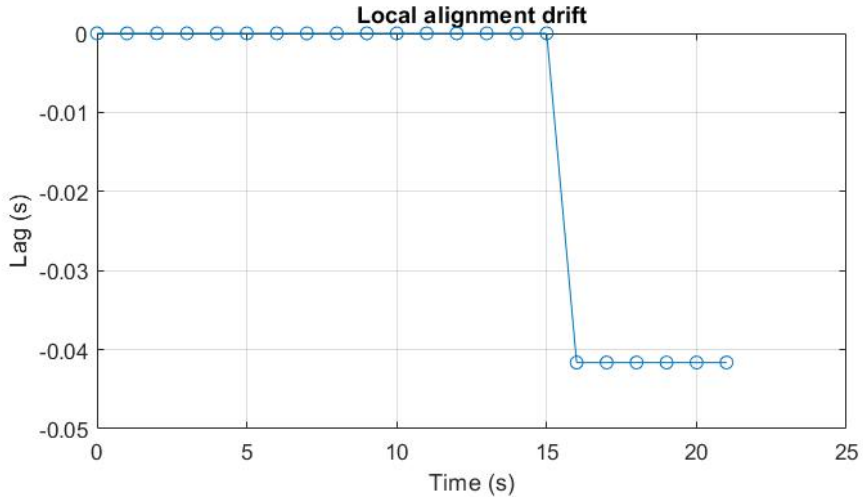


Figure 4 Example of a local alignment fix for a reading, where chunks after roughly 16 seconds had been wrongly shifted forward due to misalignment errors. To fix the chunks are shifted back by the shift value. This is done by comparing the noise and mixture audio chunk by chunk.

As seen in Figure 4, a negative local shift occurs, meaning the code will move the remaining audio back by the shift amount.

Both alignment issues and the severity of misalignment inspired the creation of an audio alignment algorithm. The global and local shift alignments successfully aligned both mixture and noise-only files. Trimming the audio files removed any unwanted sound before or after, including the loud impulse at the start. The files were now able to be used for offline processing.

4.5.3 Code Adaptation - Real-Time to Offline

Offline code is a shifted version of the two methods such that they do not rely on real-time input data. A fixed input file is used for consistent testing.

The audio files were now accurate and aligned. With the help of Dr Felix Yan, who developed the NN version, both versions of the code underwent adaptation to offline processing. The aim was to run the code on a different device to eliminate any potential inconsistency of the Raspberry Pi device or setup.

The simplified code is easier to understand and debug, removing the audio contexts and replacing them with audio file input.

The offline Baseline and NN versions of the code both received only the mixture audio file. The Baseline process to calculate SNR is the same as the real-time version, using signal separation and VAD, which means using teacher and student audio file streams. Whereas the offline NN does not need VAD, a snippet example can estimate the teacher values, meaning only the student recording of the mixture is required. For testing purposes, we used a period of 200 ms. With a sampling frequency of 48KHz, this results in 9600 samples.

4.5.4 Ground Truth Definition

For the ground truth, testing purposes require an additional sound file of only teacher-only audio too. audio file of only noise audio too. Therefore, two files, "Mixture" and "NoiseOnly", are used, which represent the two environments from the teacher and student microphones.

The real-time code has been shifted offline, breaching privacy concerns, but enabling in-depth testing of each version of code. The results will highlight the more accurate code, but what defines the accuracy? The overall SNR environment of each testing case is a predicted estimate; during conversation, the speech level fluctuates on a continuous scale. Removing all and any potential issues will generate a truth, which represents the true nature of the instantaneous SNR.

The ground truth wont use VAD or NN, it will divide the noise power from the teacher power, whether estimated or true, for every sample. The first method of ground truth uses signal separation to allow the use of a mixture audio, which contains both teacher and noise, but more specifically, the effect that either has on the other. The second assumes independent sound audio, skipping signal separation and directly using a teacher-only and noise-only sound file. Calculating the exact SNR from the sources.

Through testing the Isolated Signal Ground Truth performed better.

4.5.5 Offline Data Gathering

Now that all methods and their required inputs for testing had been established, data collection could begin.

To gather the data was quite simple; the respective audio files were input into the four different versions of the code: Baseline, NN, Signal-Separated GT, and Signal-Isolated GT. There were 12 environmental conditions and four versions of code, so 48 data outputs were expected. However, the Signal-Isolated GT does not and will not change based on the testing angle, as the noise file was recorded from the middle with a unidirectional microphone, therefore producing equal readings for each angle. The results of the collected data are presented in the next section.

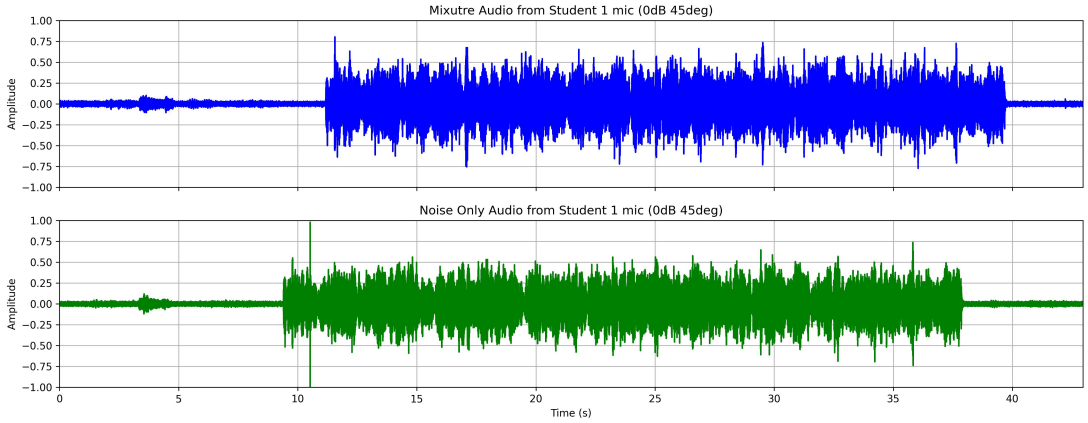
5. Results

5.1 Audio Alignment Verification

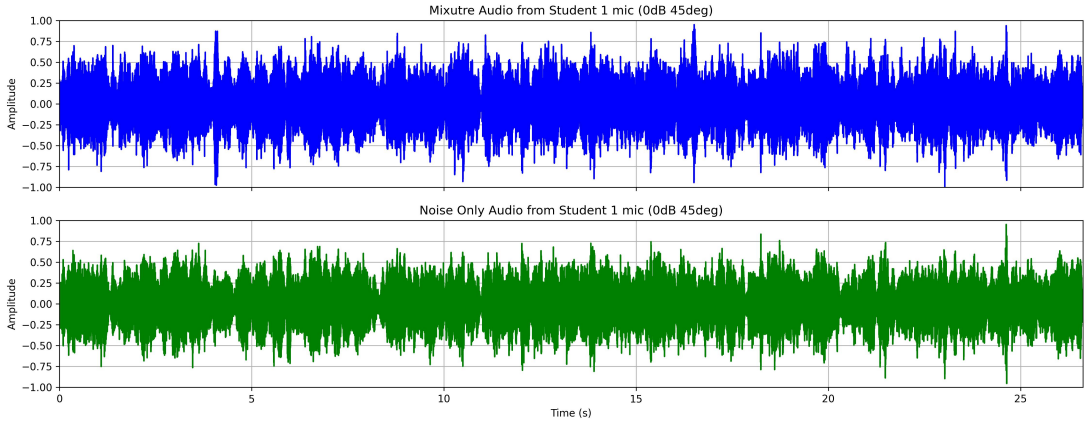
Figure 1 shows that the audio alignment algorithm identified the matching noise data in the mixture within the range 11.20 and 39.63 seconds, having an 86.2% match, which also

accounts for any local misalignments and differences in start and end time. The global phase shift (misalignment) of the matching signal was 1.818 seconds, meaning the mixture audio was brought back by that much, and the rest trimmed.

In this case, the microphones captured the data quite well, and no major local misalignments were noticeable by the eye. However, as Figure A1 shows, the code picks up a small occurrence at 24 seconds (after global alignment and trimming), and thus shifts the rest of the audio by the size needed (1 sample in this case - very minor).



(a) Audio signals before alignment.



(b) Audio signals after alignment.

Figure 5 Waveform comparison showing the mixture (blue) and noise (green) signals from Student 1's microphone before and after alignment. The alignment process synchronises both audio signals. (a) Audio signals before alignment, with untrimmed silent regions at the start and end of each recording. (b) Audio signals after alignment, only containing relevant data.

5.2 Ground Truth Selection (Separated Signal vs Isolated Signal)

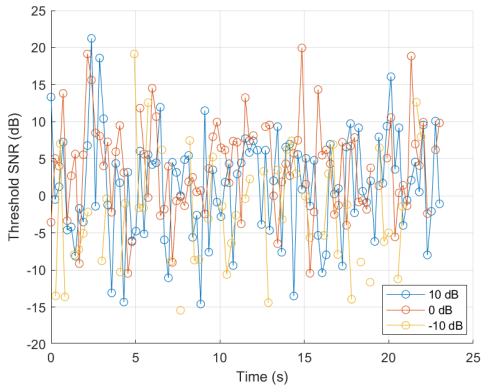
This section presents the results assessing the performance of the signal separation ground truth and Isolated signal ground truth. The former was the initial implementation, while the latter incorporated modifications to improve performance.

The signal separated ground truth produced inconsistent SNR values, with large fluctuations and deviations from the expected trend. The time domain has no consistency or pattern across the three SNR environments as shown in Figure 6a. The results show frequent over- and under-estimation of the expected SNR for each environment, suggesting the code was not accurately separating the two signals. To enforce this, in some instances, the code

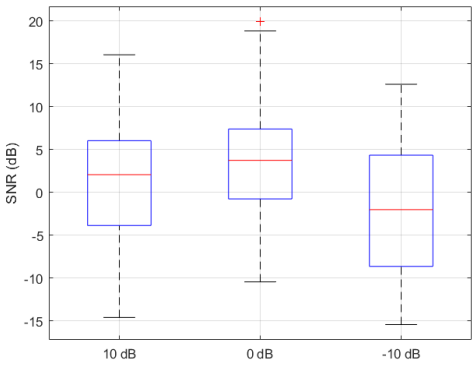
produced nans, indicating the noise was louder than the mixture, which is impossible; the only cause could be improper signal separation.

The isolated signal ground truth demonstrated a significant improvement in precision and accuracy. The estimated SNR values followed a more consistent pattern, even across all 3 SNR environments, while deviations still occur they are closer to the expected trend. Figure 6c highlights the constant SNR pattern for all tests conditions.

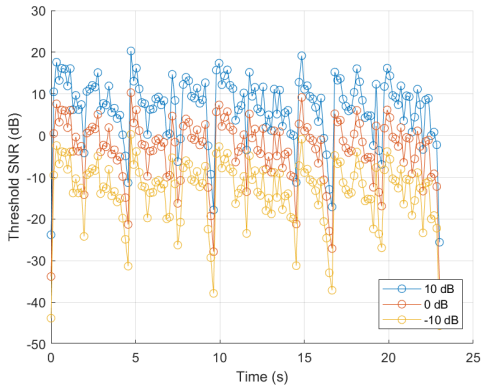
The isolated signal ground truth was selected to test the device's performance, as it accurately represented the correct nature of the ideal output.



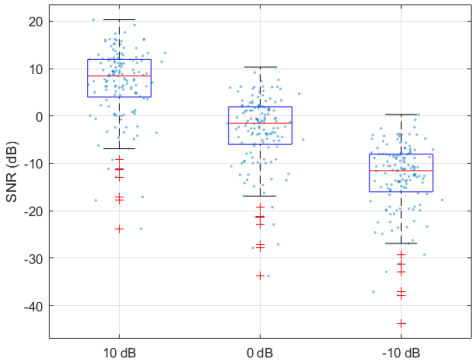
(a) Signal Separated Ground Truth: SNR vs Time.



(b) Signal Seperated Ground Truth: SNR distribution.



(c) Signal Isolated Ground Truth: SNR vs Time.



(d) Signal Isolated Ground Truth: SNR distribution.

Figure 6 Comparison of the two ground truths, the top row representing the signal separation method, while the bottom, signal isolated. (a, c) Compare the two methods SNR output over time as a waveform. Blue, red and yellow for $10\,\mathrm{dB}$, $0\,\mathrm{dB}$, and $-10\,\mathrm{dB}$ SNR environments respectively. The isolated signal ground truth shows a consistent pattern across SNR conditions. (b, d) show the isolated signal ground truth Method's improved consistency and reduced spread in SNR estimation. Accurately estimating a medium close to the respective SNR environment.

Figure 6 highlights the differences in response for the two proposed methods for obtaining the ground truth. Figures Figure 6b and Figure 6d show the box plots of the separated signal and the isolated signal ground truth, respectively. The three boxes represent $10\,\mathrm{dB}$, $0\,\mathrm{dB}$, and $-10\,\mathrm{dB}$ SNR environments, where the ground truth should match accordingly. Comparing the two plots can show the differences in the methods. The responses in Figure 6b all have large spread, and even distribution, but there is no pattern or consistency to their median SNRs; they do not match nor follow the environment SNR trend. Figure 6b

confirms that the separated signal ground truth does not accurately capture the true trend of SNR. In contrast, observing the box plots in figure 6d, it can be seen that with even spread, no skew, the SNR response and the respective median values all match the correlating environment SNR. Figure 6d shows that the isolated signal ground truth captures the true SNR response for each SNR condition.

5.3 Analysis of Variance

Table 1 Analysis of Variance (ANOVA) results for the SNR estimation experiment.

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob > F
Method	280.632	1	280.632	19.98	0.0042
SNR	321.046	2	160.523	11.43	0.0090
Angle	55.894	3	18.631	1.33	0.35
Method:SNR	14.275	2	7.138	0.51	0.62
Method:Angle	29.344	3	9.781	0.70	0.57
SNR:Angle	11.68	6	1.947	0.14	0.99
Error	84.272	6	14.045	—	—
Total	797.143	23	—	—	—

5.3.1 Angle Verification

As seen in Table 1, the angle factor produced a p-value of 0.35, greater than 0.05, indicating no statistically significant difference in SNR response across noise angles. All subsequent analyses, therefore, use data from the 180° condition.

5.3.2 Environment SNR Effect

The ANOVA results show a p-value of 0.0090 (< 0.05), indicating a statistically significant effect of environmental SNR on the measured data.

5.3.3 Method Comparison

A p-value of 0.0042 (< 0.05) indicates a statistically significant difference between the Baseline and NN SNR calculation methods on the mean measured SNR.

5.4 Cross Correlation of Signal-to-Noise Ratio Results

Following verification of the ground truth, the Baseline and NN methods were tested and compared against it. A cross-correlation of the time-domain SNR results was used to measure how closely each method's output matched the ground truth. This approach effectively evaluates the accuracy of the SNR values and the similarity of their waveform shapes over time.

Stationarity is an issue when cross-correlating two non-stationary signals; cross-correlation assumes stationary signals, and thus countermeasures are needed so it can be used on the SNR output signal. By applying the cross correlation process to each set width chunk of the SNR signal a sample response can be observed, where taking each chunk's largest correlation value can represent that segment's likeness to the ground truth. The average across all the chunks in the wave is taken. This results in the similarity score between the tested signal and the ground truth.

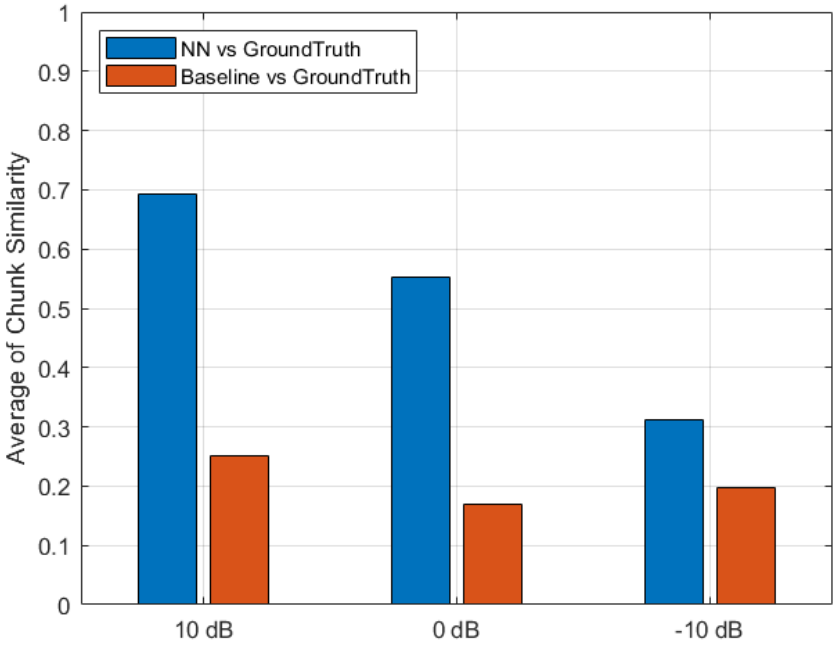


Figure 7 Comparison of the Neural Network (NN, blue) and the Baseline (red) methods against the ground truth across three SNR environments ($10\,\mathrm{dB}$, $0\,\mathrm{dB}$, and $-10\,\mathrm{dB}$). The vertical axis represents the average chunk similarity between each method's SNR and ground truth, the closer to 1, the closer the method matches the true values. A trend for the NN response can be observed. As SNR drops, the similarity between the outputs and ground truth also decreases at what seems a linear rate. However, in contrast, no trend can be seen for the Baseline behavior, for all SNR conditions, the similarity remains low and below NN for all cases.

The results in Figure 7 show that the NN method consistently achieved higher similarity scores with the ground truth than the Baseline across all SNR environments. The NN achieved values of 0.69, 0.55, and 0.31 for the $10\,\mathrm{dB}$, $0\,\mathrm{dB}$, and $-10\,\mathrm{dB}$ environments, respectively, compared with 0.25, 0.17, and 0.20 for the Baseline. Figure 7 indicates that the NN SNR response represents and matches the true SNR behavior of the environment more accuratly than the Baseline.

The superior correlation between NN estimates and the ground truth highlights its strength in maintaining temporal consistency — a limitation of the Baseline's threshold-based approach, which can lose accuracy during signal separation.

Compared to the previous work by Twinn [4] and Westbrooke [3], which relied on the Baseline algorithm, these findings suggest that a NN approach could enhance the accuracy of future classroom studies, providing stronger evidence for how noise impacts learning.

The following discussion explores these findings in greater depth, focusing on why the two methods performed differently across the SNR conditions.

6. Discussion

The NN method consistently outperformed the Baseline in all SNR environments, achieving a higher similarity to the ground truth, especially in higher SNR conditions.

Although both the Baseline and NN methods rely on thresholding (VAD) to detect if the

teacher is speaking, the NN consistently performed better due to differences in how each method handles temporal information. As mentioned, the Baseline method uses a queue-based system, where SNR estimates only update when the teacher and noise queues are both filled. However, if one queue never fills (such as when the teacher is always active), the Baseline method will not produce any SNR. It may delete old data in the queue (to make room for new data), therefore losing temporal accuracy. Splitting real-time audio into two simultaneous queues also hinders the Baseline's ability to represent data on a time scale.

These findings expand on the results stated by Twinn [4] and Westbrooke [3], who used the Baseline method to collect classroom SNR data. Their studies provided real-world insight; however, the Baseline algorithm's limitations can be justified through the variability of their results. The NN's ability to perform well highlights a potential example of a new method for enhancing acoustic measurement in New Zealand classrooms.

In contrast, the NN processes each frame continuously, keeping the SNR calculation within each time segment and allowing it to track instantaneous values. A 5-second snippet increases its ability to track the teacher's speech amidst background noise.

The ANOVA results in Table 1 highlight factors affecting the device's SNR performance. The angle factor produced a p-value greater than 0.05, indicating no statistically significant effect of noise angle on the SNR response. No significance for angle change means the device is robust against changes in noise direction, maintaining a consistent response. In contrast, the environmental SNR factor had a significant effect ($p = 0.009$), demonstrating that the device can capture changes in background noise levels as expected. Finally, the comparison between methods of obtaining SNR revealed a significant difference between the Baseline and NN approaches ($p = 0.0042$), confirming that the NN method produces more accurate SNR estimates. These results agree with Kinoshita [9], who demonstrated that NN-based spectrum estimation improves accuracy across varying noise conditions.

While the NN performs well under all conditions, a clear trend can be observed in Figure 7, where its similarity with the ground truth decreases as the environmental SNR decreases. At positive SNRs, the NN similarity tends closer to 1; however, this steadily declines with each lower SNR condition. This near-linear decrease can be explained by the NN's reliance on VAD, which uses a loudness threshold at the teacher's microphone to distinguish between noise and mixture segments in the audio. In adverse SNR environments, the higher noise levels may leak into the teacher's microphone, mistakenly triggering the VAD. As a result, noise-only segments may be treated as mixtures, offsetting the SNR estimates and reducing their similarity with the ground truth.

In practical classroom testing conditions, this sound leakage behaviour is a significant factor for background noise. The noise from neighbouring teaching zones may often reach the teacher's microphone when assessing the SNR, skewing the data. The proposed method must therefore be robust to this effect.

The pre-processing alignment matched the noise segment within the mixture with high accuracy (86.2%), and the phase shift applied (1.818 s) effectively synchronised the audio while maintaining waveform integrity. Adjusting for global and local misalignments indicates that the method is robust for this dataset. Accurate alignment is critical for SNR and cross-correlation analysis, as misalignments can shift the data and bias the results. Using both global and local alignment improves the validity of the cross-correlation similarity scores and builds confidence that the observed NN response reflects the actual SNR values.

The demonstrated NN accuracy directly supports the research objective of quantifying SNR in a classroom setting. While the current data is only relevant to offline conditions, its accuracy and stability provide a foundation for future real-time testing. Accurate SNR estimation is essential for understanding and assessing New Zealand FLEs and whether they meet established acoustic standards.

These findings have important implications for real-world classroom environments such as FLEs. Although they currently represent the system's offline performance, the NN method demonstrated superior temporal tracking and higher similarity to the ground truth, indicating higher accuracy and suggesting strong potential for use in real-time classrooms. The demonstrated accuracy highlights the feasibility and potential success of implementing a real-time version of the NN method.

7. Conclusion

This project aimed to evaluate and compare two approaches for estimating the SNR in classroom environments. A Baseline and NN method was proposed based on previous projects. Through controlled testing under varying SNR conditions, the results demonstrated that the NN consistently produced SNR estimates that closely matched the verified ground truth values, outperforming the Baseline approach in both accuracy and consistency.

The discussion highlights the NN's superior performance across all tested conditions; even offline, the data highlights the NN's potential for real-time implementation. Adding to these findings, the project demonstrates that continuous estimation of classroom SNR is obtainable with high accuracy using machine learning methods.

Beyond validating the NN's capability, the work in this project also contributes to the ongoing research improving classroom acoustics and methods of studying them, especially in FLEs. By enabling non-invasive and accurate means of estimating SNR, the proposed methods could support improvements to learning environments by informing the MoE with the results found.

Therefore, it is highly recommended that future implementations of this system adopt the NN method as the primary means of estimating SNR in classrooms.

7.1 Limitations

As discussed, the Raspberry Pi device is effectively single-channel, having only one usable stream of data from the student microphone. There is extensive evidence that single-channel systems are highly susceptible to background noise, which can distort or mask the desired signal. This limitation could explain the variability observed in some of the real-time results seen in this project.

Without the use of multiple channels, advanced signal-processing methods such as beam-forming or Time Difference of Arrival (TDOA) cannot be implemented. These methods are commonly used to reduce noise and enhance the clarity of the target signal by focusing on the direction of arrival of sound. The absence of these techniques restricts the device's ability to fully separate the two signals (teacher and noise).

Additionally, all testing was carried out in a controlled laboratory using simulated SNR environments. While this allowed for reliable and consistent results, the lack of testing in real classrooms means the results may not fully represent actual teaching conditions. The

testing also used a limited dataset with fixed SNR levels of 10 dB, 0 dB, and -10 dB, whereas real-world data would include a continuous and dynamic range of SNR values.

7.2 Future Work

The project's original scope was to test and implement real-time methods of estimating SNR. Now, with the knowledge and verification of the NN's accuracy and potential, even offline, future versions should implement real-time microphone inputs instead of an audio file, protecting and maintaining the privacy considerations while providing valuable insight into the SNR in FLEs over the course of a teaching day.

Another promising direction for future development of this project is implementing a delay-and-sum beamforming technique using a Time Difference of Arrival (TDOA). TDOA refers to the time difference at which a sound wave reaches each microphone, allowing the system to triangulate the specific direction of the sound source. Beamforming enables a microphone array to focus on sounds arriving from a particular direction while attenuating noise from all others.

Integrating TDOA and beamforming involves applying each microphone's time delay so that all the audio signals align in time before being summed. As a result, the target signal constructively interferes, while all other signals from other directions destructively interfere, reducing unwanted noise and elevating the desired sound.

In the context of this project, such an approach could be used to isolate the teacher's voice more effectively before calculating the SNR. This would enhance the device's ability to measure SNR accurately in real-time FLEs where background noise sources come from multiple directions.

7.3 Limitations

As discussed, the Raspberry Pi device is effectively single-channel, having only one usable stream of data from the student microphone. There is extensive evidence that single-channel systems are highly susceptible to background noise, which can distort or mask the desired signal. This limitation could explain the variability observed in some of the real-time results seen in this project.

Without the use of multiple channels, advanced signal-processing methods such as beamforming or Time Difference of Arrival (TDOA) cannot be implemented. These methods are commonly used to reduce noise and enhance the clarity of the target signal by focusing on the direction of arrival of sound. The lack of such techniques restricts the system's ability to fully separate the teacher's voice from background noise in complex classroom environments.

Additionally, all testing was carried out in a controlled laboratory using simulated SNR environments. While this allowed for reliable and consistent results, the lack of testing in real classrooms means the results may not fully represent actual teaching conditions. The testing also used a limited dataset with fixed SNR levels of $10\,\mathrm{dB}$, $0\,\mathrm{dB}$, and $-10\,\mathrm{dB}$, whereas real-world data would include a continuous and dynamic range of SNR values.

7.4 Future Work

The project's original scope was to test and implement real-time methods of estimating SNR. Now, with the knowledge and verification of the NN's accuracy and potential, even offline, future versions should implement real-time microphone inputs instead of an audio

file, protecting and maintaining the privacy considerations while providing valuable insight into the SNR in FLEs over the course of a teaching day.

Another promising direction for future development of this project is implementing a delay-and-sum beamforming technique integrated with Time Difference of Arrival (TDOA). TDOA refers to the time difference at which a sound wave reaches each microphone, allowing the system to triangulate the specific direction of the sound source. Beamforming enables a microphone array to focus on sounds arriving from a particular direction while attenuating noise from all others. Integrating TDOA and beamforming involves applying each microphone's time delay so that all the audio signals align in time before being summed. As a result, the target signal constructively interferes, reinforcing the desired sound, while all other signals from other directions destructively interfere, reducing unwanted noise.

In the context of this project, such an approach could be used to isolate the teacher's voice more effectively before calculating the SNR. This would enhance the device's ability to measure SNR accurately in real-time FLEs where background noise sources come from multiple directions.

References

[1] O. Wilson, *Classroom acoustics : a New Zealand perspective*. Oticon Foundation in New Zealand, 2002, 2002.

[2] M. Donn, G. Dodd, and S. Leggett, "The acoustic performance of modern learning environments vs. single cell classrooms," in *Proceedings of the 21st International Congress on Sound and Vibration (ICSV21)*, 2014. [Online]. Available: https://www.researchgate.net/publication/307637517

[3] K. M. Westbrooke, "Signal-to-noise ratios in aotearoa flexible learning environments," University of Auckland, Acoustics Research Group, Tech. Rep., 2024.

[4] B. M. T. Twinn, "Speech audibility in flexible learning environments of aotearoa: Factors that determine signal-to-noise ratio," Tech. Rep., 2024.

[5] American National Standards Institute and Acoustical Society of America, *ANSI/ASA S12.60-2010/Part 1: Acoustical Performance Criteria, Design Requirements, and Guidelines for Schools*. Melville, NY: Acoustical Society of America, 2010, approved April 13, 2010 by the American National Standards Institute.

[6] J. S. Bradley, "The intelligibility of speech in elementary school classrooms," *Canadian Acoustics*, vol. 36, no. 3, pp. 31–35, 2008.

[7] A. C. Neuman, M. Wroblewski, J. Hajicek, and A. Rubinstein, "Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults," *Ear and Hearing*, vol. 31, pp. 336–344, 6 2010.

[8] B. Yen, C. T. J. Hui, E. Bergin, E. Jensen, S. C. Purdy, W. Keith, Y. Hioka, J. Whitlock, G. Dodd, M. D. Acoustics, and N. Z. M. of Education, "Development of a continuous classroom signal-to-noise ratio measurement system," University of Auckland, Acoustics Research Group, Auckland, New Zealand, Tech. Rep., 2023.

[9] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online wpe dereverberation," in *Interspeech 2017*, 2017, pp. 384–388.

[10] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6875–6879.

[11] A. Meijer, M. R. Benard, A. Woonink, D. Baskent, and E. Dirks, "The auditory environment at early intervention groups for young children with hearing loss: Signal to noise ratio, background noise, and reverberation," *Ear and Hearing*, 2025.

[12] A. Weisser and J. M. Buchholz, "Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions," *The Journal of the Acoustical Society of America*, vol. 145, pp. 349–360, 1 2019.

[13] J. Whitlock, "Classroom acoustics research in new zealand: Past, present and future," *New Zealand Acoustics*, vol. 21, no. 4, pp. 25–28, 2008. [Online]. Available: https://www.oticon.org.nz

[14] Post Primary Teachers' Association (PPTA), "Flexible learning spaces: An experiment on our education system?" PPTA Te Wehengarua Annual Conference, Wellington, New Zealand, Tech. Rep., 2017.

[15] L. M. Wang and L. C. Brill, "Speech and noise levels measured in occupied k–12 classrooms," *The Journal of the Acoustical Society of America*, vol. 150, pp. 864–877, 8 2021.

[16] S. R. Bistafa and J. S. Bradley, "Reverberation time and maximum background-noise level for classrooms from a comparative study of speech intelligibility metrics," *The Journal of the Acoustical Society of America*, vol. 107, pp. 861–875, 2 2000.

[17] S. Nittrouer and A. Boothroyd, "Context effects in phoneme and word recognition by young children and older adults," *The Journal of the Acoustical Society of America*, vol. 87, no. 6, pp. 2705–2715, 1990.

[18] Standards Australia and Standards New Zealand, *Acoustics: Recommended Design Sound Levels and Reverberation Times for Building Interiors (AS/NZS 2107:2016)*. SAI Global Limited under licence from Standards Australia Limited and Standards New Zealand, 2016.

[19] B. Yen, *Classroom SNR Calculation Device: User Manual*, University of Auckland, Acoustics Research Centre, 2023.

Appendix A

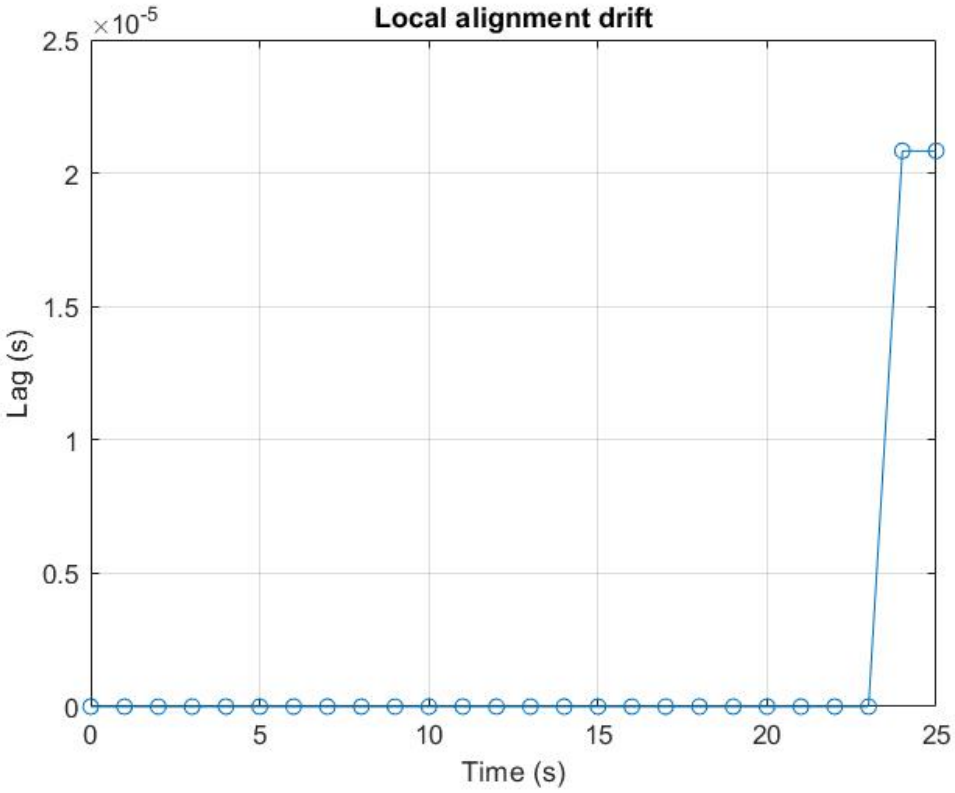


Figure A1 Example of local phase shift needed to align mixtures. It shows a jump around 23 seconds, indicating all data after must be shifted forward.