Ollscoil na hÉireann, Corcaigh

Coláiste na hOllscoile Corcaigh
University College Cork, Ireland

Implementation of an AI-Assisted Sonification Algorithm on an Edge Device

Presented by

Feargal O'Sullivan

For the degree of

MEngSc

Head of School/Discipline: Prof. Jorge Oliveira/Prof. Peter Parbrook
Supervisors: Dr. Emanuel Popovici and Prof. Andriy Temko

Department of Electrical and Electronic Engineering
School of Engineering and Architecture
University College Cork

2023

Abstract

Oxygen deprivation at birth leads to brain injury, which can have serious consequences. It is the dominant cause of seizures. Quickly and accurately detecting seizures is a challenging problem for neonates. A severe shortage of medical professionals with the necessary expertise for Electroencephalogram (EEG) analysis leads to significant delays in decision-making and hence treatment. These problems are made worse in disadvantaged communities. Artificial intelligence (AI) techniques have been proposed to automate the process and compensate for the lack of available expertise. However, these models are 'black boxes', and their lack of explainability dampens the wide adoption by medical professionals. AI-assisted sonification adds explainability to any such automated methodology, empowering medical professionals to make accurate decisions regardless of their level of expertise in EEG analysis. The feasibility of an implementation of an AI-assisted sonification algorithm on an edge device is presented and analyzed. A lightweight derived algorithm for resource-constrained implementation scenarios is also evaluated and presented, suggesting suitability for further ultra-low power, mobile and wearables implementations. Furthermore, a neural network is analysed for the potential of low-precision implementation, enabling inference on specialised hardware.

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Feargal O'Sullivan, 2023

Acknowledgements

I would like to thank my supervisors, Dr Emanuel Popovici and Dr Andriy Temko, for their guidance during this project. I would also like to thank Dr Sergi Quintana-Gomez, for his advice and technical suggestions.

I would like to thank my parents for their support and sacrifices during and long before this research, without whom it would not have been possible.

Contents

List of Tables

LIST OF TABLES

List of Figures

Glossary

AC	Alternating Current
ADSR	Attack Delay Sustain and Release
aEEG	Amplitude Integrated EEG
AI	Artificial Intelligence
APT	Average Processing Time
AUC	Area Under the Curve
BCE	Binary Cross Entropy
CI	Confidence Interval
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CV	Cross Validation
DSP	Digital Signal Processing
ECG	Electrocardiogram
EEG	Electroencephalogram
EMG	Electromyography
EOG	Electrooculogram
FCNN	Fully Convolutional Neural Network
FFT	Fast Fourier Transform
HIE	Hypoxic-ischemic Encephalopathy
IIR	Infinite Impulse Response
ML	Machine Learning
NICU	Neonatal Intensive Care Unit
NPU	Neural Processing Unit
PSD	Power Sprectral Density
ReLU	Rectified Linear Unit
ROC	Reciever Operating Characteristic
STFT	Short Time Fourier Transform
TFD	Time Frequency Distribution
WHO	World Health Organisation

Publications

1. **F. O'Sullivan**, S. G. Quintana, A. Temko and E. Popovici, "An implementation of an AI-assisted sonification algorithm for neonatal EEG seizure detection on an edge device," 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Ioannina, Greece, 2022, pp. 01-04, doi: 10.1109/BHI56158.2022.9926876. **(Won Best Paper Award, 3rd Prize)**

2. S. Gomez-Quintana, **F. O'Sullivan**, A. Factor, E. Popovici, A. Temko: Sonif.AI: empowering the medical professional with fast and accurate interpretation of neonatal EEG, poster presented at the 9th Congress of the European Academy of Paediatric Society, Barcelona, 7-10 Oct 2022. Abstract published in Frontiers in Pediatrics. ISBN: 978-2-88971-024-9 DOI: 10.3389/978-2-88971-024-9 **(Impact Factor 3.569)**

3. T. Nguyen, A. Daly, **F. O'Sullivan**, S. G. Quintana, A. Temko and E. Popovici, "A real-time and ultra-low power implementation of an AI-assisted sonification algorithm for neonatal EEG," Accepted to be published in the IEEE 9th International Workshop on Advances in Sensors and Interfaces (IWASI)

1

Introduction

1.1 Overview

Deaths in neonates (within 28 days of birth) are of major concern. The World Health Organisation estimate that in 2019 there were 2.4 million neonatal deaths, with the majority occurring within the first week of life. 1 million deaths occur within the first 24 hours (WHO, 2020). Death rates are worse in developing countries, where 29.4 per 100 live births are affected. This is in contrast to developed countries, where 3.5 neonates per 1000 live births are affected (Ye et al., 2016).

One of the leading causes of a high death rate in babies is due to birth or perinatal asphyxia (deprivation of oxygen). A common cause of brain asphyxia is hypoxic-ischemic encephalopathy (HIE), where blood flow to the brain is impaired, and oxygen delivery is obstructed. Brain asphyxia is the fifth biggest cause of mortality for children under the age of 5, resulting in 8.5% of deaths (Lawn et al., 2007). Most of these deaths are preventable with sufficient care (Oza et al., 2014). There is a higher rate of HIE incidences in poorer countries than wealthier ones, where there are as high as 26 per 1000 live births in developing countries and between 1 to 8 per 1000 live births in the developed world (Douglas-Escobar and Weiss, 2015).

Early detection of neonatal seizures is an essential yet difficult clinical task (Murray et al., 2008). Failure to detect such events within an optimal time window may reduce treatment efficiency and ultimately increase mortality and morbidity rates (Pavel et al., 2022).

Studies have shown that clinicians achieve less than 10% accuracy in diagnosing neonatal seizures based on clinical signs alone (blinking or abnormal eye movements) (Murray et al., 2008). When detecting a condition through clinical signs is impossible, physiological signals must be measured and analysed. The most common tool for monitoring and detecting abnormal brain function is the visual analysis of EEG recordings. It records brain activity through electrodes placed on the surface of the scalp and is the only reliable way to detect seizures (Murray et al., 2008).

The expensive equipment and expertise required to analyse EEG for seizures are only available in tertiary care hospitals. Even if the expertise is available, reviewing and analyzing the recording takes a long time. It is reported in (Brogger et al., 2018) that it takes from one-seventh up to half of the original recording length for a medic to review/analyze an EEG recording, where individual recordings are typically several hours long.

In underfunded health systems, the necessary equipment may not be accessible.

While in developed countries, expertise is the main limiting factor. A survey carried out in (Boylan et al., 2010) showed that while 90 % of surveyed hospitals in the US and Europe had access to EEG equipment, only half of Neonatal Intensive Care Unit (NICU) personnel had formal training in interpreting the signal. Only 9 % of them felt comfortable making a diagnosis based on data. The lack of internal expertise leads to most diagnoses being made by specialist external neurophysiologists. This delayed treatment likely caused an increased morbidity rate as administration of anti-epileptic drugs within 24 hours of a patient's first seizure is crucial for a positive outcome (Pavel et al., 2022).

The challenges associated with detecting neonatal seizures and the lack of available expertise across clinical settings have prompted research into developing automated neonatal seizure detection algorithms (Temko et al., 2011). These algorithms are designed to be decision-support tools; they should alert clinical staff when suspected seizure events are detected. Recent advances in deep learning allow for fast and accurate detection of seizures for neonates (Daly et al., 2021; O'Shea et al., 2020), with high performance (95%- 97% AUC). These techniques are particularly interesting as they allow real-time online analysis of EEG. However, there is more to consider when evaluating models than absolute performance. Explainability is a concept where an AI model's output can be interpreted and understood by humans at a reasonable level (Holzinger et al., 2019). These "black-box" techniques lack explainability, a key feature for pervasive adoption in a medical setting (Kundu, 2021).

Along with AI that provides an objective assessment of EEG, new methods of subjective EEG analysis have emerged in sonification to support and complement visual EEG assessment (Väljamäe et al., 2013). In (Gómez-Quintana et al., 2021), a scheme in which AI is used as a "watchdog" for the presence of seizure followed by a review using AM/FM sonification mechanism is presented. If the AI inference engine detects a seizure, it generates an alarm, and a medical professional can review the previous EEG with a compression ratio of 20 (the ratio of the input EEG duration to the output audio duration).

However, this sonification mechanism does not (re)use the probabilities generated by the AI inference engine. A new AI-assisted sonification algorithm was presented in (Gomez-Quintana et al., 2022), which combines the benefits of fast, objective AI analysis with the acuity of the human ear to detect seizure frequency evolution in time. In this algorithm, the probabilistic outputs of the AI inference engine are used to identify the regions of EEG likely to contain a seizure event. By variably processing and compressing EEG, greater overall compression can be achieved while still detecting brief seizures.

1.2 Aim and Scope

This work aims to implement an EEG sonification algorithm on an edge device. This will enable pervasive and ubiquitous EEG analysis. This thesis presents the first implementation of the AI-assisted sonification algorithm in (Gómez-Quintana et al., 2021) on a resource-constrained, low-cost edge device. The implementation helps to identify the most computationally intensive blocks. An example of a computationally intensive operation is to remove ECG interference, time delay is estimated every 8 seconds by means of a convolution. This function along with several others was shown to take up the majority of the execution time. Based on these observations, a new lightweight algorithm is derived, drastically reducing the computation time. The AI element of the algorithm is also optimised for a resource-constrained device.

1.3 Research Contribution

1. Deployment of an existing sonification algorithm to an edge device to allow for cheap deployment in resource-constrained settings. Optimisations are made to the algorithm by lowering the sampling rate from 256 Hz to 32 Hz and making some computations parallel. Inspecting the spectral output of both algorithms shows that performance is unaffected. It's shown that further improvement is required to increase the usability of the algorithm in a clinical environment.

2. A lightweight algorithm derived from the previously deployed algorithm has been proposed to further decrease processing time in a resource-constrained setting. Based on the timing results from the deployment of the original algorithm, alternative, less computationally intensive functions were proposed. In some cases where the function's value was limited and no suitable alternative was found, the function was removed. A survey was conducted to compare the performance of both surveys. Performance between the two algorithms was deemed to be the same.

3. Optimisation of the CNN used in the sonification algorithm is also explored. The network is quantised to a variety of precisions (8 bit, 16 bit), and loss in performance is measured. Quantise aware training is carried out in an effort to reduce quantisation error. It's found that the network can be

quantised to 8 weights and 16 bit activation functions without any loss in performance.

1.4 Thesis Outline

The remaining chapters of this dissertation are as follows:

Chapter Two presents state-of-the-art and common practices when assessing newborn brain health. Automated seizure detection methods are explored. An overview of seizures and their causes is given.

Chapter Three implements a previously presented sonification algorithm on an edge device. Optimisations are made to the initial implementation. The non-optimised and optimised versions of the algorithm are then compared in terms of output and execution time.

Chapter Four focuses on further decreasing the execution time. A lightweight version of the algorithm is proposed and tested.

Chapter Five optimises the neural network for deployment on an edge device using low-precision operations.

Finally, *Chapter Six* gives conclusions and avenues for future work.

2

State of the Art

The period immediately after birth is of particular interest to medical professionals. Abnormal developments in vital organs like the brain, heart or lungs have the potential to cause permanent health implications or death. In the postnatal period, neonates are not fully developed, and often, clinical signs do not manifest. Therefore monitoring physiological signs is paramount to successful diagnosis and treatment.

Seizures are common symptoms of neurological abnormality in neonates. Seizures affect a person's development and have a lasting effect on the quality of their life over a lifetime. The health implications are severe, and such seizures must be cared for as clinical emergencies. Seizures occur when underlying neurological disorders cause abnormal neuron firing patterns. The current gold standard in clinical practice is the visual analysis of EEG. However, its review is time-consuming and requires a high level of expertise that's not widely available.

The below sections present a broad overview of this traditional analysis, a review of current AI methods to automate the process and, finally, an exploration of sonification methods. All of these methods offer pros and cons in terms of explainability, performance and human involvement. Recent developments combining AI and sonification are also explored.

2.1 The Brain

The human brain is a massive control centre, making decisions and autonomously running vital systems to keep us alive. From the power of 100 billion neurons (Herculano-Houzel, 2009), our nervous system gathers information and pilots the rest of our body.

Neurons are connected to each other via nerve tissue to transmit and receive information (Sanei and Chambers, 2013) and glial cells, which support and maintain the neurons' environment. Most neurons are comprised of a cell body, an axon for information transmission and dendrites for information reception. A neuron transmits information through electrical potential. A typical range for the potential is -70mV in resting conditions and +40mV when neurons are in depolarisation mode (Barnett and Larkman, 2007). A neuron can connect to 10000 neighbouring neurons even in simple connection patterns (Kandel et al., 2000). These large networks are what enable humans to do complex tasks.

Early life is critical for brain development. The brain develops continuously until adulthood, but the majority of development in terms of volume and cognitive functions occurs in the first few years of life (Johnson, 2001; Stiles and Jernigan, 2010). Newborns may suffer perinatal asphyxia during birth leading to permanent damage (Fenichel, 1983). HIE can have a lasting effect on many facets of brain function (cognition, motor, behavioural, visual, and hearing) (Miller and Ferriero, 2009).

2.2 EEG Acquisition

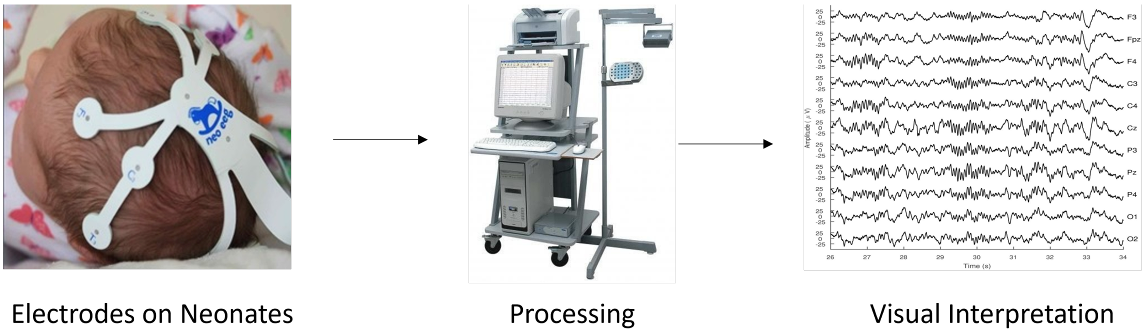


Figure 2.1: EEG acquisition flow from electrode to interpretation

Figure 2.1 shows a high-level diagram detailing the process from EEG acquisition to interpretation. Electrodes are metal plates that measure electrical activity in the brain through direct contact with the scalp. Electrodes are usually located on patients' heads using the 10-20 system. In this methodology, the numbers 10 and 20 refer to the distances between them. They can either be 10 % or 20 % of the total distance of the skull apart (measured from front-back or right-left). In this way, the electrodes are always positioned proportionately, and results can be inferred generally across individuals. An ordered arrangement of EEG electrodes is called a montage. Due to the newborn's smaller head sizes, a smaller modified montage is used (Jasper, 1958; Shellhaas et al., 2011).

The voltage at each electrode in a referential montage is measured with respect to the referential electrode (usually placed on the ears). Often, in clinical practice, a bipolar montage is used. In a bipolar montage, an electrode's voltage is measured with respect to its surrounding electrodes. This gives a better representation of localised brain activity. Figure 2.2 illustrates both the referential and bipolar montages. The 8 channel montage relies on 9 electrodes, while the 18 channel relies on all 19 electrodes.

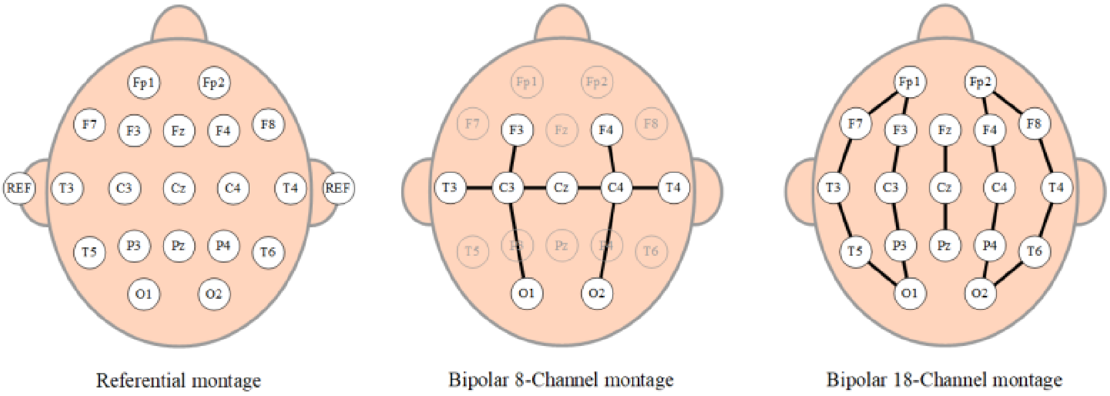


Figure 2.2: Diagram of bipolar and referential EEG montages (Gomez-Quintana et al., 2022)

A firing neuron has electrical potential in the millivolt range ((Barnett and Larkman, 2007)), but due to a dampening effect as the electrical signal propagates through the scull and surrounding soft tissue by the time the electrical signal is measured by an electrode, it is in the order of hundreds of microvolts (Aurlien et al., 2004). To increase signal strength, an abrasive gel removes the outermost layers of skin and increases the conductivity between the electrode and the skin (Lloyd et al., 2015). Dry electrodes have recently been investigated as a potential replacement (Di Flumeri et al., 2019; O'Sullivan et al., 2019).

When analysing EEG frequency, the content is grouped into 5 categories: delta, theta, alpha, beta and gamma. The frequency ranges are shown in Table 2.1. As previously stated, seizure activity occurs below 13 Hz (Kitayama et al., 2003), so the delta, theta and alpha ranges dominate. Activity in the ranges is useful for assessing a patient's health. For example, delta and theta bands dominate when awake, but alpha and beta activity may occur during sleep (Tsuchida et al., 2013). Generally, spectral density decrease with frequency. Figure 2.3 shows half an hour of healthy 8 channel EEG and the corresponding PSD per channel.

Table 2.1: EEG frequency bands

Frequency Band	Frequency Range (Hz)
Delta	0.5-4
Theta	4-8
Alpha	8-14
Beta	14-30
Gamma	30-100

EEG is measured in the microvolt (μV) range. The electrical signal is weak and is susceptible to many artefacts. Electrode disconnection, AC supply, and other biosignals (ECG, EOG, EMG) interfere with the EEG. ECG is a repetitive electrical signal and is in the range of millivolts (mV) (White and Van Cott, 2010). Hence, it is one of the most persistent causes of interference. ECG artefacts can resemble seizures. However, it is possible to distinguish seizures as their frequency evolves over time.

Visual interpretation of EEG in the time domain is a complex process (Husain, 2005). A patient's EEG may be monitored over several days. Seizures

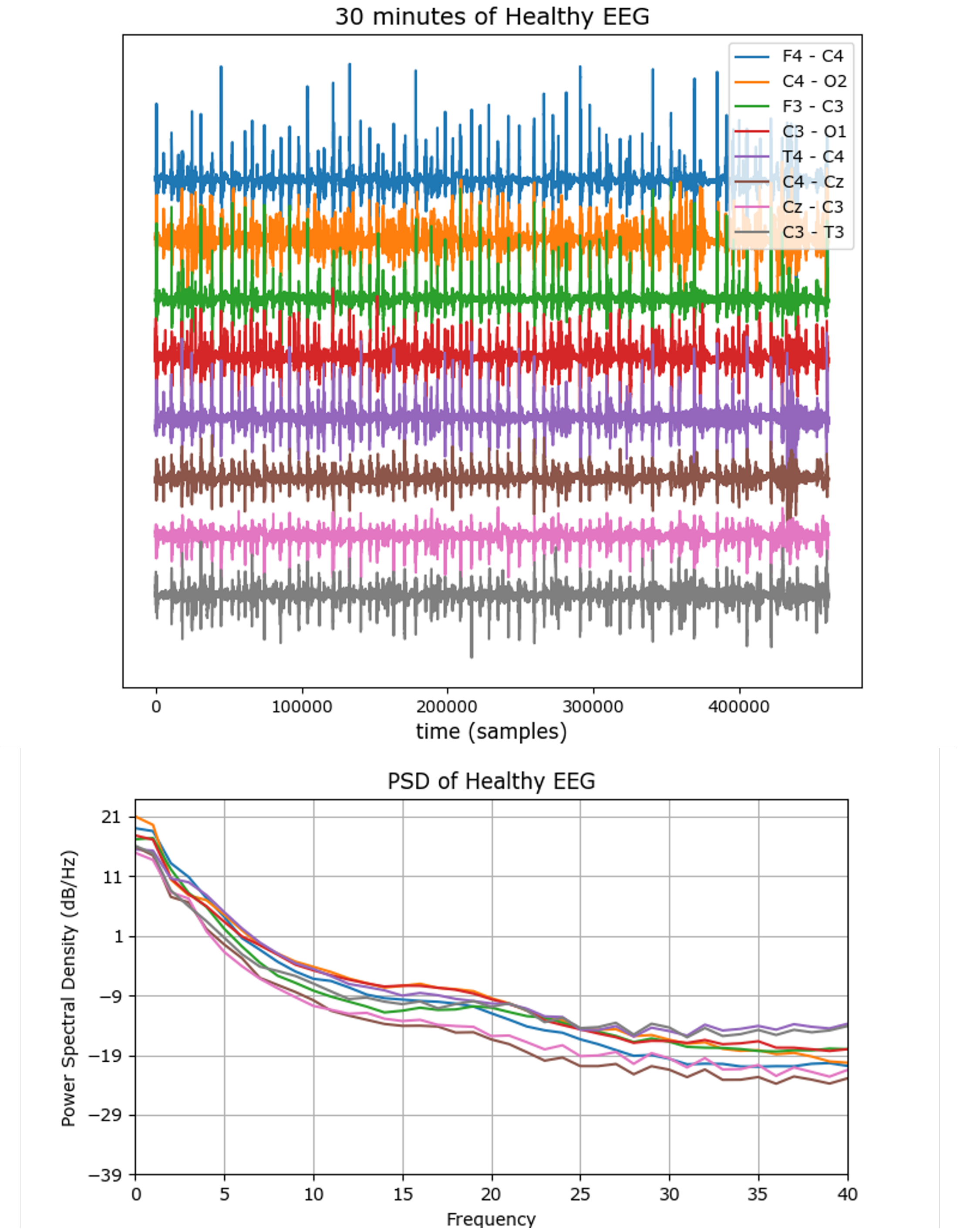


Figure 2.3: PSD of healthy EEG

are relatively rare events and are time-consuming to locate. It's been shown that there is substantial disagreement even amongst experts when diagnosing seizures from EEG (Stevenson et al., 2019). Amplitude-integrated EEG (aEEG) is often used to decrease the burden on a clinical practitioner. This method temporally smooths and compresses the signal's energy, making it easier to interpret (Rakshasbhuvankar et al., 2015). Although aEEG makes it easier to review several hours of EEG at once quickly, it also increases the number of false alarms due to energy artefacts and leads to misdiagnoses of short seizures (Rennie et al., 2004; Zhang et al., 2011).

2.3 Seizures

Neonatal seizures are common symptoms of HIE, strokes and infections (Delanty et al., 1998; Ramantani et al., 2019). The typical mortality rate for neonates suffering from seizures is 10 % (range 7-16 %). Of those who survive, 50 % suffer permanent disability (McBride et al., 2000; Nagarajan et al., 2010; Scher et al., 1989; Uria-Avellanal et al., 2013). The rate of reported seizure instances in the NICU varies from 1.5 to 3.5 out of 1000 patients (Eriksson and ZetterstrÖM, 1979; Lanska et al., 1995; Ronen et al., 1999). Neonatal seizure detection is particularly challenging. Adults present clinical signs including involuntary limb movement, breathing cessation and blinking. Only 10 % of neonatal seizures are detectable using clinical signs alone. Brain monitoring with EEG is the only way to accurately detect seizures (Murray et al., 2008).

Seizures occur when a neuron's activation becomes uncontrolled, and the surrounding network begins to discharge synchronously (Scharfman, 2007). The related frequency content is between 0.5 and 13 Hz (Kitayama et al., 2003). In contrast, a healthy patient's neurons communicate independently, and the discharge frequency can be modelled as random noise (Rankine et al., 2006). The distinct electrical patterns in the brain during a seizure make them detectable through EEG. A comparison of single-channel seizure and non-seizure EEG is made in Figure 2.4.

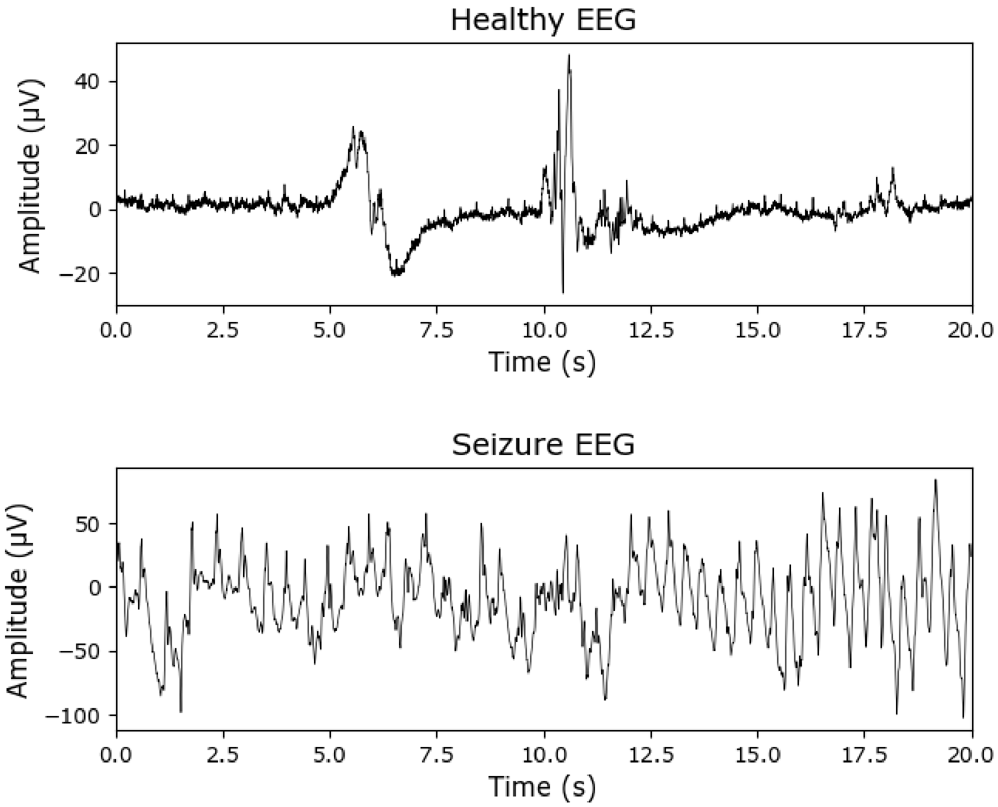


Figure 2.4: Comparsion of healthy EEG and EEG containing a seizure in a single channel.

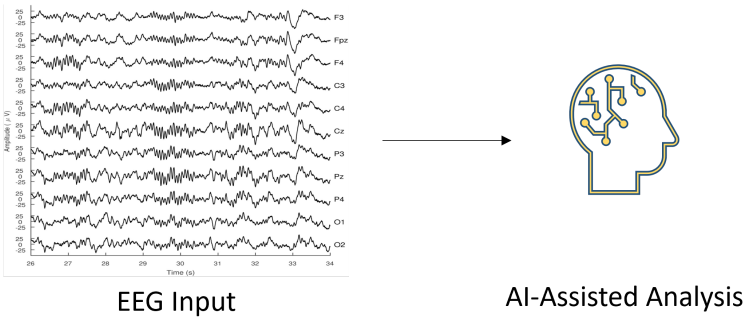


Figure 2.5: AI assisted decision making

2.4 AI for Automated Seizure Detection

AI is becoming popular to assist medical professionals in decision-making across all domains (Figure 2.5) (Benjamens et al., 2020). AI has the potential to fill the gaps in the availability of highly specialised expertise, decreasing the time taken for a diagnosis and hence the time to treatment. Neurophysiologists detect seizures by looking for repetitive abnormal frequency and amplitude patterns in the signal (Scharfman, 2007). Figure 2.4 shows a repetitive pattern corresponding with a seizure. Seizures can be distinguished from background healthy EEG due to their evolving frequency and amplitude (Patrizi et al., 2003). Literature defines seizures as events greater than ten seconds (Clancy and Legido, 1987).

Using domain knowledge, early rule-based seizure detection algorithms were developed, exploiting frequency, correlation, entropy and pattern repetition to make a classification (Faul et al., 2005; Gotman et al., 1997; Liu et al., 1992; Roessgen et al., 1998). Over time computing power increased, and it became possible to train statistical models using gradient descent to iteratively adjust parameters to minimise error. Models trained in this fashion are referred to as machine learning models. Using machine learning, state-of-the-art performance in automated seizure detection was greatly improved (Aarabi et al., 2006; Ansari et al., 2016; Tapani et al., 2019; Temko et al., 2011; Thomas et al., 2010). All of these algorithms utilise traditional machine learning techniques consisting of a feature extraction stage to select features that best summarise information related to seizures and model trained to make a classification based on the features.

Further increases in computing power have allowed for the training of deep learning models. Deep learning approaches do not require manual handcrafted features. Because features are automatically extracted, it is unclear what causes a decision, and these models lack explainability. Convolutional Neural Networks (CNN) have been used to achieve state-of-the-art performance on the neonatal seizure detection problem (Daly et al., 2021; O'Shea et al., 2020). They originated in the 1980s (Fukushima and Miyake, 1982), but their use has become widespread with the increase in deep learning methods. In CNNs, kernels are convoluted with the input to enhance seizure patterns while maintaining the spatial relationship between input components. After several convolution layers, a decision is made either by a convolutional layer or, in a fully convolutional neural network, by global pooling.

All AI models have some form of error. (Stevenson et al., 2019) showed that even expert doctors disagree when diagnosing seizures. If the ground truth labels are subjective in nature, there is an upper limit to the AI's performance. In a clinical environment, the cost of a wrong diagnosis leads to unnecessary treatment or even death. Using models with a high level of explainability allow the traceback of errors to the underlying clinical causes (Linardatos et al., 2020). For example, in (Ahmed et al., 2016), the features contributing to wrong decisions are analysed so the people using the models can interpret the decisions better. It is preferable for clinicians to understand the reason why decisions are made. However, this level of explainability is not achievable with deep learning approaches. To include explainability in the decision process, traditional machine learning models can be used at the cost of lower performance.

2.5 Sonification

Sonification is the use of non-speech audio to perceptualise data. The fully automated AI seizure detection methods discussed above relieve clinicians of the tedious task of manually assessing EEG. Although to get the best performance, models which are not explainable must be used. Integrating these models with sonification can combine high-performance levels with explainability as the human is kept in the loop to interpret the output. Because the cost of a wrong diagnosis is so high, a human must always be kept in the loop. In the best-case scenario, a neurophysiologist would review the EEG. However, due to the lack of available expertise, this is difficult to put into practice. Sonification offers a solution as the EEG can be brought into the audible domain and intuitively interpreted with minimal training. The human ear has evolved to detect sound patterns accurately and can easily differentiate a seizure's evolving frequency from artefacts and background EEG. Sonifying EEG often temporally compresses the signal yielding short review times. EEG sonification algorithms are analogous to a stethoscope, where a doctor can hear abnormalities in the heart.

The seizure patterns have a characteristic evolving temporal and spatial evolution (Clancy and Legido, 1987; Rose and Lombroso, 1970). Since most of the neonatal EEG's frequency content lies in the range 0.5-13 Hz (Kitayama et al., 2003), EEG is inaudible. A sonification algorithm is used to shift the frequency of EEG into the range of 20 Hz to 20 kHz (Huttunen et al., 2007) to make these patterns detectable in audio. The phase vocoder was first used in EEG sonification with a constant compression rate (Temko et al., 2014a). Later, it was observed that increasing the rate of temporal compression improves the algorithm's performance as seizures and non-seizures are more separable at higher frequencies (Gomez et al., 2018). Although increasing the audio speed increases the algorithm's general performance, it can lead to missed short seizures. The effects of this problem were mitigated in (Gomez-Quintana et al., 2022), where AI was used to locate seizures and variably compress the EEG accordingly. Using this methodology, greater temporal compression was achieved without affecting the detection of short seizures. This methodology combined the excellent performance of a CNN with the explainability of sonification algorithms.

Other sonification methods are presented in the literature. The most basic of which increases the input signal's sampling rate (Khamis et al., 2012; Olivan et al., 2004). More elaborate methods based on tone synthesis (Baier et al., 2007; Hermann et al., 2002), a mapping from EEG to musical notes

(Loui et al., 2014), and a voice-like synthesiser (Parvizi et al., 2018). These approaches are limited compared to (Gomez-Quintana et al., 2022) because they are only guided by the EEG and do not convey all the information in the signal.

3

Implementation on an Edge Device

A major obstruction to pervasive 24 hour EEG monitoring is the need for more interpretation expertise. aEEG has been used as an alternative to reduce the necessary level of expertise required; however, its use has led to missing short and low-amplitude seizures (Hellström-Westas et al., 2006).

The previously proposed sonification algorithm promises to mitigate these errors while requiring practitioners to undergo virtually no training (Gomez-Quintana et al., 2022). For this technology to become widely adopted, it must be low cost and available to clinicians at the cotside. Edge devices are computers located physically close to the data source with enough computing resources to process data. The following chapter presents an implementation of the algorithm on an edge device. This device can be integrated with existing EEG acquisition equipment cheaply.

A Raspberry Pi 3B+ was chosen as an edge device due to its integrated audio output and connectivity options, including Ethernet, Wi-Fi and Bluetooth, high processing capabilities, and versatility in prototyping the system. The integrated connectivity options allow the Pi to seamlessly interact with existing EEG acquisition systems for seamless integration into various hospital setups. The Raspberry Pi contains a Broadcom BCM2837B0 CPU (Figure 3.1).

Some optimisations are made to reduce clinicians' wait time. The Average Processing Time (APT) was measured for the entire algorithm and each function to evaluate their contribution to the overall execution time.

Figure 3.1: Picture of Raspberry Pi 3B+ chosen for implementation

3.1 Python vs MATLAB

This chapter focuses on migrating and optimising a MATLAB implementation of an algorithm to Python. MATLAB and Python are similar in some ways as they are high-level, interpreted languages.

In contrast to MATLAB, Python is a free, open-source language with many available libraries and packages. Python's libraries become particularly advantageous when using AI models. Python is also more portable than MATLAB, meaning without any changes the same code can be run in different environments independent of of operating system or even device. Its portability makes it easier to implement on an edge device.

3.2 Algorithm Complexity

An algorithm's complexity measures the number of calculations an algorithm requires as the input size grows. It is calculated by finding the number of multiply-accumulate operations an algorithm requires for an input of size N. Using "Big O Notation", we can then simplify the expression and infer how execution time will increase for a change in N. For example, if there are $2N^2 + N$ basic operations, this is represented using "Big O Notation" as $O(N^2)$. Constant scaling factors and lower powers of N are removed because, for large values of N, the N^2 term dominates (Sipser, 1996). In this example, the number of basic operations to be computed will grow as the square of the input size.

This technique is used for some of the functions below to investigate interesting trends caused by the algorithm's complexity.

3.3 System Overview

The AI-assisted sonification algorithm is composed of an AI engine to detect seizures and a Digital Signal Processing (DSP) block to turn the EEG into sound. The Raspberry Pi processes both components. However, the AI component runs online, outputting the probability of a seizure occurring in an 8s chunk every second. In comparison, the DSP block executes offline, on a long continuous EEG recording and the corresponding AI probabilities. Figure 3.2 shows a high-level signal processing flow diagram.

The AI engine is a Convolutional Neural Network (CNN) (O'Shea et al., 2020), which operates online to continuously detect the likelihood of a patient having a seizure. The EEG data and AI probabilities are then stored in memory until a clinician wishes to review the recording. The DSP block utilises the probabilities to sonify the EEG data, focusing the listener's attention on the segments most likely to contain seizure events. The DSP block processes the data offline.

By pushing the algorithm to the edge, the proposed system can be easily integrated with existing EEG acquisition systems over a serial link (shown in Figure 3.3).

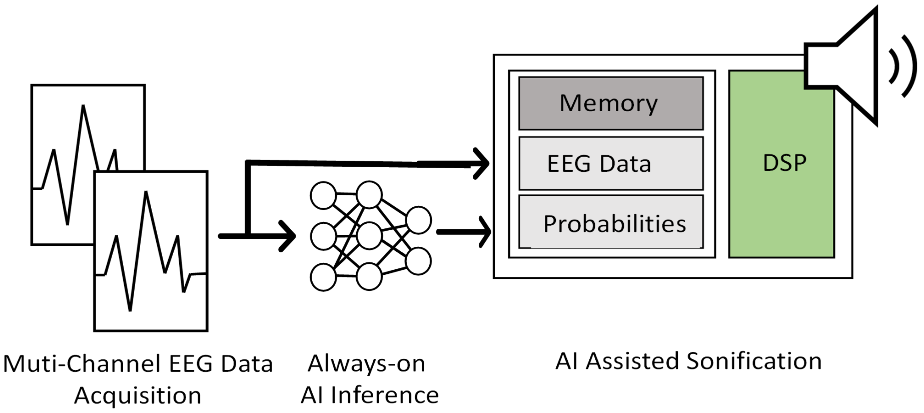


Figure 3.2: High-level sonification overview showing the signal acquisition, inference and signal processing flow

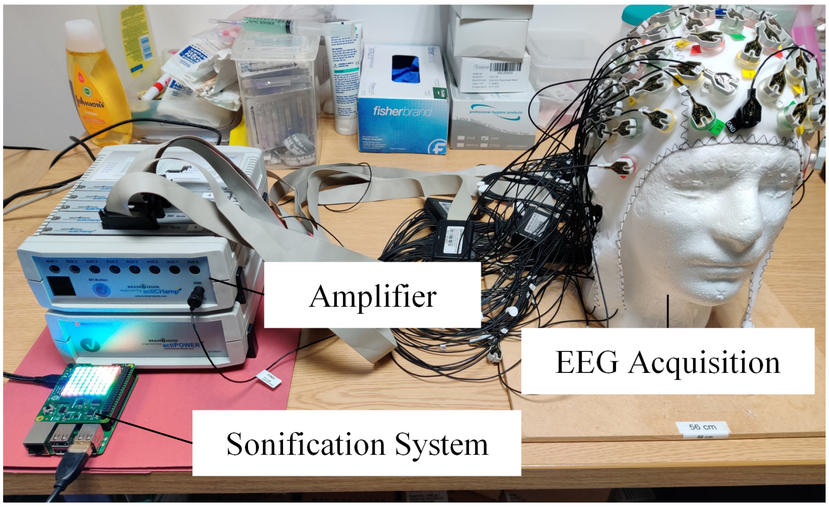


Figure 3.3: System adaptation to an existing EEG acquisition system

3.4 Fully Convolutional Neural Network

A Fully Convolutional Neural Network (FCNN) developed in (O'Shea et al., 2020) is used for the AI engine. Although it is common practice to use convolutional layers early on for feature extraction and fully connected layers as the final classification layers, this architecture uses convolutional layers for feature extraction and classification. The architecture is presented in Figure 3.4. The FCNN was trained on a separate proprietary dataset and later tested on the publicly available Helsinki dataset consisting of recordings from 79 neonates each of 1-2 hours long (Stevenson et al., 2019). On the Helsinki dataset, the model achieved an AUC (Section 5.6) of 95.6%.

3.5 DSP Block

An AI-assisted phase vocoder algorithm that variably compresses and sonifies the EEG signals to allow users to perceive seizures audibly was first

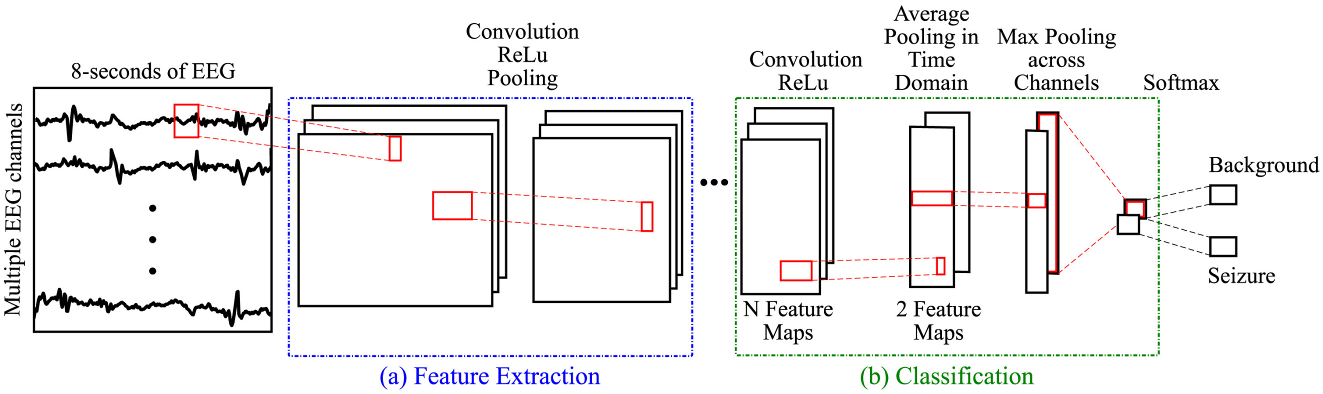


Figure 3.4: FCNN presented in (O'Shea et al., 2020) used to predict the likelihood of a seizure

introduced in (Gómez-Quintana et al., 2021). The phase vocoder maps the inaudible EEG signal frequencies from 0.5 - 13Hz (Kitayama et al., 2003) into the audible range of 20 - 20kHz (Rankine et al., 2006). While the algorithm can use any AI model which provides a probability output, this implementation and analysis use the AI model developed in (O'Shea et al., 2020). A block diagram of the signal processing flows associated with the algorithm in (Gómez-Quintana et al., 2021) is shown in Figure 3.5. All channels are individually processed to ensure a high signal-to-noise ratio. AI Attenuation, Spectral Subtraction and the Phase Vocoder all use the output from the CNN to variably process and focus attention on likely seizure events. The main blocks of the two algorithms are described below.

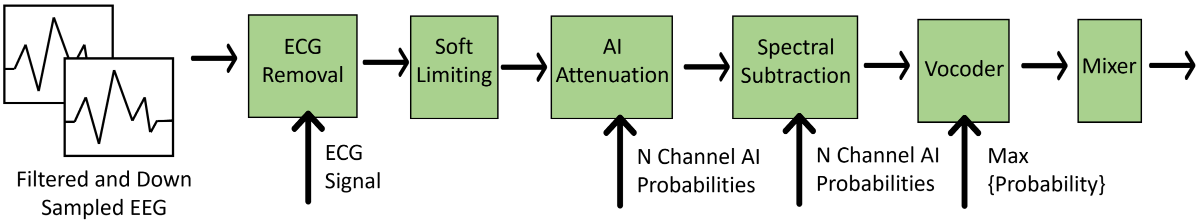


Figure 3.5: Block diagram showing signal processing flow in the DSP block

3.5.1 Filtering

Filtering reduces interference and noise from a signal based on frequency. A digital Infinite Impulse Response (IIR) filter attenuates frequencies outside 0.5 to 13 Hz. The filter calculates its output by mixing a delayed version of the signal and the signal itself with a delayed version of the filter's output. The formula for applying an IIR filter is shown in Equation (3.1).

$$y[t] = \sum_{i=0}^{N} b_i \cdot x[t-i] - \sum_{j=1}^{M} a_j \cdot y[t-j] \tag{3.1}$$

The 50 Hz supply voltage effect is strong, requiring an additional notch filter to attenuate it to the necessary level. The filtering block will be executed after acquisition and before the AI/CNN inference and will be "always-on", continuously processing incoming EEG. The following block then stores the filtered EEG data in memory for later access.

3.5.2 Downsampling

Downsampling lowers a signal's sampling frequency by reducing the number of samples in the signal while maintaining important information. The EEG data is sampled to 256 Hz and is downsampled to 32 Hz during sonification. The original sampling rate is an integer multiple of the desired frequency; hence, decimation can be used. To decimate the signal from 256 to 32 Hz, every 8^{th} sample is taken.

In the first implementation, the downsampling happens in the vocoder block. However, the Nyquist sampling theory states that to represent a signal accurately, the sampling rate must be at least twice its frequency.

To avoid aliasing, a signal must be low pass filtered to at least half the sampling rate before downsampling. Thus the downsampling function can be placed anywhere after the filtering in the algorithm.

3.5.3 ECG Removal

Sonification of ECG artefacts yields a high pitch which can sound similar to a seizure causing a false positive. The artefact's pitch is constant, and with practice, a listener can distinguish the evolving pitch belonging to seizures. To reduce the listener burden, a temporally varying parametric model can be used to attenuate the presence of the interfering ECG signal. ECG amplitudes are much larger than EEG signals (in the mV range versus µV for EEG). ECG may contain the same frequency components as EEG, so it can not be removed using a standard frequency filter. The interference can be modelled with the following parametric equation, considering ECG as a delayed version of the measured ECG with varying amplitude:

$$\hat{S}_{EEG}[t] = S_{EEG}[t] - \alpha \cdot S_{ECG}[t - \delta] \tag{3.2}$$

Where S_{EEG} is the EEG signal plus the interference from ECG; $\hat{S}_{EEG}$ is an estimation of the clean signal;S_{ECG} is the patient's ECG signal. The amount of interference from S_{ECG} is modelled by the parameters α and δ. α determines the magnitude of the ECG to be subtracted based on the strength of interference at that point in time. δ is the time delay between the ECG signal and its interference. Both α and δ vary temporally, so these parameters are recalculated every 8 seconds for 16-second segments. The time delay is estimated using convolution, making this function computationally intensive.

3.5.4 Soft Limiting

Artefacts are commonly introduced to EEG signals due to electrode disconnection or other unwanted electrical impulses. The signal is first normalised by dividing by 100 mV as seizures are expected to occur below this value. A dynamic range compressor is applied to attenuate signals outside of this range.

Dynamic range compression is a parametric model that uses an envelope instead of the signal to reduce distortion while attenuating the signal into the desired range. It is widely used in music to keep a song's volume consistent (Giannoulis et al., 2012).

An envelope is used to obtain a smoothed version of a signal's intensity. The function uses an ADSR (Attack, Decay, Sustain and Release) envelope for intensity estimation. The envelope is applied to the signal by filtering a kernel (the envelope's impulse response) over the signal's instantaneous power. The general form of the ADSR's impulse response is shown in Figure 3.6.

The shape of the impulse response can be defined as:

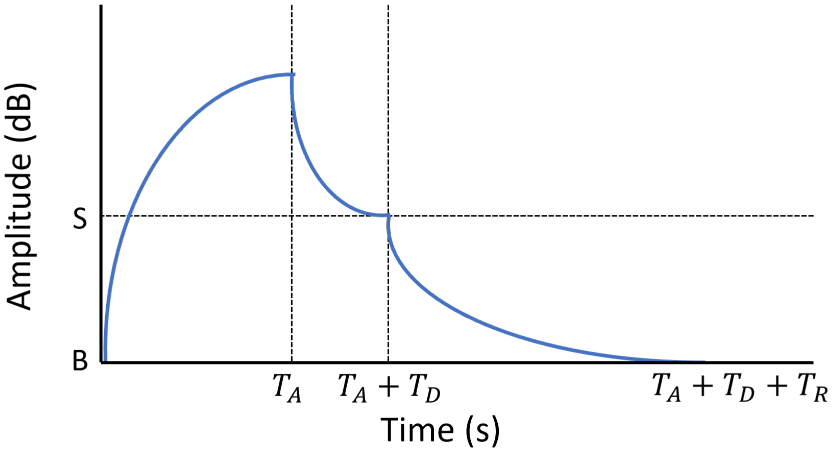


Figure 3.6: General form of the impulse response for an ADSR envelope with Attack, Delay, and Release times (T_A, T_D, T_R) and the Sustain level S

$$h[t]_{dB} = \begin{cases} -\frac{B}{T_A}t + B & t \leq T_A \\ \frac{S}{T_D}(t - T_A) & T_A < t \leq T_D + T_A \\ \frac{B-S}{T_R}(t - T_A - T_D) + S & T_D + T_A < t \leq T_D + T_A + T_R \end{cases} \tag{3.3}$$

The ADSR parameters T_A, T_D, T_R determine the kernel length. To apply the kernel as a digital filter the sampling frequency, f_s, is used to convert the parameter from seconds into samples. Therefore the total number of coefficients, N, is equal to the length of the impulse response in samples:

$$N = (T_A + T_D + T_R) \cdot f_s \tag{3.4}$$

By examining the equation for an IIR filter (3.1), it can be seen that for each data sample, where there is a total of N coefficients, there are $N \cdot N$ real multiplication operations and $N \cdot (N + 1)$ real additions. The filter's complexity can be represented using big O notation as $O(N^2)$. Since the complexity is high, if a signal has a high sampling rate, execution times for this function may be long.

3.5.5 AI Attenuation

The EEG signal is attenuated based on the predicted likeliness of a seizure occurring at each point by the CNN. The attenuation factor at each point is proportional to the probability of seizure. A segment with a probability $p[n] = 1$ will be unchanged. The maximum attenuation occurs when $p[n] = 0$ and is determined by the variable Max. Max is a user-set parameter and was chosen to be 20 dB for these experiments. Equation (3.5) shows how this is applied to the input EEG $x[n]$.

$$y[n] = x[n] \cdot 10^{\left(\frac{Max}{20}\right)\cdot(p[n]-1)} \tag{3.5}$$

Variably attenuating the signal increases the clarity of the audio, making seizures easier to detect as non-seizures are attenuated.

3.5.6 Spectral Subtraction

Spectral subtraction aims to remove the background EEG from the signal in the frequency domain, making the seizure content clearer. Background EEG is normal EEG, not containing any seizures. The contribution of background EEG to the spectral profile is calculated using the probabilities for each window, $p_{seizure}$, obtained for the signal, $x(t)$, and by converting the signal into the frequency domain using the Short Time Fourier Transform

(STFT). Where $X[n,k] = STFT\{x[t]\}$, $Pxx[n,k] = |X[n,k]|^2$, and the background EEG's spectral contribution, $T[k]$, can be calculated as:

$$\mathrm{T[k]} = \sum_n P_{xx}[\mathrm{n,k}] \frac{1 - p_{seizure}[\mathrm{n}]}{\Sigma_n (1 - p_{seizure}[\mathrm{n}])} \tag{3.6}$$

The gain to be applied to the input signal can be calculated as follows:

$$G[n,k] = \begin{cases} 1 & \text{if } P_{XX}[n,k] \geq T[k] \\ \frac{P_{XX}[n,k]}{T[k]} & \text{otherwise} \end{cases} \tag{3.7}$$

The gain can then be applied (Equation (3.8)). The first term increases the attenuation factor by a maximum of 10 % (when the probability of seizure is zero).

$$\mathrm{Y[n,k]} = 10^{\mathrm{p[n]}-1} \cdot \mathrm{G[n,\ k]} \cdot \mathrm{X[n,k]} \tag{3.8}$$

If the effect is more significant than the seizure's, the signal is attenuated proportionally to the likelihood of seizure. Otherwise, the signal is left unchanged. The inverse STFT can then be calculated to synthesise the denoised time series signal.

3.5.7 Variable Speed Phase Vocoder

The phase vocoder is an essential block in the sonification process. It was first used to shift the EEG's frequency into the audible domain without any temporal compression (Temko et al., 2014b). Additionally, the phase vocoder can variably compress the signal, focusing a listener's attention on seizure events. In this work, the Vocoder is used on EEG data. However, it was initially proposed for time-stretching speech signals (Flanagan and Golden, 1966).

The magnitude of the signal is linearly interpolated at a varying compression rate. After altering the frequency content, the phase difference between points in the input signal is measured to keep horizontal coherence between the phase neighbouring Fast Fourier Transform (FFT) bins.

Variable compression is achieved by taking the probability of the EEG containing a seizure, as generated by the AI inference, and allotting more listening time to predicted seizure events. Sections of the EEG where no seizure is present are compressed more than where seizures are likely to be present. The compression ratio, $R[n]$, can be calculated based on the ratio of the input sampling frequency, Fs_{in} , the output sampling frequency, Fs_{out} and the AI dependent variable speed parameter, $VS[n]$:

$$R[n] = \frac{Fs_{\text{out}} / Fs_{\text{in}}}{VS[n]} \tag{3.9}$$

Equation (3.10) gives the formula for calculating the relative playback speed, VS[n], at a point in time.

$$VS[n] = Vmax \left(\frac{Vmin}{Vmax} \right)^{AI[n]} \tag{3.10}$$

The minimum compression factor, Vmin, is 60, while the maximum, Vmax, is 3600. The values for Vmin and Vmax were previously selected and experimented on in (Gomez-Quintana et al., 2022). AI[n] varies the compression factor according to the likeliness of a seizure at each point in time. This variable compression directs the listeners' attention to where it is most needed allowing efficient and accurate detection of seizures.

After the variable speed is calculated the output duration, T_o, can be calculated from the input duration, T_i:

$$T_o = T_i \frac{1}{N} \sum_{n=0}^{N-1} VS[n] \tag{3.11}$$

T_i and T_o can be measured in either STFT frames or seconds.

Equation (3.11) has important consequences when measuring the execution time of the function, as the number of outputted STFT frames is calculated on the variable speed. It follows that the function's complexity depends not only on the length of the input sequence but also on variable speed (which depends on the value of the AI probabilities).

It can be easily shown that the complexity of internal functions is linearly dependent on the number of output STFT frames. The complexity of the vocoder, taking the STFT frames as input and where T_o is measured in STFT frames, is $O(T_o)$. Substituting (3.11) into this equation and removing constants, the complexity is shown to be dependent on AI:

$$O(T_o) = O\left(T_i \sum_{n=0}^{N-1} \left(\frac{Vmin}{Vmax} \right)^{AI[n]} \right) \tag{3.12}$$

The expansion and compression rate (ER, CR) in the Vocoder is given as follows:

$$ER = \frac{1}{CR} = \frac{T_o}{T_i} = \frac{1}{N} \sum_{n=0}^{N-1} VS[n] \tag{3.13}$$

Even for patients with a high seizure burden, most of the EEG recording is considered non-seizure. The imbalance between seizure and non-seizure events generally leads to temporal compression, where $T_o < T_i$.

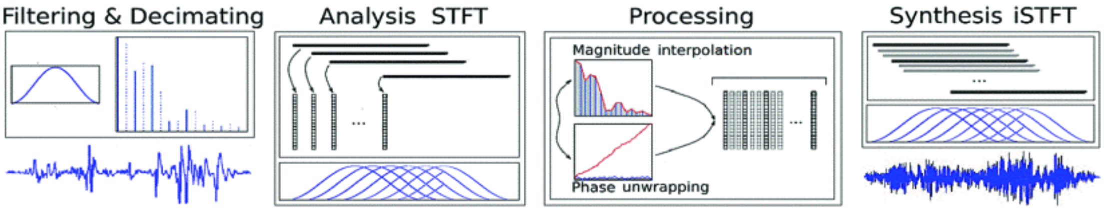


Figure 3.7: Signal processing flow in the Phase Vocoder (Gomez et al., 2018)

3.5.8 Mixer

The role of the mixer is to reduce an eight-channel EEG signal to stereo sound for review. A delay is introduced between the left and right ear, known as the Haas delay, which allows a listener to determine the location of the electrode/seizure in the neonate's brain. It has been shown that using this approach, clinicians can differentiate between the left, central and right hemispheres of sonified pairs of electrodes (Middlebrooks, 2015). Theoretically, it could be used to locate seizures anywhere on the head.

3.6 Optimisations

Some practical optimisations can be made without altering the structure of the algorithm to decrease execution time while leaving functionality completely unchanged.

3.6.1 ECG Removal

ECG removal is one of the most computationally intensive functions, as the time delay between the interfering reference ECG signal and EEG signal varies temporally and must be estimated periodically using a convolution. In the first implementation, the signals were split into 16s segments, then iteratively converted to the frequency domain using an FFT and convoluted together to find the time delay. Computing FFTs and convolutions in parallel can significantly speed up these processes. When computing iteratively, all the computing resources are assigned to one FFT even if not all are needed, and the next can not begin until the first is finished. In parallel computing, resources are shared over all FFTs at the same time. Therefore, computing resources are used efficiently and execution time reduces. Converting the time series data to matrices facilitates this process (Figure 3.8).

The ECG and EEG are converted to matrices with the FFT window size and the number of frames as the matrix's dimensions. The FFT can be taken of the whole signal in the matrix, and the resultant matrices can be convoluted by multiplication. The remaining steps in the process described in section 3.5.3 can be executed in matrix form.

3.6.2 Early Down-sampling

The input signal was not downsampled in the MATLAB implementation until the phase vocoder. The signal must only remain at the 256 Hz sampling rate during the filtering stage to avoid aliasing. By downsampling to a

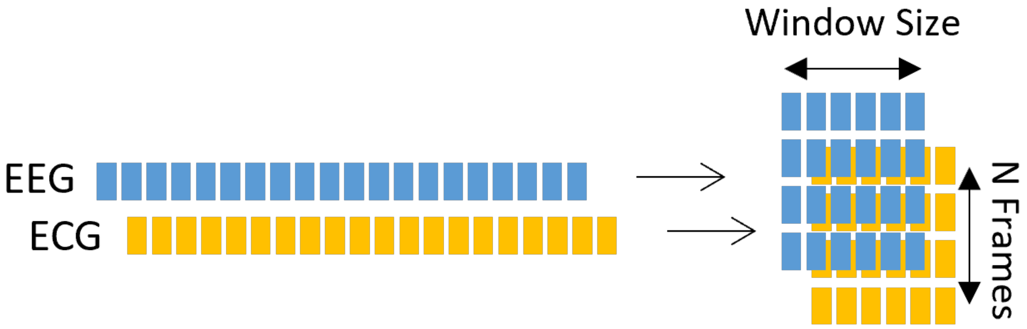


Figure 3.8: Conversion of an EEG channel and an ECG signal from 1D time series signals to a matrix representation

32 Hz signal directly after, all the proceeding stages need to process 8 times fewer data samples. For algorithms with high complexity, it is particularly beneficial in reducing execution time. Moving downsampling earlier means that the CNN and the DSP block have the exact same preprocessing steps, preventing the system from doing the same operation twice.

3.7 Results

Execution time results were recorded on a Raspberry Pi 3B+. The Pi's CPU has a max frequency of 1.4 GHz (RaspberryPi, 2022), although it is dynamically varied to as low as 600 MHz to save on power consumption. In order to yield more consistent and reproducible results, the clock is fixed to a frequency of 800 MHz. Timing results are presented for three implementations with increasing optimisation levels in Table 3.1. The first implementation has no optimisation and is directly translated from the MATLAB code. The second implementation includes an improved ECG removal function. The final implementation down samples immediately after filtering to reduce the number of data samples to be processed.

To verify that these optimisations have not affected the algorithm, spectrograms generated using the final implementation are presented for both a seizure and non-seizure recording.

3.7.1 Average Processing Time

To reduce percentage error while measuring a function's execution time, the Average Processing Time (APT) for 1 hour of EEG is measured and averaged over ten iterations.

The execution time for every function except the Vocoder depends only on the length of the inputted data. The phase vocoder has a variable length output, and the number of computations the phase vocoder has to do depends on the values of the AI probability (Equation (3.12)). Therefore the APT varies greatly from file to file. To get a consistent measure, the APT of the phase vocoder must be taken over the entire Helsinki dataset and normalised to 1 hour. APT results are shown in Table 3.1.

3.7.2 Verification

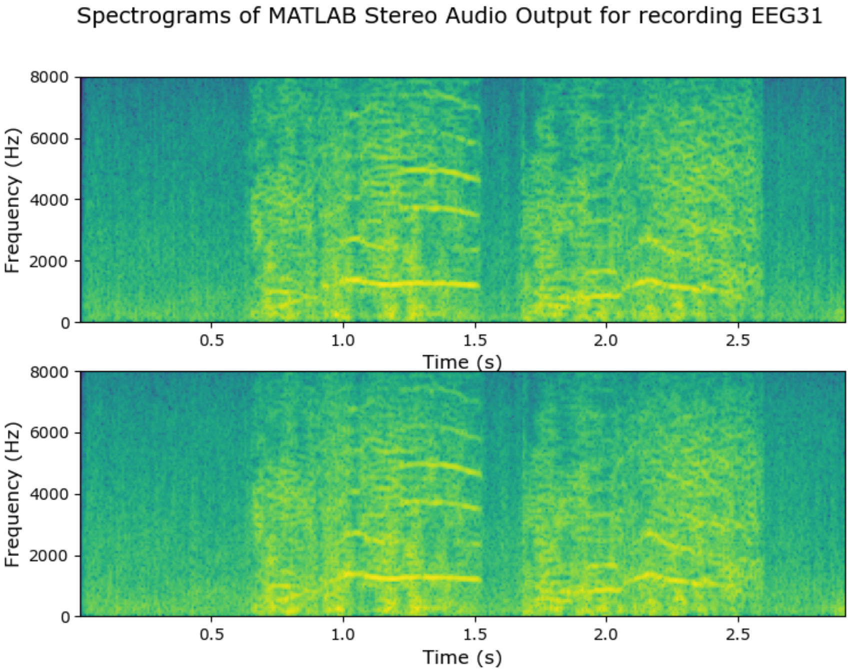
This work was translational, and the output of optimisations should give a 1:1 mapping when compared with the MATLAB implementation. The changes can be verified objectively by plotting the audio output of both implementations. A spectrogram is a visual representation showing how a signal's frequency content changes over time. It is calculated by taking the Short Time Fourier Transform (STFT) of a signal which consists of many

Table 3.1: Average processing time for 1 hour of EEG with both algorithms on the Raspberry Pi 3b+

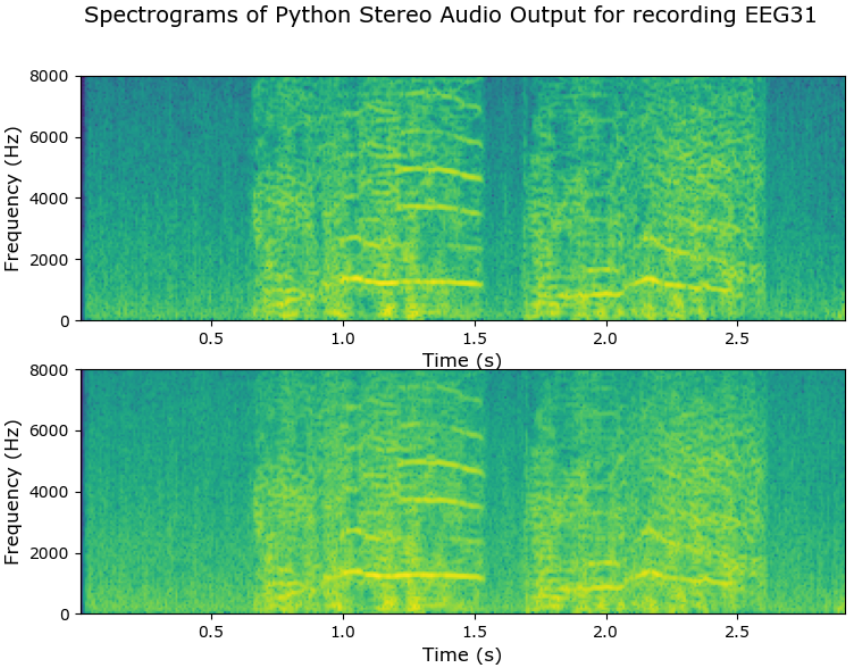
	First Implementation (s)	ECG Removal (s)	Resampling (s)
ECG Removal	191.76	28.24	3.41
Mixer	5.77	-	0.58
AI Post Processing	0.29	-	0.29
Spectral Subtraction	62.12	-	9.84
Attenuation	10.41	-	0.99
Vocoder	12.09	-	12.09
Limiting	126.64	-	2.41
Total:	409.08	245.56	29.61

overlapping Fast Fourier Transforms (FFT). Sonified seizures are differentiated from background EEG due to their high pitch sound with evolving frequency. Hence spectrograms are the ideal way of visually interpreting the audio, as frequency can be observed over time.

Figures 3.9 and 3.10 show the spectral output for two files, one containing a seizure and one not. Figure 3.9 shows EEG file 31, which contains two distinct seizures. While 3.10 shows the output for EEG file 3, which contains no seizures. In each figure, (a) displays the MATLAB output while (b) shows Python. The spectrograms for both files show minimal differences, and it can be concluded that algorithmic performance is not affected. Spectrograms are displayed here for two files to illustrate that the optimisations to the algorithm do not affect performance. However, while testing the spectral output was systematically reviewed for every file in the dataset.

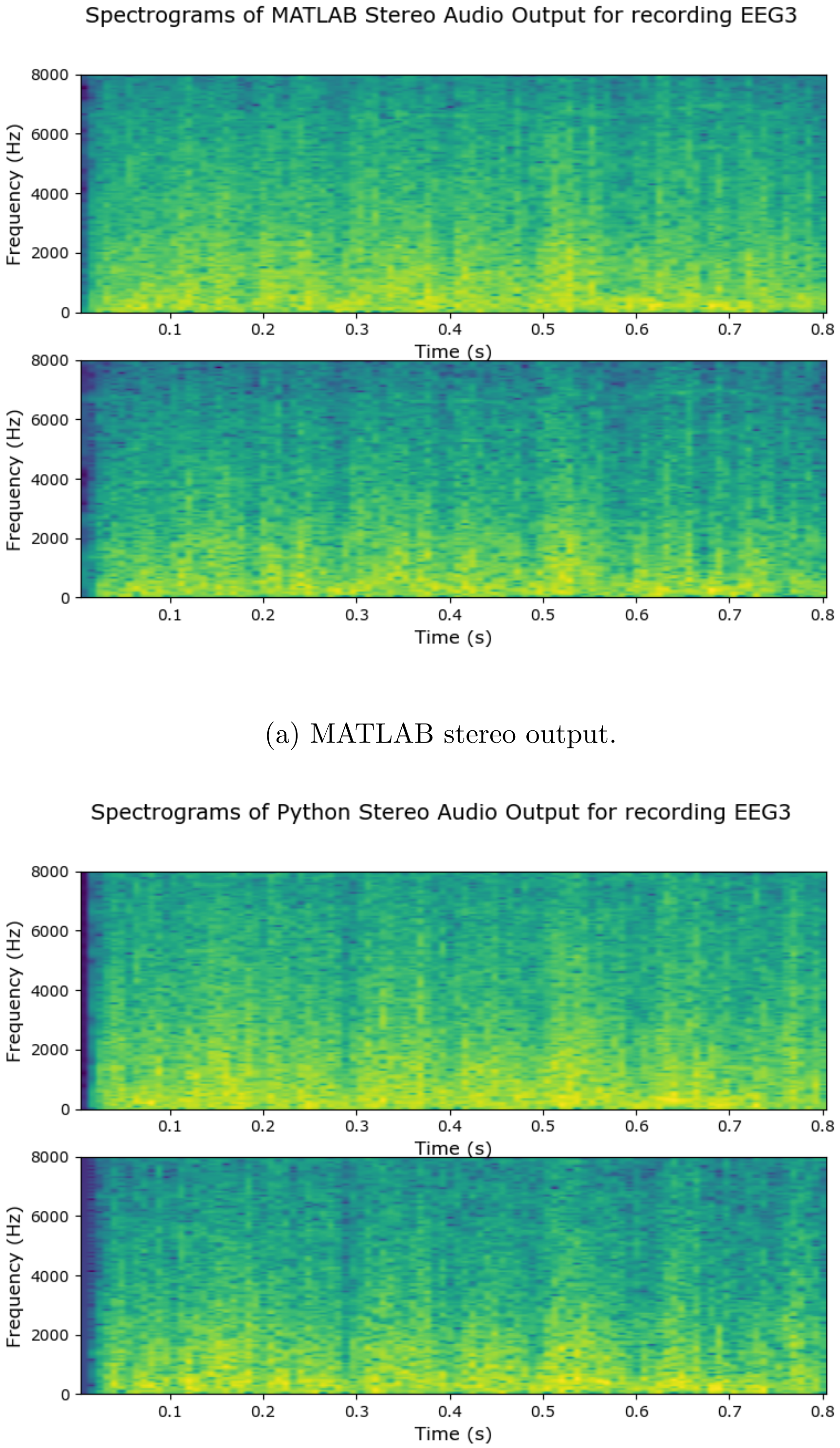


(a) MATLAB stereo output.



(b) Python stereo output.

Figure 3.9: (a) MATLAB generated audio and (b) Python generated audio for an EEG recording containing seizures.



(a) MATLAB stereo output.

(b) Python stereo output.

Figure 3.10: (a) MATLAB generated audio and (b) Python generated audio for an EEG recording containing no seizures.

3.8 Discussion

Examining the first implementation's results, it becomes clear that the ECG removal function is a bottleneck for performance. It takes up 46 % of the total execution time. Optimisations made to this function improved its execution time by 6.8x and the total execution time by 1.7x. Moving the downsampling function earlier reduces the overall execution time of the algorithm by 8.3x. However, it reduces the execution time of the soft limiting function by 52.5x. This improvement can be explained by the high complexity of the function, $O(N^2)$, as derived in section 3.5.4. As shown in Equation (3.4), downsampling the signal by a factor of 8 reduces the number of coefficients, N, by the same factor.

Although the APT for the algorithm was reduced by 13.8x, it still takes an average of 29.61 seconds to sonify an hour of EEG. Moreover, on average, the sonification will compress 1 hour of EEG into a 4.9-second audio clip. The most optimised version of the algorithm takes an average of 6x longer to process EEG than for a clinician to review the generated audio. The sonification algorithm was designed to facilitate a fast review of many hours of EEG. This wait time is unacceptable in a busy hospital environment.

The contribution of each function to the overall APT is shown in Figure 3.11. The Vocoder makes up 41 % of the execution time. It is the core block in the sonification process, and this is proportionate to its value. The subsequent three slowest blocks are Spectral Subtraction (33%), ECG Removal (12%) and Limiting (8%). These steps are non-essential and are included in the algorithm to improve audio clarity. In all cases, alternative methods should be investigated to determine if performance can be maintained by using simpler versions or removing the function.

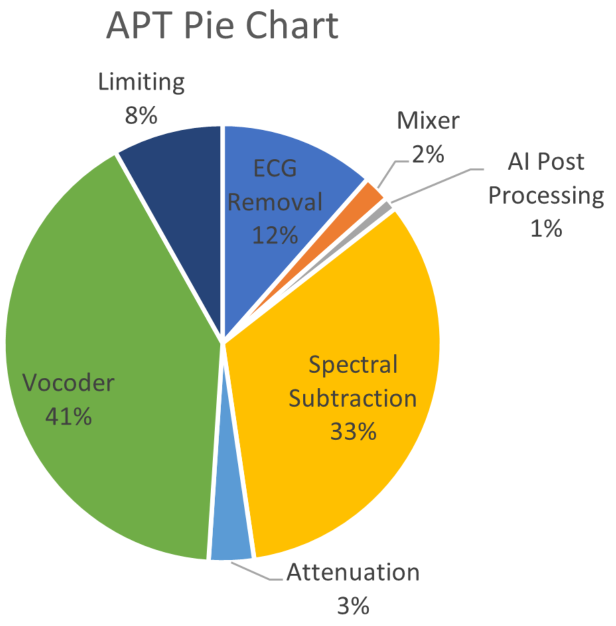


Figure 3.11: Pie chart showing the APT of each function as a proportion of the overall APT

Moving the mixer earlier and operating using single-channel processing should be investigated to reduce the execution time of all functions, including the phase vocoder. Following this framework, a version of the algorithm more suited for the resource constraints of an edge device could be developed.

3.9 Conclusion

A first implementation of a seizure sonification algorithm on an edge device is presented. The first implementation was slow, so several optimisations were made. ECG removal was sped up by unwrapping an iterative loop to produce results in parallel, resulting in a 1.7x reduction in total execution time. The input signal was downsampled earlier to reduce computational load. The reduced input decreased the algorithm's average execution time by 8.3x and the limiting function's APT by 52.5x due to its high complexity. The optimisations lead to a total reduction in execution time of 13.8x. The final implementation was validated by comparing its spectral output against the original MATLAB implementation.

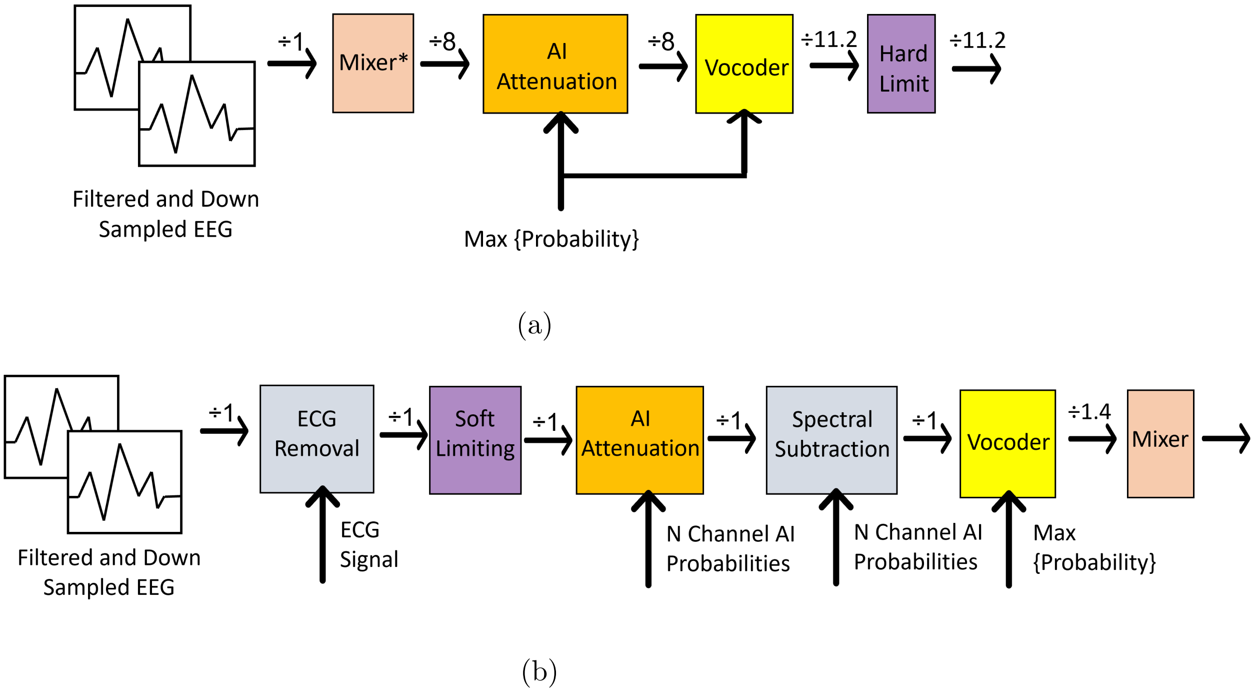
Although significant improvements in overall execution time were made, the program still has a significant execution time. For pervasive use of the sonification algorithm, this must be reduced.

4

Modified Algorithm Towards Lightweight Implementation

The previous chapter optimised the first implementation to reduce execution time by 13.8x. Even with these optimisations, it takes an average of 29.62 seconds to sonify 1 hour of continuous 8-channel EEG. A clinician may have to review several hours of EEG for many patients in the NICU. In this case, sonification will take in the order of minutes to process the raw EEG for review. Despite the significant improvement in execution time presented in the previous chapter, it takes 6 times longer to process EEG than to review the generated audio. To improve the processing further, the algorithm's structure must be changed.

Data compression already occurs in the algorithm and can be exploited to reduce the number of data samples processed at each stage. Data compression occurs in the phase vocoder, which compresses the data temporally by an average of 1.4x over the Helsinki dataset, and also in the mixer, which



(a)

(b)

Figure 4.1: Block diagram of (a) the proposed lightweight algorithm and (b) the original algorithm showing the accumulative data compression at each stage

reduces the number of channels to produce stereo sound (two channels). All other stages can run faster by rearranging the algorithm to reduce data at more computationally intensive stages. Moving these blocks to the start of the algorithm significantly reduces processing time. In cases where functions still run slowly, suitable alternatives are considered.

The following chapter focuses on utilising these ideas to propose a lightweight algorithm. The lightweight algorithm is compared with the original algorithm in terms of average processing time and classification performance (as measured by a survey). An in-depth statistical analysis is performed to evaluate differences in the surveys.

4.1 Overview

An overview of the two algorithms is shown in Figure 4.1 (a) and (b) for the lightweight and original algorithms, respectively. The algorithms are colour coded to make it easy to identify changes. Corresponding functions have the same colour in both algorithms, while functions which are removed entirely are grey in 4.1 (b). The accumulated data compression is presented at the inputs and outputs of each block. Details of the changes and the added data compression follow.

4.2 Data Compression

Although both data and temporal compression reduce the duration of the output audio, it is important to distinguish between them. Temporal compression occurs either in the phase vocoder or by increasing the playback speed of the audio, essentially decreasing the length of the output in seconds. Data compression reduces the number of data points in the signal, decreasing the length of the output in samples. Data compression is accomplished by channel reduction (in the mixer), compression in the phase vocoder, and downsampling of the signal.

The input EEG signals are 1-2 hours in duration and are downsampled to 32 Hz across all 8 channels. Processing this large quantity of data is computationally intensive. The data can be compressed significantly by rearranging some of the signal processing blocks. Reducing the number of data points leads to a proportionate decrease in the average processing time of the algorithm.

In Figure 4.1, the accumulative compression is shown relative to the algorithm's 8-channel 32 Hz input signal at all points. There is no data compression in the original algorithm until the last two stages of the algorithm (the mixer and the phase vocoder). The late data compression results in the most computationally intensive blocks processing the most data. The phase vocoder variably compresses the EEG signal to draw the listener's attention to likely seizures resulting in a temporally warped signal with average data compression of 1.4x over the Helsinki data set. The mixer reduces the number of channels in the signal, going from 8-channel EEG to 2-channel stereo sound. These blocks can be rearranged to significantly speed up the algorithm.

4.3 Simplifications

Some functions in the original algorithm were 'over-engineered', and simpler alternatives can be implemented, which, although theoretically worse, within the practical constraints of this problem, their performance is similar for binary classification of seizures.

4.3.1 Mixer

In the modified version of the algorithm the mixer is moved first to reduce the data to be processed. The original mixer uses the Haas delay, adding a slight delay between the left and right ears, allowing for spatial location of the seizure in the neonate's brain. This delay puts the AI probabilities out of sync with the EEG signal and, therefore, had to be removed to move the mixer first. This functionality does not affect the binary classification results used for performance evaluation. Without the Haas delay, stereo sound is no longer advantageous; hence mono mixing is used to decrease the number of EEG channels further to one. As such, the lightweight algorithm no longer processes each channel individually. In figure 4.1, the lite algorithm's mixer is referred to as mixer* to differentiate between the two mixers.

4.3.2 Limiting

Artefacts are commonly introduced to EEG signals due to electrode disconnection or other unwanted electrical impulses. Seizures are expected to occur below 100 mV with these artefacts resulting in high amplitude spikes. In the original algorithm, the soft limiter reduces the signal gain progressively before reaching this threshold to reduce the signal distortion. In the lightweight algorithm, a hard limit is used, where values outside the limit are set equal to the limit. Hard limiting results in increased signal distortion. However, seizure EEG occurs below the limit and only affects artefacts. The distortion is known as clipping, where values above the limit become a flat line at the limit (Figure 4.2).

4.4 Function Removal

In cases where slow functions were deemed ineffective in comparison to their computational costs, they were completely removed. This only applied to two functions where similar processing had already been undertaken and it was absolutely necessary to improve performance.

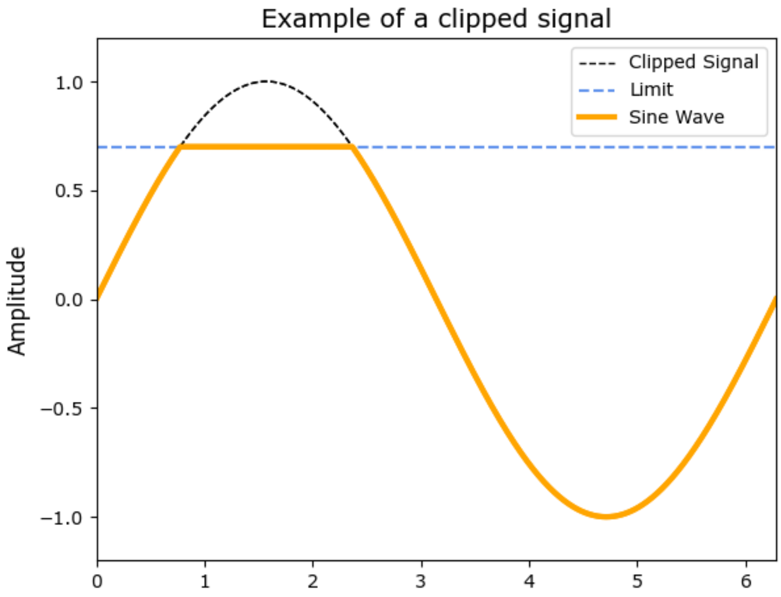
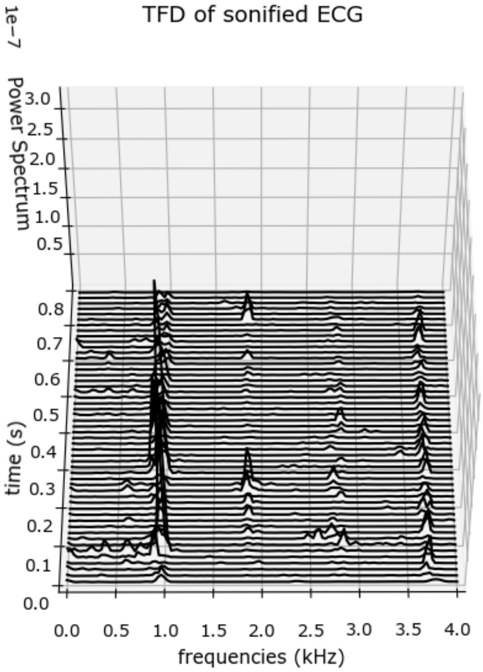


Figure 4.2: Example of a sine wave being clipped above a limit of +0.7

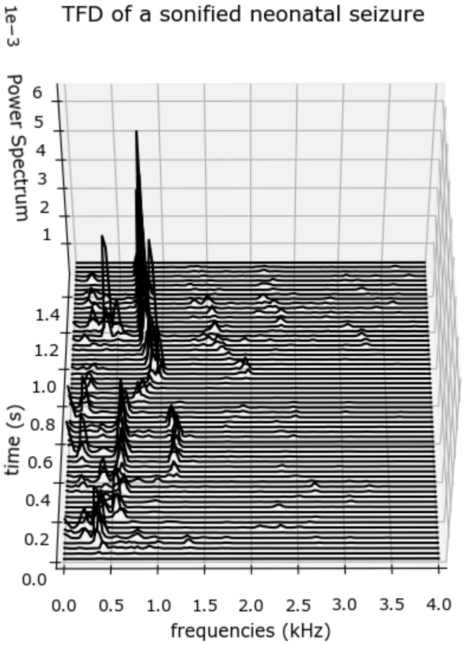
4.4.1 ECG Removal

Sonified ECG artefacts present as high-pitched notes, which can sound similar to seizures to untrained individuals. However, a listener can easily distinguish between ECG interference and seizure patterns because the frequency of seizures evolves over time.

The differences in frequency evolution are observed empirically when sonified ECG and seizures are presented using time-frequency distribution (TFD) plots. Figures 4.3(a) and (b) are obtained by plotting the square of the STFT magnitude in three dimensions. They show the TFD of sonified EEG, but this is a feature of the raw signal, not the sonification process.



(a) TFD of sonified ECG artefact



(b) TFD of sonified seizure

Figure 4.3: (a) TFD of sonified ECG artefact (b) TFD of sonified seizure

4.4.2 Spectral Subtraction

Spectral subtraction was removed as it was observed not to have an effect that justifies its execution time. The AI Attenuation function still processes the signal based on the AI probabilities, and most of the benefits of AI-based processing can be maintained with only one function.

4.5 Experimental Setup

The algorithm's effectiveness is subjective as it involves interpretation from a human listener. Hence, a human must be involved in the final testing process to verify that the aforementioned changes do not affect the algorithm's classification performance. The subjective classification performance is measured with a survey. Execution times of the lightweight algorithm's functions are also measured and compared to that of the originals. Timing and survey results are obtained on the Helsinki dataset. A detailed explanation of the methodology used follows.

4.5.1 Dataset

The publicly available Helsinki dataset was used to conduct the survey. It consists of 79 full-term neonates from Helsinki University Hospital NICU (Neonatal Intensive Care Unit) (Stevenson et al., 2019). There is a single recording for each baby with a median duration of 74min (IQR: 64 to 96min) with a total of 112 h of EEG recordings in the dataset. Three expert doctors independently annotate each 1 second segment containing either a seizure/non-seizure. Where the experts disagreed majority vote was taken.

An average of 460 seizures were annotated by each expert, with 39 babies having seizures by consensus. The EEG signals have a 24 bit resolution and were sampled at 256 Hz.

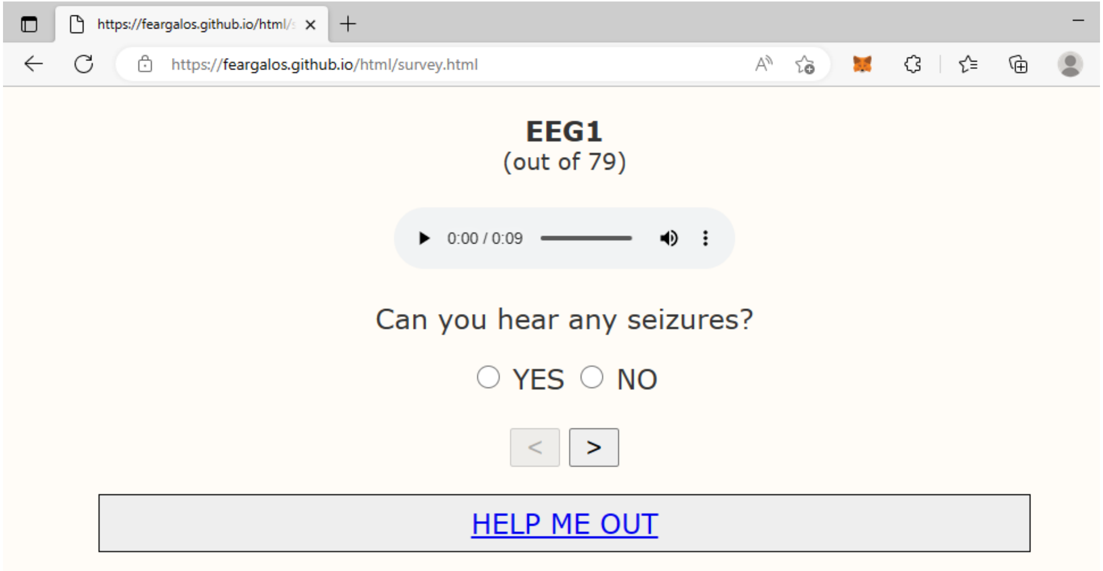
4.5.2 Average Processing Time Methodology

Testing was done using the same process outlined in the previous chapter. When measurements were taken, the Raspberry Pi was used with a clock frequency fixed at 800 MHz. The average processing time (APT) results for implementing the two algorithms are presented in Table 4.1. The performance of every block except the vocoder is solely dependent on the number of EEG samples to be processed. The execution time of the phase vocoder depends on a patient's seizure burden. Therefore it is necessary to get results over the whole dataset and normalise for 1 hour of EEG.

4.5.3 Survey Methodology

An audio file was generated from each of the 79 EEG recordings in the Helsinki database. Participants listened to each audio file and were then asked if they heard a seizure in the file or not. The majority vote was then taken amongst participants to make a classification. Figure 4.4 shows the simple user interface used to carry out the survey. It was identical for both surveys.

The original AI-sonification algorithm described in Chapter 3 was verified in (Gomez-Quintana et al., 2022) using a survey of over 30 participants, including a medical and non-medical cohort. Changes to the algorithm in the development stage were evaluated using small in-lab surveys (n =


https://feargalos.github.io/html/
https://feargalos.github.io/html/survey.html
0:00 / 0:09
<
>
HELP ME OUT

EEG1
(out of 79)

Can you hear any seizures?

○ YES ○ NO

Figure 4.4: Screenshot of an example survey question

5). Similarly to the exploratory surveys, the survey to validate the Lite algorithm used a small group of non-clinical participants (n=7). In order to reduce statistical variance from comparing two groups (the original 30 and the new 7 participants), the survey was repeated on the same group using both the original and proposed lightweight algorithm.

The survey participants were split into two groups, A and B. The order in which the groups did the survey was stratified, so Group A listened to the output of the adapted algorithm first, while Group B did the survey based on the original algorithm first. The participants in each group took the other survey several days later. This was done to mitigate the effects of fatigue and novelty on results when comparing the performance of the two algorithms.

4.5.4 Evaluation Metrics

The execution time, AI performance and survey results must all be interrogated using robust metrics. Although the AI performance and execution time are measured similarly in other chapters, some statistical measures must be introduced to properly interpret the survey results. The measures used are as follows.

4.5.4.1 Sensitivity and Specificity

The survey results are interpreted by taking a majority vote of all participants. The dataset is imbalanced, containing more samples with seizures than not. Sensitivity (a measure of true positive rate) and specificity (a measure of false positive rate) will be used instead to measure performance better. Equations (4.1) and (4.2) show the respective formulas.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4.1}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \tag{4.2}$$

4.5.4.2 Variance

Variance is a measure of the spread or dispersion of a data set and is used to show the spread of each participant's performance. It is calculated as the average squared deviation of each data point from the data set's mean. A significant variance indicates that the data points are spread out over a wide range, while a small variance indicates that the data points are concentrated closer to the mean. It is useful for understanding data distribution and comparing different data sets' dispersion. The formula for variance, σ^2, is given in Equation (4.3), where x is a data point, μ is the mean of the data set, and N is the number of data points in the set.

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} \tag{4.3}$$

4.5.4.3 Binary Cross Entropy Loss

Binary Cross Entropy (BCE) is commonly used to measure the loss of model predictions in binary classification problems (Ruby and Yendapalli, 2020). It measures dissimilarities between predictions and labels. It is given by the formula:

$$BCE = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot \log \hat{y_i} + (1 - y_i) \cdot \log (1 - \hat{y_i}) \tag{4.4}$$

Where y_i is the label and $\hat{y_i}$ is the maximum probability of a seizure occurring in any channel, ch, at the time, t:

$$\hat{y_i} = \max_t \{AI[t]\} = \max_t \left\{ \max_{ch} \{AI[ch, t]\} \right\} \tag{4.5}$$

4.5.4.4 Cohen's Kappa

The Kappa statistic is used to measure the inter-rater agreement level between annotators (McHugh, 2012). For this work, the raters are the majority vote from the survey participants. Specifically, when measuring reliability amongst two annotators, Cohen's Kappa is used. A score of 1 indicates perfect agreement, while a score of -1 indicates perfect disagreement. The Kappa statistic is defined by the probability of agreement, p, and the probability of agreement by chance, p_e:

$$\kappa = \frac{p - p_e}{1 - p_e} \tag{4.6}$$

And SD_κ is calculated as:

$$SD_\kappa = \sqrt{\frac{p(1-p)}{(1-p_e)^2}} \tag{4.7}$$

The confidence interval can be calculated with the following expression:

$$CI(\alpha) = \kappa \pm Z_S(n, \alpha) \cdot \frac{SD_\kappa}{\sqrt{n}} \tag{4.8}$$

Where $Z_S(n, \alpha)$ is the z score for n = 79 patients and α is the estimation error for a 95% confidence interval (=0.05).

To compare agreement levels between different sets of raters, it is necessary to measure the statistical significance to ensure that the two kappa scores are similar. The t statistic is used to compare distributions with the Kappa means and variance. When both distributions have the same number of samples, the t statistic is given by:

$$t = \frac{\kappa_1 - \kappa_2}{\sqrt{\frac{SD_{\kappa_1}^2 + SD_{\kappa_2}^2}{n}}} \tag{4.9}$$

Where κ_i and SD_{κ_i} are the mean and SDs of the i^{th} distribution, and n is the number of samples. The p-value is calculated from the t statistic using the inverse of the cumulative density function.

4.5.4.5 Area Under the Curve (AUC)

AUC is commonly used to measure a models predictive power based on the ability of the probabilistic output to measure how separable two classes in a binary classifier are. The classification threshold is varied between 0 and 1 and the sensitivity and specificity metric are calculated. Plotting the sensitivities and specificities gives the ROC curve.

In this work it is used to measure the discriminating power of different features on classifications from the surveys. When used in this way it can be interpreted similarly to a correlation metric. A value close to one indicates the feature was highly correlated with the associated outcomes, while an AUC close to 0.5 indicates no correlation.

4.6 Average Processing Time Results

The average processing time was measured to obtain a consistent measure of the improvements of overall execution time after the algorithms modifications. Results are shown in Table 4.1.

Figure 4.5 shows the APT of each function as a percentage of the corresponding function in the original. It gives a visual illustration of the individual contribution of each function to the overall speedup. AI Post Processing was unaffected, while the Limiting APT decreased more than 200 times. Overall, the lightweight algorithm is an order of magnitude faster than the original. This speedup proportionally decreases energy consumption and drastically increases the system's responsiveness. The 13.3x

Table 4.1: Average processing time for 1 hour of EEG with both algorithms on the Raspberry Pi 3b+

Function	**Original (s)**	**Lightweight(s)**
ECG Removal	3.41	-
Mixer	0.58	0.37
AI Post Processing	0.29	0.29
Spectral Subtraction	9.84	-
Attenuation	0.99	0.11
Vocoder	12.09	1.44
Limiting	2.41	0.01
Total:	29.62	2.23

decrease in APT is due to data compression and simplifications.

4.7 Survey

The survey results are interpreted using the previously defined statistical measures in terms of agreement and absolute performance. The answers of the 7 participants were combined by means of a majority vote and compared to the expert's annotations.

4.7.1 Survey Results

Table 4.2 shows the confusion matrix calculated using this methodology for both surveys. The confusion matrix presents the actual number of seizures and non-seizures vs the predicted quantities, hence showing the number of True Positives, False Positives, True Negatives and False Negatives.

The sensitivity and specificity are derived from the confusion matrix for each algorithm are presented in Table 4.3. The lightweight and original

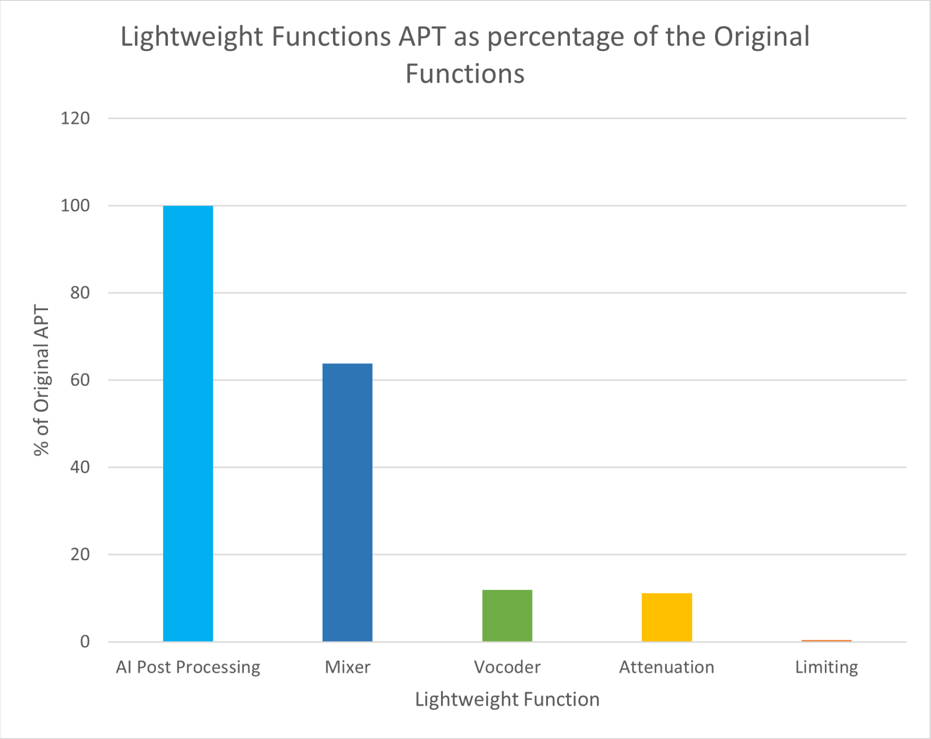


Figure 4.5: APT of corresponding functions in the original and lightweight algorithm normalised to respective function's original APT

Table 4.2: Confusion matrix of survey results for both the lightweight and original algorithm

		Predicted			
		Original		Lightweight	
		Seizure	Non-Seizure	Seizure	Non-Seizure
Actual	Seizure	40	6	40	6
	Non-Seizure	6	27	5	28

algorithms show the same sensitivity, while the lightweight algorithm shows an improved specificity. This difference in specificity is due to a single additional false positive by the original algorithm.

Table 4.3: Specificity and Sensitivity results from survey

	Original	Lightweight
Sensitivity	0.870	0.870
Specificity	0.818	0.848

Table 4.4 shows results on an individual basis. In contrast to the majority voting results, the average specificity of the lightweight algorithm is lower than the original but has a significantly higher variance. Majority voting compensates for the variance and causes an increase in performance.

Some interesting trends arise from survey stratification. Both groups A (participants P1-P3) and B (participants P4-P7) achieved a higher individual sensitivity in the second survey they completed. Sensitivity is a more critical measure when it comes to seizure detection. It measures the proportions of annotated seizures correctly detected in the survey. In turn, the second survey each participant completes presents a decrease in their individual specificity.

Kappa values are calculated and plotted with confidence intervals to make comparisons in the agreement between annotators and the algorithms. There is no statistically significant disagreement between annotators and the original algorithm, and annotators and the lightweight as the p-value is calculated as 0.223.

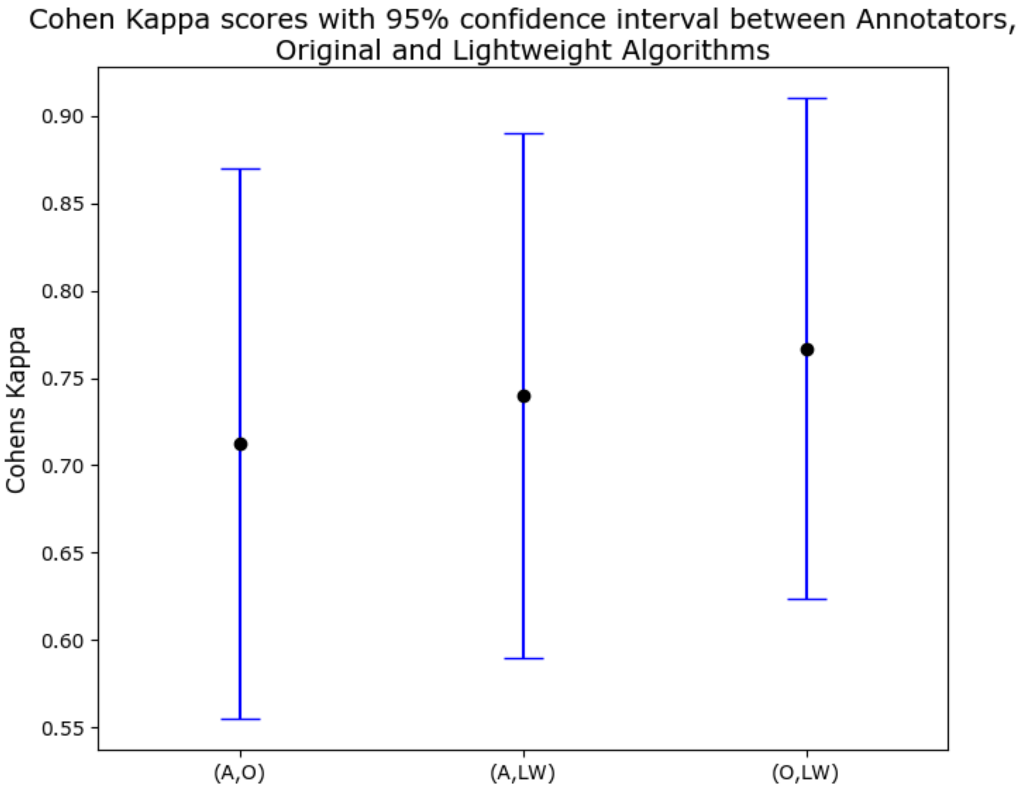


Figure 4.6: Cohen's Kappa with 95% confidence interval

Table 4.4: Individual participant sensitivity and specificity with variances from survey

	Original		Lightweight	
	Sensitivity	Specificity	Sensitivity	Specificity
P1	0.870	0.757	0.717	1.000
P2	0.935	0.666	0.760	0.970
P3	0.891	0.696	0.826	0.848
P4	0.81	0.576	0.870	0.758
P5	0.717	0.848	0.804	0.697
P6	0.760	0.970	0.891	0.454
P7	0.739	0.848	0.869	0.606
Average ± Variance	0.829 ± 0.008	0.766 ± 0.018	0.820 ± 0.004	0.762 ± 0.038

4.7.2 Error Analysis

Although performance is shown to be similar for both algorithms, the participants interpreted the output for 9 of the recordings differently (2 of which contained seizures).

The Cohen Kappa score measures the agreement level between algorithms as 0.767. Where a kappa score between 0.61 and 0.8 indicate substantial agreement.

To analyse the causes of disagreement, features are extracted from EEG segments the algorithms and disagreed on. The chosen features are the Binary Cross Entropy of the AI predictions, the number of annotators who disagreed and the certainty of decision for both algorithms calculated. The AUC or discriminating power of each feature is given in Table 4.5. The certainty is measured by the fraction of votes past the decision threshold. This metric's max value is 0.5 (representing total confidence), and the minimum is 0 (representing an even split of votes). The power of these features to discriminate between the classifications the algorithms agreed (70 patients) and disagreed on (9 patients) is measured using AUC. The leading cause of disagreement between the surveys is a high degree of uncertainty amongst participants (an AUC of 0.902 and 0.865 for the lightweight and original algorithms).

Table 4.5: Average ± CI95 for annotator disagreement, Binary Cross Entropy (BCE) of AI predictions and the certainty of decisions for the recordings the algorithms agreed and disagreed on. The final row shows the AUC score, a measure of each metric's discriminating power

	Disagreement	BCE	LW Certainty	Original Certainty
Agreed (n=70)	0.2 ± 0.104	0.260 ± 0.0411	0.414 ± 0.0319	0.410 ± 0.0329
Disagreed (n=9)	0.556 ± 0.405	0.449 ± 0.183	0.119 ± 0.109	0.198 ± 0.102
AUC	0.681	0 .735	0.902	0.865

4.8 Discussion

EEG monitoring is the gold standard for assessing a patient's brain health. It is non-invasive, and its use in neonatal encephalopathy is widespread (Pisani and Pavlidis, 2018). The necessary interpretation expertise is often unavailable in clinical settings (Boylan et al., 2010). Even where there are onsite expert clinicians, expertise is not available 24/7. These problems are exacerbated in the developing world (Haider and Bhutta, 2006). Without the relevant expertise, healthcare professionals use aEEG (simplified EEG) or clinical signs for seizure detection. These alternative interpretation methods contain limited information and lead to missed seizures (Hellström-Westas et al., 2006; Murray et al., 2008; Rennie et al., 2004; Zhang et al., 2011).

Even when sufficient expertise is available, timely treatment is critical to preventing long-term damage due to encephalopathy. (Pavel et al., 2022) studies the effects of varying treatments on neonatal seizures. Only 17 % of the babies who received medication had received it within 1 hour of a seizure occurring. It was shown that babies who received treatment within an hour had a lower seizure burden after treatment than those who received treatment outside this window.

Due to the widespread need for timely care, upskilling healthcare professionals in asphyxia treatment has been ranked as the second highest research priority for improving newborn health and birth outcomes by 2025 (Yoshida et al., 2016).

To overcome obstacles in detection, the original sonification algorithm was proposed. It enabled a human listener to classify EEG recordings as seizure/non-seizures with the same proficiency as expert doctors with virtually no training. It improved over previous sonification methodologies as it possessed a variable time compression factor, enabling high temporal compression without missing short seizures (Gomez-Quintana et al., 2022).

The technology needed to be made amenable for implementation on an edge device to make this process pervasive in EEG monitoring. An edge implementation allows the cotside analysis of sonified EEG. Using low-cost hardware in resource-constrained environments where the necessary expertise is scarce is particularly advantageous. A lightweight version of the algorithm was developed to these ends. The following subsections deeply analyse the disagreement between the algorithms and their causes, along with changes in execution time and the limitations of the presented work.

4.8.1 Execution Time

The execution time of the lightweight algorithm is 13.3x faster than that of the original. This performance increase is due to decreased data at each point in the algorithm and, in some cases altering functions to be more amenable for execution on an edge device. The average wait time for a medic to sonify 1 hour of EEG is now 2.23 seconds. Since 1 hour of EEG is compressed to 5 seconds of audio on average, the processing time is less than half of the review time. The wait time is now deemed acceptable to put into clinical practice.

Compared to the first implementation presented in the previous chapter, the lightweight algorithm is 183.4x faster. Figure 4.7 visually shows the improvements in execution time due to algorithmic changes discussed in the current and previous chapter.

4.8.2 Algorithm Performance

The algorithm's performance is measured in absolute terms using sensitivity and specificity. Both algorithms have a higher sensitivity than specificity. Sensitivity is more important in clinical practice as the risk of unnecessary treatment associated with false positives is less severe than the risk of death

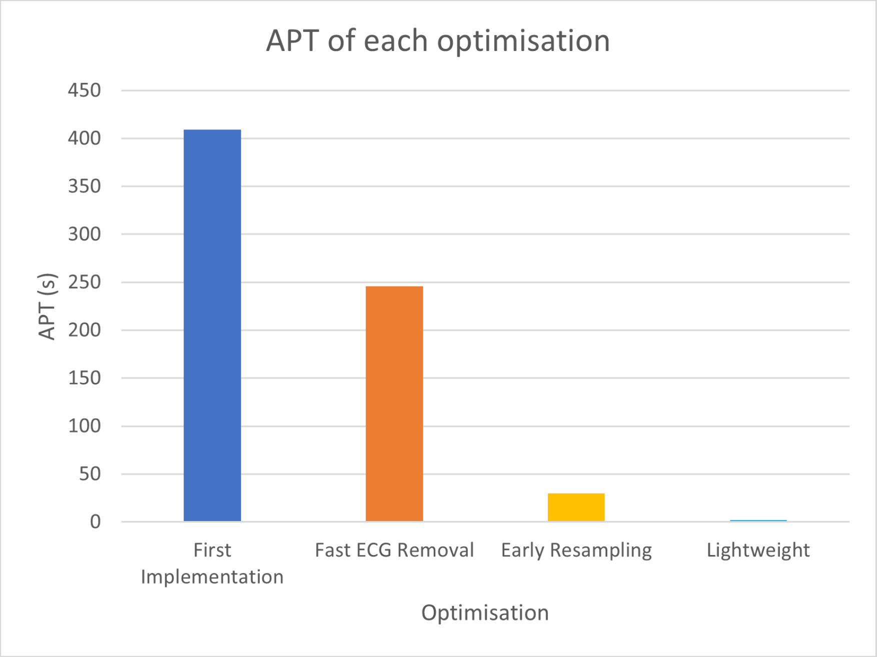


Figure 4.7: Average Processing Time shown for each Optimisation

associated with missed seizures. The sensitivity and specificity are high for both algorithms and are in line with what was previously reported.

The surveys were stratified to ensure a fair evaluation of the two algorithms. All participants achieved a higher sensitivity in the second survey. The increase in detected seizures shows participants become more tuned to the sounds of seizures over time. Although the increase in sensitivity is accompanied by a decrease in specificity, sensitivity is more important in the medical domain making this change desirable.

The algorithm's performance is also measured in terms of the agreement between the lightweight, original and annotators. All the Kappa's are greater than 0.6 and indicate substantial agreement. The highest agreement level was between the lightweight and original algorithms. This was expected and indicates no fundamental change in the sonification algorithm, as the core vocoder block was unchanged. The p-value comparing the distribution between annotators and the lightweight, and the annotators and the original algorithm shows no statistically significant difference in agreement between the two algorithms and the ground truth.

4.8.3 Error Analysis

Table 4.5 shows that disagreement between annotators contributes to differing classifications in the two surveys, with an AUC of 0.681. When the algorithms disagreed, often the annotators did too. This implies that some disagreement is due to the subjective nature of the ground truth. On average, at least one annotator would consider the output of both algorithms correct.

The algorithm's disagreement is also a function of the FCNN's performance. The AI's performance was measured using binary cross entropy. With an AUC of 0.735, when the AI performance is low there is an increase in the disagreement between the two algorithms.

The most significant factor in differing classifications is the certainty among survey participants, with AUCs of 0.902 and 0.865 for the lightweight and original algorithms, respectively. The relatively poor performance in the areas mentioned above likely contributed to the uncertainty. The low decision certainty for these nine samples indicates they are difficult to classify using the existing algorithms and that the lightweight algorithm improved on the specificity of the original by chance.

4.8.4 Survey Limitations

The major limitation of this survey is the small number of participants involved (n = 7). A majority vote is taken to reduce variance due to the small number of participants. Additionally, the end goal of the project is to be used by medical personnel but the 7 participants are from non-clinical backgrounds and may not fully understand the condition they are trying to diagnose. The previously conducted survey showed that clinical and non-clinical participants perform at different operating points in terms of sensitivity and specificity. Further work must be carried out to guarantee the performance of the lightweight algorithm on a clinical cohort.

The survey is conducted using the publicly available Helsinki dataset, consisting of 79 EEG recordings 1-2 hours long. The use of a public dataset is beneficial as it allows results to be easily reproduced. However, it comes with some limitations. The dataset is curated, and further research must be conducted to test the solution in a real clinical setting. Audio generation is also not tested for long EEG (>2 hours), which may be required in clinical practice.

A notable change in the lightweight algorithm is the absence of the ECG removal function. A comparative survey must also be completed on other datasets to ensure performance is not only maintained because there are low levels of ECG interference present in this dataset.

Survey participants' performance is expected to increase with practice. So although participants' initial performance was similar, it is uncertain whether this holds for the upper bound of performance after participants undergo extensive training.

4.8.5 Survey Availability

The surveys used for both implementations, along with the entire database of generated audio files, are available at the following links:

Original: https://sergigomezquintana.github.io/EEGsoundSurvey/

Lightweight: https://feargalos.github.io/

4.9 Conclusion

To further decrease the execution time of the algorithm, a lightweight version was proposed. Improvements were made by exploiting existing data compression in the vocoder and mixing the channels. The role of each function was examined, and some were optimised by using alternative methods or removed if deemed not performance critical. The changes in the algorithm resulted in a 13.3x decrease in execution time. The average processing time to sonify 1 hour of EEG is now 2.23 seconds.

A survey was carried out comparing the performance of the two algorithms. Results were similar, showing a sensitivity and specificity 0.870 and 0.848, respectively, for the lightweight algorithm and 0.870 and 0.818 for the original. The level of agreement between both algorithms and the ground truth annotations was measured using Cohen's Kappa. It was shown there is no statistically significant disagreement as the p-value between the two Kappas was 0.223 (>0.05). The Kappa score between the two algorithms is 0.767, indicating substantial agreement. The causes of disagreement between the two algorithms were investigated using AUC. Uncertainty was the most significant factor when comparing files the algorithms agreed and disagreed on. The AUC for uncertainty is 0.902 for the lightweight and 0.865 for the original. The high level of uncertainty among patients the algorithms disagreed on implies that the lightweight algorithm's improved specificity was obtained by chance.

The improvement in APT is significant and makes the algorithm more usable in a clinical setting. Making the algorithm amenable for edge deployment enables pervasive cheap, cotside EEG monitoring for neonates. Improvements in monitoring lead to an improved standard of care and the potential for decreased morbidity rates.

5

Low Precision CNN Inference

In the previous chapters, a neural network is used to achieve variable compression during sonification. AI inference is fundamental to the algorithm, and the network must be optimised to run on an edge device to decrease the memory and computational resources required. Large neural networks with many parameters are used to learn complex patterns and achieve high-performing models. The network's size presents a difficulty when deploying models on resource-constrained devices where memory is scarce and often only low-precision operations are available. The research area focused on deploying machine learning to the edge is known as Tiny ML. The idea is to decrease the number of bits used to represent the network (Quantisation) and remove unnecessary weights (Pruning) to allow faster low-memory inference. This chapter investigates the feasibility of operating the network at lower precision levels.

5.1 TensorFlow Ecosystem

The TensorFlow ecosystem is an open-source framework developed by Google for training, optimising and deploying models. TensorFlow can be used on machines with high computational power (servers, desktops). TensorFlow Lite compresses and optimises neural networks trained with TensorFlow for deployment on mobile devices (TensorFlow, 2022a). The optimised model can be deployed using TensorFlow Lite Micro on bare metal microcontrollers with C++ (David et al., 2021) (Figure 5.1).

This work optimised a model trained using TensorFlow in (Daly et al., 2021) using quantisation in TensorFlow lite.

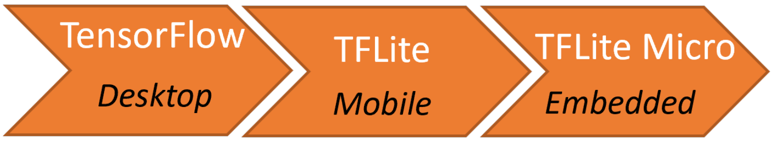


Figure 5.1: Tensorflow Ecosystem

5.2 Quantisation

Quantisation is a common optimisation technique applied to reduce the precision of a model's weights to reduce the model's size and computational requirements. It is the process of mapping a large range of values to a smaller one. Quantisation error is defined as the finite error in the value mapping between the original and smaller range. As well as for model optimisation, quantisation is widely used in signal processing and data compression. The quantisation process and the resultant error are shown in Figure 5.2 from the digital sampling of an analogue signal. The same principle applies to quantising neural network weights.

The neural network weights are stored by default as 32-bit floating point numbers when the network is trained. The network size and inference time can be improved by representing weights as 8 or 16-bit integers. The smallest possible datatype in C is an 8-bit integer. Moreover, TensorFlow Lite is the framework used for quantisation in this work, and it supports 8-bit integers as the maximum quantisation. Although some weight-sharing schemes have been proposed, where two weights are stored in the same integer address, to deploy models at even lower precisions, the minimum precision considered in this work is 8 bits.

Although it is advantageous to quantise the weights for the execution on the Raspberry Pi, to move onto smaller, more energy-efficient microcontrollers, it may become a necessity. The limitation occurs because some CPUs and Neural Processing units (NPU) cannot make computations on 32-bit representations. Although 32 bit NPUs exist they are not common, as these chips are often meant for ultra-low power computing where lower precision operations offer memory and speed advantages.

5.2.1 Post-Training Quantisation

Post-training quantisation directly converts a model's weights to a lower precision without altering the training process. It is convenient. However, the resultant quantisation error can be high.

Quantisation can be used after training to represent a model's weights using

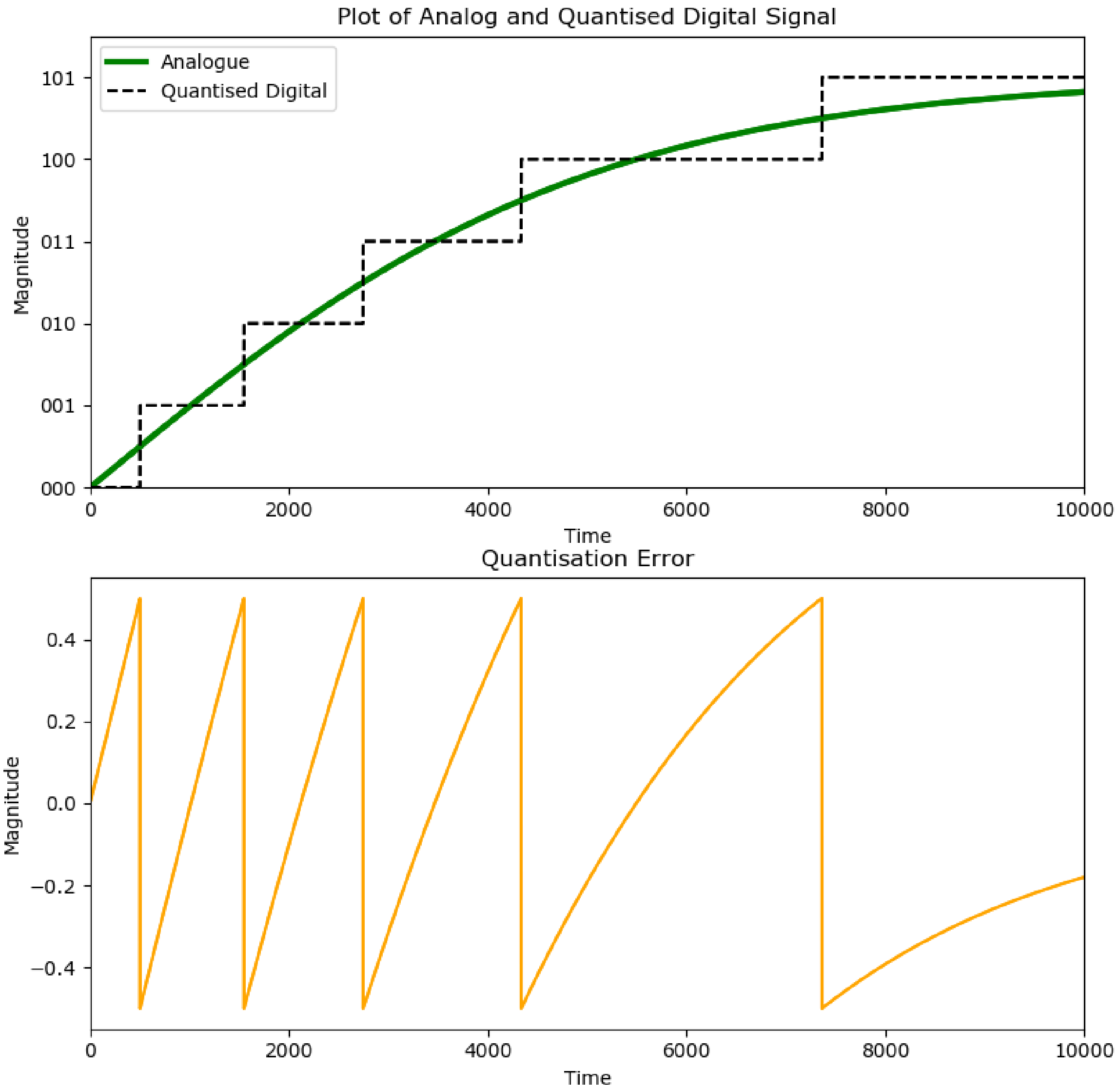


Figure 5.2: The continuous analogue signal $5Tanh(x)$, where $x \in [0, 10000]$, and the discrete quantised version. The y-axis is given in base-2 format. The bottom graph shows the associated quantisation error

8-bit integer two's complement representation in the range [-128, 127]. The zero point is set to zero or any number between [-128, 127]. The scale is calculated using a representative dataset containing a few hundred randomly selected unlabelled input data points to calibrate the quantisation (Jacob et al., 2018). Figure 5.3 shows an example mapping.

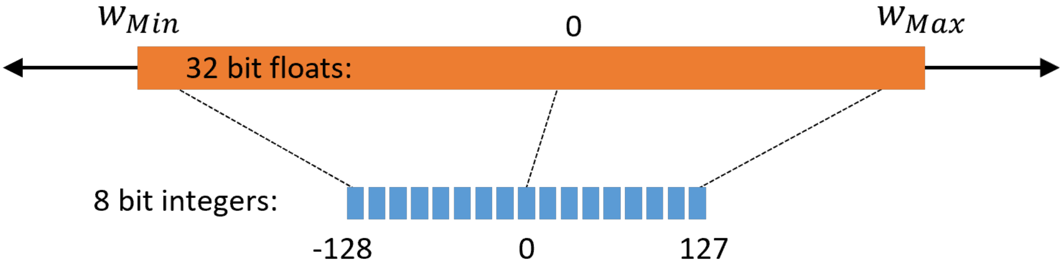


Figure 5.3: Float 32 to Int 8 Quantisation Mapping

The relationship can be expressed mathematically as:

$$\text{float value } = (\text{ int value } - \text{ zero point }) \times \text{ scale} \tag{5.1}$$

5.2.1.1 Full Integer

Full integer quantisation forces all operations and weights to either 16 or 8-bit weights. For example, in an 8-bit quantisation scheme, weights and activations are quantised to 8-bit integers, and only integer operations are used. For some embedded platforms, it is necessary to use this form of quantisation (TensorFlow, 2022c).

5.2.1.2 8-bit weights with 16-bit Activations

This is a special case of mixed precision full integer quantisation. When the quantisation error is due to the activations alone, this quantisation type can significantly improve a quantised model's accuracy without a significant increase in model size. It is beneficial when the activation functions are sensitive to quantisation. It is an experimental framework created by Tensorflow, and its kernel operations are not yet optimised. As such, inference runs noticeably slower than the full 8-bit integer quantisation on a CPU (TensorFlow, 2022b).

5.2.1.3 Dynamic-Range Quantisation

This type of quantisation dynamically de-quantises activations during inference and yields minimal performance loss. Because the activations are floating point, they can not be used in integer-only hardware like an NPU.

5.2.2 Quantisation Aware Training

Steps can be taken to reduce quantisation errors as part of the training process (Jacob et al., 2018). In quantise aware training, low precision operations are simulated during forward propagation, while backward propagation is unchanged.

5.3 Preproccessing

The raw EEG signal is filtered using a 0.5 to 12.8 Hz band pass filter. Seizures are not expected to occur outside this range. Immediately after filtering, the signal is downsampled from 256 Hz to 32 Hz. The AI's preprocessing stage is identical to that of the sonification algorithm. Figure 5.4 illustrates the preprocessing signal flow as a block diagram.

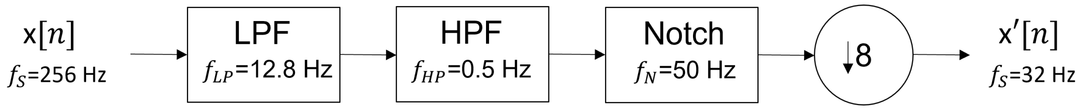


Figure 5.4: Block diagram of preproccessing stages

5.4 Fully Convolutional Neural Network (FCNN)

The AI probabilities used for the survey were generated using the FCNN developed in (O'Shea et al., 2020). The current state of the art has since been improved in (Daly et al., 2021), where an AUC of 96.4 % was achieved on the Helsinki database (Stevenson et al., 2019), which constitutes a 0.8 % improvement over the previous model. The architecture of the two models is fundamentally the same. However, residual connections are added to obtain a deeper model. The state-of-the-art model is optimised for deployment in this work. The changes in architecture from the two networks are studied.

Generally speaking, FCNN is advantageous over a traditional CNN as it contains fewer parameters than a network with fully connected layers, making training and inference easier. Residual connections mitigate the effect of a vanishing gradient problem to allow for a deeper network (He et al., 2015). Additionally, the input can be of any size because there are no fully connected layers. This makes for easier deployment as the number of channels varies greatly.

5.4.1 Residual Connections

Residual connections allow the network to learn a mapping between the block's input and output. They are advantageous because they increase the maximum network depth before performance is limited by the vanishing gradient problem. For FCNNs, an increase in network depth increases the receptive field and the effective input size, which corresponds to an increase in performance.

5.4.2 Feature Extraction Blocks

Feature extraction blocks consist of 3 convolutional layers with a Rectified Linear Unit (ReLU) activation function and average pooling.

The convolutional layers extract meaningful patterns from the EEG signal while maintaining the spatial relationship in the waveform. A 2D convolution is applied with a 1D kernel with all N-channels of the EEG. The 1D kernel processes each channel individually. A ReLU activation function is applied, which enables the neural network to learn non-linear patterns. Average pooling is used to reduce the dimensionality of the outputted feature maps. These blocks, shown visually in Figure 5.5, can be stacked to maximise the network's performance.

5.4.3 Classification Block

The other block used in the network is the classification block (Figure 5.6). The network is fully convolutional and, as such, has no fully connected layer to make a classification. Instead, the patterns extracted by the convolutions are classified using a global pooling layer. Average pooling reduces dimensionality across time, and max pooling reduces dimensionality across the channels to find the channel with the most 'seizureness'. All channels are processed individually, and the channel most closely resembling a seizure is

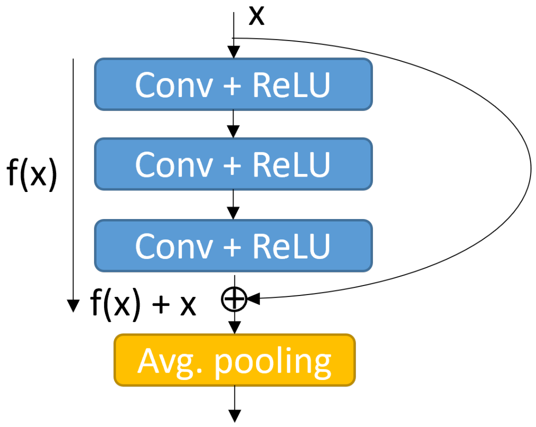


Figure 5.5: Feature Extraction Block of CNN

then taken to make a classification. Loss is calculated against the max probability of seizure, and weak labels can be used to train the network. Weak labels, in this case, are annotations that specify when a seizure occurs but not in what channel. There is much more data available with weak labels than strong ones. The network's ability to train on more data contributed to its state-of-the-art performance.

After global pooling, the softmax function converts the output into a probability. Since average and max pooling are used for classification, the network can operate on any input length with any number of channels. However, increasing the window length affects the receptive field, which in turn affects performance.

This block is placed at the end of the network and only occurs once.

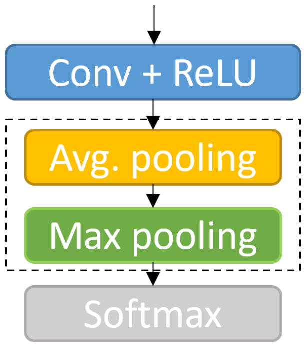


Figure 5.6: Final Classification Block of FCNN

5.4.4 Network Comparison

The network optimised in this chapter has three extra convolutional layers compared with the network used to generate probabilities in the previous chapters. The increase in depth corresponds to an increase in performance. The difference is shown visually in Figure 5.7.

5.5 Training Procedure

The network is finetuned using quantise aware training from the weights of the best previously performing model to reduce quantisation error. The original model was trained on a proprietary dataset not available to the author of this work. The network has to be retrained on the Helsinki dataset to obtain a baseline to get a fair measure of performance loss. Details on finetuning follow, but the sole goal of this retraining is to decrease quantisation error, not absolute performance. So no innovations are presented in terms of the network architecture.

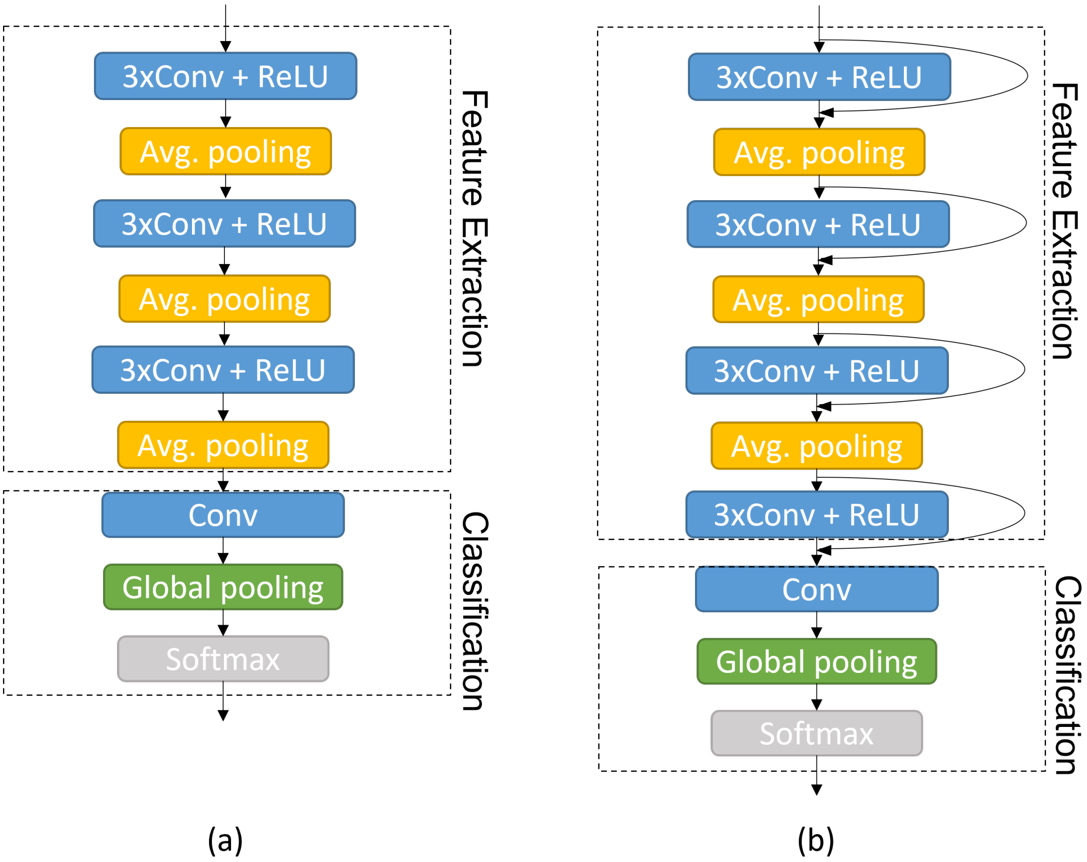


Figure 5.7: Comparison of both FCNN architectures

5.5.1 K Fold Cross Validation

Patient-independent K-fold cross-validation is used to measure model performance. K is selected as 5 for these experiments. The patients are divided into five groups (each containing approximately 20 % of the data). One fold is used for testing, while the others are used for training. The test fold is alternated until all folds are used for the test set. No data from any baby is used simultaneously in training and testing. This simulates the clinical scenario where patient generalisation is key and evaluates the model's patient-independent performance (Saeb et al., 2016).

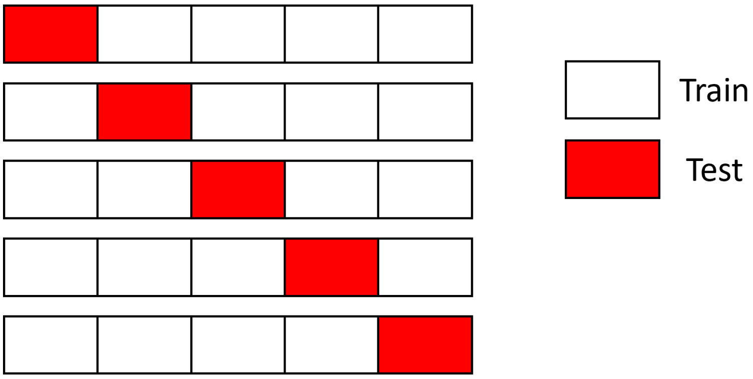


Figure 5.8: Illustration of alternating test fold

5.5.2 Optimiser

RAdam (Rectified Adam) is used as an optimiser. This is similar to Adam but accounts for variance in the adaptive learning rates early in the learning process (Liu et al., 2019).

5.5.3 Loss

Binary cross entropy, or log loss (as defined in Equation (4.4)), is used to measure loss. The function is widely used for classification problems. The parameters are then updated through backpropagation to minimise the loss calculated from the maximum predicted probability of seizure in all channels. The network works on weak labels, where annotations are not channel specific, so the loss is calculated based on the max probability of seizure in any channel.

5.5.4 Data Augmentation

To reduce generalisation error, data augmentation is used. Vertical, horizontal, and magnitude scaling are randomly applied. Cutmix and mixup are also applied. Cutmix is when part of the windowed EEG is swapped out with another section of EEG which can be taken for any patient in the dataset or either class. If segments from different classes are combined, the labels are mixed in the ratio of the segements. Mixup is similar except the two different segments are linearly combined to form a new segment. The labels are mixed in the same ratio.

5.6 AUC

The metric used to measure the performance of the neural networks is the Area Under the Curve (AUC) calculated from the Receiver Operating Characteristic (ROC). The ROC plots the sensitivity against specificity at different decision thresholds for the model's probabilistic output (Bradley, 1997). AUC is advantageous over accuracy for imbalanced datasets where accuracy may give misleading results. For example, a model performing well on the majority class will have high accuracy, even if it performs poorly on the minority class (Huang and Ling, 2005). Seizures are rare events, and a model producing many false negatives (missing seizures) will still possess a high accuracy.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5.2}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \tag{5.3}$$

The perfect model would score an AUC = 1, giving a right angle in the ROC. The worst obtainable score is an AUC = 0.5, meaning the classifier has no predictive power and is the equivalent of tossing a coin to detect seizures. The worst model would yield a ROC consisting of a straight line, from the top left corner to the bottom right corner of the graph. An example ROC curve is shown in Figure 5.9.

5.7 Results

The results reported in (Daly et al., 2021) are generated by taking an ensemble of three sets of weights from different training rounds. The ensemble achieves 0.11 % better than the best-performing single model. For implementation, the best-performing model is taken as the ensemble's marginal increase in performance is not justified by the three-fold increase in inference time. The new baseline performance is an AUC of 96.32 %. The results

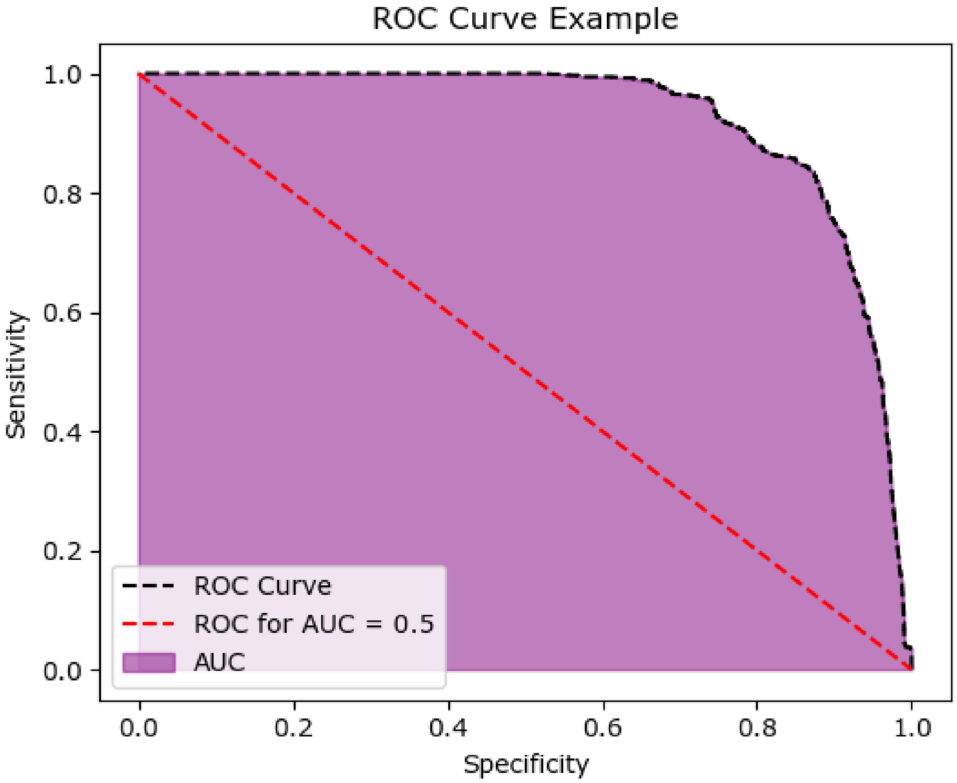


Figure 5.9: Example Receiver Operating Characteristic curve and the corresponding AUC. The worst case scenario, AUC = 0.5, is also shown

of each quantisation level are presented in Table 5.1. The quantisation error for the post-training quantisation results is reported as the fall in AUC score according to this baseline, while the performance loss of the quantise aware method is measured relative to the Helsinki Baseline.

Table 5.1: Results for varying levels of precision and quantisation methods

	AUC (%)	**Error (%)**
Baseline	96.32	-
Dynamic	96.29	0.03
Int 8x16	96.38	-0.06
PT Full Integer	63.88	32.44
Helsinki Baseline	96.61	-
QA Full Integer	88.84	7.77

5.8 Discussion

Some models are more suited to quantisation than others, and this model seems to quantise well up to a point.

The model performance after the dynamic quantisation is almost identical to the baseline. It uses floating-point fallback, so although it enables some model compression and faster inference, it cannot run on integer-only hardware. This limits its potential for implementation

Full integer quantisation at 8-bit precision significantly reduced AUC from 96.32 % to 63.88 %. The quantisation error of 32.44 % is unacceptable. Int 8 weights, with int 16 activations, were investigated as an alternative. This type of quantisation led to no loss in performance. The biggest source of error was found to be from the Batch Normalisation Layers. TF Lite treats

Batch Normalisation as an activation function, and quantisation error is significantly reduced at the higher precision level. Although the kernels are not currently optimised for general CPUs, specialised hardware like the Arm Ethos U55 chip can run this model efficiently (Arm, 2022).

Despite the potential for quantising a model with 16-bit activations, microcontrollers with NPUs are more widely available for full 8-bit quantised networks. To reduce quantisation error, quantise aware training was used on the Helsinki dataset. The performance of the 8-bit quantised model was 88.84 %, a fall in AUC of 7.77 % compared to the Helsinki baseline. Although the quantisation error was decreased by a factor of 4.2, the fall in AUC is still unacceptable.

The algorithm's performance has been shown to rely on the CNNs performance (Gomez-Quintana et al., 2022). Therefore, a fall in AUC will directly affect the sonification algorithm's classification performance. The network should be quantised to 16 bits to ensure performance is unaffected while minimising memory and inference time.

5.9 Conclusions

The Fully Convolutional Neural Network enables variable compression during sonification and is optimised for size. As inference is a key part of the sonification algorithm, a range of quantisation schemes are investigated to achieve a high level of optimisation. The network can be quantised to 16-bit activations with 8-bit weights allowing full integer operations with no loss in performance. Quantising the model opens up the potential for a future ultra-low power embedded systems implementation and integration with highly-specialised integer-only hardware.

6

Conclusions and Future Directions

The brain is a person's essential control centre and is vulnerable in the early stages of life. Health complications in the first few years of life can cause permanent disability. Asphyxia after birth usually presents in the form of seizures. Detecting these seizures is a challenging clinical task. Since only ten per cent of neonatal seizures show clinical signs, EEG must be used to monitor brain health. Visual EEG analysis is the current gold standard, but it is challenging to interpret and requires a highly skilled neurologist to make diagnoses. Even when expertise is available, it is not available round the clock, nor is it trivial to treat a patient promptly due to the time taken to make a diagnosis. These problems are made worse in disadvantaged communities. Solutions for automating EEG interpretation using AI alone have been proposed. However, these models must choose between explainability and performance (deep learning vs traditional machine learning), but all options push the human out of the decision-making loop. Sonification is advantageous as it keeps a human involved in the decision process. This thesis focused on deploying an AI-assisted sonification algorithm on an edge device. The algorithm combines the absolute performance of a CNN while adding interpretability from the sonification element. Pushing this algorithm onto an edge device makes it available cheaply to underfunded communities where it is needed most.

The algorithm was first implemented on a Raspberry Pi 3B+ with Python. The average time to sonify 1 hour of EEG was 409.1 seconds. It was seen that the 'ECG Removal' function was a bottleneck to performance. It was optimised by changing slow iterative calculations to fast parallel matrix operations. The average processing time to sonify 1 hour of EEG then decreased to 245.6 seconds. Finally, the signal was downsampled early in the sonification process because the frequency content of seizures is generally less than 13 Hz, so processing at the original sampling rate of 256 Hz is redundant. This decreased the average processing time to 29.6 seconds. This represents a 13.8x reduction in execution time from the first implementation of the algorithm. The spectral output was compared, and it was shown after optimisations that the algorithm produced nearly identical outputs.

From these results, it was concluded that this version of the algorithm was not amenable to an edge implementation. On average, processing time was 6 times longer than review time, making the algorithm unusable in the real world. To improve execution time, more drastic changes needed to be made to the algorithm. Data compression existed in the algorithm already but was not yet exploited to improve processing time. It was utilised by reordering the blocks of the algorithm. The mixer and phase vocoder reduce data using channel reduction and temporal compression, reducing the number of EEG samples to be processed by each proceeding stage. The execution time of each function was analysed, and their contribution to the algorithm was qualitatively assessed in comparison. In cases where the execution time outweighs the function's value, a suitable alternative was implemented instead. And in some cases, the functions were removed. The fundamental structure of the algorithm was changed, and the performance had to be evaluated by means of a survey. The survey showed no statistical significance in the performance of the two algorithms. The algorithms produced different classifications for 9 of the 79 audio files in the survey. The underlying causes of these misclassifications were examined by looking at different features AUC between the files the algorithms agreed and disagreed on. Although disagreement between annotators (68.1 %) and AI performance (73.5 %) played a role, the biggest contributor was found to be uncertainty among survey participants (90.2 % for the lightweight and 86.5 % for the original). This suggests underlying performance was unaffected, and any disagreement between algorithms was due to uncertainty and chance. This further confirms the performance of the lightweight algorithm.

The changes resulted in a 13.3x reduction in execution time. Compared to the first implementation of the algorithm, the lightweight version is 183.4x faster. The average wait time is now acceptably low for adoption in a clinical setting.

The CNN used in the sonification algorithm is optimised for deployment on an edge device. The model's weights are quantised to 8 bits while the activation is quantised to 16-bit integers without any loss of performance. The model was retrained using quantise aware training in an attempt to fully quantise the model to 8 bits. The loss in performance was still too large. The mixed precision 8 and 16 bit model is still able to run on integer-only hardware and opens up future possibilities to put the network on an NPU.

Although work has been carried out moving this algorithm to the edge, an implementation in C/C++ would be faster and more power efficient than in Python. Specialised hardware, such as a dedicated DSP processor and a Neural Processing Unit, could be used to decrease execution time further.

References

Aarabi, A., Wallois, F., and Grebe, R. (2006). Automated neonatal seizure detection: a multistage classification system through feature selection based on relevance and redundancy analysis. *Clinical Neurophysiology*, 117(2):328–340. 16

Ahmed, R., Temko, A., Marnane, W., Lightbody, G., and Boylan, G. (2016). Grading hypoxic–ischemic encephalopathy severity in neonatal eeg using gmm supervectors and the support vector machine. *Clinical Neurophysiology*, 127(1):297–309. 17

Ansari, A. H., Cherian, P., Dereymaeker, A., Matic, V., Jansen, K., De Wispelaere, L., Dielman, C., Vervisch, J., Swarte, R., Govaert, P., et al. (2016). Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor. *Clinical Neurophysiology*, 127(9):3014–3024. 16

Arm (2022). Arm ethos https://www.arm.com/products/. 89

Aurlien, H., Gjerde, I., Aarseth, J., Eldøen, G., Karlsen, B., Skeidsvoll, H., and Gilhus, N. (2004). Eeg background activity described by a large computerized database. *Clinical Neurophysiology*, 115(3):665–673. 10

Baier, G., Hermann, T., and Stephani, U. (2007). Event-based sonification of eeg rhythms in real time. *Clinical Neurophysiology*, 118(6):1377–1386. 18

Barnett, M. W. and Larkman, P. M. (2007). The action potential. *Practical neurology*, 7(3):192–197. 8, 10

Benjamens, S., Dhunnoo, P., and Meskó, B. (2020). The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ digital medicine*, 3(1):118. 15

Boylan, G., Burgoyne, L., Moore, C., O'Flaherty, B., and Rennie, J. (2010). An international survey of eeg use in the neonatal intensive care unit. *Acta paediatrica*, 99(8):1150–1155. 3, 64

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159. 85

Brogger, J., Eichele, T., Aanestad, E., Olberg, H., Hjelland, I., and Aurlien, H. (2018). Visual eeg reviewing times with score eeg. *Clinical neurophysiology practice*, 3:59–64. 2

Clancy, R. R. and Legido, A. (1987). The exact ictal and interictal duration of electroencephalographic neonatal seizures. *Epilepsia*, 28(5):537–541. 15, 18

Daly, A., O'Shea, A., Lightbody, G., and Temko, A. (2021). Towards deeper neural networks for neonatal seizure detection. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 920–923. 3, 16, 74, 79, 86

David, R., Duke, J., Jain, A., Janapa Reddi, V., Jeffries, N., Li, J., Kreeger, N., Nappier, I., Natraj, M., Wang, T., et al. (2021). Tensorflow lite micro: Embedded machine learning for tinyml systems. *Proceedings of Machine Learning and Systems*, 3:800–811. 74

Delanty, N., Vaughan, C. J., and French, J. A. (1998). Medical causes of seizures. *The Lancet*, 352(9125):383–390. 13

Di Flumeri, G., Aricò, P., Borghini, G., Sciaraffa, N., Di Florio, A., and Babiloni, F. (2019). The dry revolution: Evaluation of three different eeg dry electrode types in terms of signal spectral features, mental states classification and usability. *Sensors*, 19(6):1365. 10

Douglas-Escobar, M. and Weiss, M. D. (2015). Hypoxic-ischemic encephalopathy: a review for the clinician. *JAMA pediatrics*, 169(4):397–403. 2

Eriksson, M. and ZetterstrÖM, R. (1979). Neonatal convulsions incidence and causes in the stockholm area. *Acta Pædiatrica*, 68(6):807–811. 13

Faul, S., Boylan, G., Connolly, S., Marnane, W., and Lightbody, G. (2005). Chaos theory analysis of the newborn eeg-is it worth the wait? In *IEEE International Workshop on Intelligent Signal Processing, 2005.*, pages 381–386. IEEE. 15

Fenichel, G. M. (1983). Hypoxic-ischemic encephalopathy in the newborn. *Archives of neurology*, 40(5):261–266. 9

Flanagan, J. L. and Golden, R. M. (1966). Phase vocoder. *Bell System Technical Journal*, 45(9):1493–1509. 33

Fukushima, K. and Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and Cooperation in Neural Nets: Proceedings of the US-Japan Joint Seminar held at Kyoto, Japan February 15–19, 1982*, pages 267–285. Springer. 16

Giannoulis, D., Massberg, M., and Reiss, J. D. (2012). Digital dynamic range compressor design—a tutorial and analysis. *Journal of the Audio Engineering Society*, 60(6):399–408. 29

Gomez, S., O'Sullivan, M., Popovici, E., Mathieson, S., Boylan, G., and Temko, A. (2018). On sound-based interpretation of neonatal eeg. In *2018 29th Irish Signals and Systems Conference (ISSC)*, pages 1–6. IEEE. 18, 35

Gómez-Quintana, S., Cowhig, G., Borzacchi, M., O'Shea, A., Temko, A., and Popovici, E. (2021). An eeg analysis framework through ai and sonification on low power iot edge devices. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 277–280. IEEE. 4, 26

Gomez-Quintana, S., O'Shea, A., Factor, A., Popovici, E., and Temko, A. (2022). A method for ai assisted human interpretation of neonatal eeg. *Scientific Reports*, 12(1):1–13. 4, 10, 18, 19, 21, 33, 53, 65, 89

Gotman, J., Flanagan, D., Zhang, J., and Rosenblatt, B. (1997). Automatic seizure detection in the newborn: methods and initial evaluation. *Electroencephalography and clinical neurophysiology*, 103(3):356–362. 15

Haider, B. A. and Bhutta, Z. A. (2006). Birth asphyxia in developing countries: current status and public health implications. *Current problems in pediatric and adolescent health care*, 5(36):178–188. 64

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition https://arxiv.org/abs/1512.03385. 79

Hellström-Westas, L., Rosén, I., De Vries, L., and Greisen, G. (2006). Amplitude-integrated eeg classification and interpretation in preterm and term infants. *NeoReviews*, 7(2):76–87. 21, 64

Herculano-Houzel, S. (2009). The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in human neuroscience*, page 31. 8

Hermann, T., Meinicke, P., Bekel, H., Ritter, H., Müller, H. M., and Weiss, S. (2002). Sonifications for eeg data analysis. Georgia Institute of Technology. 18

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312. 3

Huang, J. and Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310. 86

Husain, A. M. (2005). Review of neonatal eeg. *American journal of electroneurodiagnostic technology*, 45(1):12–35. 11

Huttunen, T., Seppälä, E. T., Kirkeby, O., Kärkkäinen, A., and Kärkkäinen, L. (2007). Simulation of the transfer function for a head-and-torso model over the entire audible frequency range. *Journal of Computational Acoustics*, 15(04):429–448. 18

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713. 77, 78

Jasper, H. H. (1958). The ten-twenty electrode system of the international federation. *Electroencephalogr. Clin. Neurophysiol.*, 10:370–375. 9

Johnson, M. H. (2001). Functional brain development in humans. *Nature Reviews Neuroscience*, 2(7):475–483. 8

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., Hudspeth, A. J., Mack, S., et al. (2000). *Principles of neural science*, volume 4. McGraw-hill New York. 8

Khamis, H., Mohamed, A., Simpson, S., and McEwan, A. (2012). Detection of temporal lobe seizures and identification of lateralisation from audified eeg. *Clinical Neurophysiology*, 123(9):1714–1720. 18

Kitayama, M., Otsubo, H., Parvez, S., Lodha, A., Ying, E., Parvez, B., Ishii, R., Mizuno-Matsumoto, Y., Zoroofi, R. A., and Snead III, O. C. (2003). Wavelet analysis for neonatal electroencephalographic seizures. *Pediatric neurology*, 29(4):326–333. 11, 14, 18, 26

Kundu, S. (2021). Ai in medicine must be explainable. *Nature medicine*, 27(8):1328–1328. 3

Lanska, M. J., Lanska, D. J., Baumann, R. J., and Kryscio, R. J. (1995). A population-based study of neonatal seizures in fayette county, kentucky. *Neurology*, 45(4):724–732. 13

Lawn, J. E., Manandhar, A., Haws, R. A., and Darmstadt, G. L. (2007). Reducing one million child deaths from birth asphyxia–a survey of health systems gaps and priorities. *Health Research Policy and Systems*, 5(1):1–10. 1

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18. 17

Liu, A., Hahn, J., Heldt, G., and Coen, R. (1992). Detection of neonatal seizures through computerized eeg analysis. *Electroencephalography and clinical neurophysiology*, 82(1):30–37. 15

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J. (2019). On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*. 84

Lloyd, R., Goulding, R., Filan, P., and Boylan, G. (2015). Overcoming the practical challenges of electroencephalography for very preterm infants in the neonatal intensive care unit. *Acta paediatrica*, 104(2):152–157. 10

Loui, P., Koplin-Green, M., Frick, M., and Massone, M. (2014). Rapidly learned identification of epileptic seizures from sonified eeg. *Frontiers in human neuroscience*, 8:820. 19

McBride, M. C., Laroia, N., and Guillet, R. (2000). Electrographic seizures in neonates correlate with poor neurodevelopmental outcome. *Neurology*, 55(4):506–514. 13

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282. 56

Middlebrooks, J. C. (2015). Sound localization. *Handbook of clinical neurology*, 129:99–116. 35

Miller, S. P. and Ferriero, D. M. (2009). From selective vulnerability to connectivity: insights from newborn brain imaging. *Trends in neurosciences*, 32(9):496–505. 9

Murray, D. M., Boylan, G. B., Ali, I., Ryan, C. A., Murphy, B. P., and Connolly, S. (2008). Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 93(3):F187–F191. 2, 13, 65

Nagarajan, L., Palumbo, L., and Ghosh, S. (2010). Neurodevelopmental outcomes in neonates with seizures: a numerical score of background encephalography to help prognosticate. *Journal of child neurology*, 25(8):961–968. 13

Olivan, J., Kemp, B., and Roessen, M. (2004). Easy listening to sleep recordings: tools and examples. *Sleep medicine*, 5(6):601–603. 18

Oza, S., Lawn, J. E., Hogan, D. R., Mathers, C., and Cousens, S. N. (2014). Neonatal cause-of-death estimates for the early and late neonatal periods for 194 countries: 2000–2013. *Bulletin of the World Health Organization*, 93:19–28. 1

O'Shea, A., Lightbody, G., Boylan, G., and Temko, A. (2020). Neonatal seizure detection from raw multi-channel eeg using a fully convolutional architecture. *Neural Networks*, 123:12–25. 3, 16, 24, 25, 26, 79

O'Sullivan, M., Temko, A., Bocchino, A., O'Mahony, C., Boylan, G., and Popovici, E. (2019). Analysis of a low-cost eeg monitoring system and dry electrodes toward clinical use in the neonatal icu. *Sensors*, 19(11):2637. 10

Parvizi, J., Gururangan, K., Razavi, B., and Chafe, C. (2018). Detecting silent seizures by their sound. *Epilepsia*, 59(4):877–884. 19

Patrizi, S., Holmes, G. L., Orzalesi, M., and Allemand, F. (2003). Neonatal seizures: characteristics of eeg ictal activity in preterm and fullterm infants. *Brain and Development*, 25(6):427–437. 15

Pavel, A. M., Rennie, J. M., de Vries, L. S., Blennow, M., Foran, A., Shah, D. K., Pressler, R. M., Kapellou, O., Dempsey, E. M., Mathieson, S. R., et al. (2022). Neonatal seizure management: is the timing of treatment critical? *The Journal of Pediatrics*, 243:61–68. 2, 3, 65

Pisani, F. and Pavlidis, E. (2018). The role of electroencephalogram in neonatal seizure detection. *Expert review of Neurotherapeutics*, 18(2):95–100. 64

Rakshasbhuvankar, A., Paul, S., Nagarajan, L., Ghosh, S., and Rao, S. (2015). Amplitude-integrated eeg for detection of neonatal seizures: a systematic review. *Seizure*, 33:90–98. 13

Ramantani, G., Schmitt, B., Plecko, B., Pressler, R. M., Wohlrab, G., Klebermass-Schrehof, K., Hagmann, C., Pisani, F., and Boylan, G. B. (2019). Neonatal seizures—are we there yet? *Neuropediatrics*, 50(05):280–293. 13

Rankine, L., Stevenson, N., Mesbah, M., and Boashash, B. (2006). A nonstationary model of newborn eeg. *IEEE Transactions on biomedical engineering*, 54(1):19–28. 14, 26

RaspberryPi (2022). Datasheet for raspberry pi 3b+ https://static.raspberrypi.org/files/product-briefs/raspberry-pi-model-bplus-product-brief.pdf. 37

Rennie, J., Chorley, G., Boylan, G., Pressler, R., Nguyen, Y., and Hooper, R. (2004). Non-expert use of the cerebral function monitor for neonatal seizure detection. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 89(1):F37–F40. 13, 65

Roessgen, M., Zoubir, A. M., and Boashash, B. (1998). Seizure detection of newborn eeg using a model-based approach. *IEEE Transactions on Biomedical Engineering*, 45(6):673–685. 15

Ronen, G. M., Penney, S., and Andrews, W. (1999). The epidemiology of clinical neonatal seizures in newfoundland: a population-based study. *The Journal of pediatrics*, 134(1):71–75. 13

Rose, A. L. and Lombroso, C. T. (1970). Neonatal seizure states: A study of clinical, pathological, and electroencephalographic features in 137 full-term babies with a long-term follow-up. *Pediatrics*, 45(3):404–425. 18

Ruby, U. and Yendapalli, V. (2020). Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng*, 9(10). 56

Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., and Kording, K. P. (2016). Voodoo machine learning for clinical predictions. *Biorxiv*, page 059774. 84

Sanei, S. and Chambers, J. A. (2013). *EEG signal processing.* John Wiley & Sons. 8

Scharfman, H. E. (2007). The neurobiology of epilepsy. *Current neurology and neuroscience reports*, 7(4):348–354. 14, 15

Scher, M. S., Painter, M. J., Bergman, I., Barmada, M. A., and Brunberg, J. (1989). Eeg diagnoses of neonatal seizures: clinical correlations and outcome. *Pediatric Neurology*, 5(1):17–24. 13

Shellhaas, R. A., Chang, T., Tsuchida, T., Scher, M. S., Riviello, J. J., Abend, N. S., Nguyen, S., Wusthoff, C. J., and Clancy, R. R. (2011). The american clinical neurophysiology society's guideline on continuous electroencephalography monitoring in neonates. *Journal of clinical neurophysiology*, 28(6):611–617. 9

Sipser, M. (1996). Introduction to the theory of computation. *ACM Sigact News*, 27(1):27–29. 23

Stevenson, N. J., Tapani, K., Lauronen, L., and Vanhatalo, S. (2019). A dataset of neonatal eeg recordings with seizure annotations. *Scientific Data*, 6. 13, 16, 25, 52, 79

Stiles, J. and Jernigan, T. L. (2010). The basics of brain development. *Neuropsychology review*, 20(4):327–348. 8

Tapani, K. T., Vanhatalo, S., and Stevenson, N. J. (2019). Time-varying eeg correlations improve automated neonatal seizure detection. *International journal of neural systems*, 29(04):1850030. 16

Temko, A., Marnane, W., Boylan, G., O'Toole, J. M., and Lightbody, G. (2014a). Neonatal eeg audification for seizure detection. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4451–4454. IEEE. 18

Temko, A., Marnane, W., Boylan, G., O'Toole, J. M., and Lightbody, G. (2014b). Neonatal eeg audification for seizure detection. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4451–4454. 32

Temko, A., Thomas, E., Marnane, W., Lightbody, G., and Boylan, G. (2011). Eeg-based neonatal seizure detection with support vector machines. *Clinical Neurophysiology*, 122(3):464–473. 3, 16

TensorFlow (2022a). Tenssorflow lite https://www.tensorflow.org/lite/guide. 74

TensorFlow (2022b). Tflite dynamic range www.tensorflow.org/lite/performance/post_training_quant. 78

TensorFlow (2022c). Tflite integer https://www.tensorflow.org/lite/performance /post_training_integer_quant. 77

Thomas, E., Temko, A., Lightbody, G., Marnane, W., and Boylan, G. (2010). Gaussian mixture models for classification of neonatal seizures using eeg. *Physiological measurement*, 31(7):1047. 16

Tsuchida, T. N., Wusthoff, C. J., Shellhaas, R. A., Abend, N. S., Hahn, C. D., Sullivan, J. E., Nguyen, S., Weinstein, S., Scher, M. S., Riviello, J. J., et al. (2013). American clinical neurophysiology society standardized eeg terminology and categorization for the description of continuous eeg monitoring in neonates: report of the american clinical neurophysiology society critical care monitoring committee. *Journal of Clinical Neurophysiology*, 30(2):161–173. 11

Uria-Avellanal, C., Marlow, N., and Rennie, J. M. (2013). Outcome following neonatal seizures. In *Seminars in Fetal and Neonatal Medicine*, volume 18, pages 224–232. Elsevier. 13

Väljamäe, A., Steffert, T., Holland, S., Marimon, X., Benitez, R., Mealla, S., Oliveira, A., and Jordà, S. (2013). A review of real-time eeg sonification research. 4

White, D. M. and Van Cott, C. A. (2010). Eeg artifacts in the intensive care unit setting. *American journal of electroneurodiagnostic technology*, 50(1):8–25. 11

WHO (2020). Newborns: improving survival and well-being. *Geneva: World Health Organization*. 1

Ye, J., Zhang, J., Mikolajczyk, R., Torloni, M. R., Gülmezoglu, A., and Betran, A. (2016). Association between rates of caesarean section and maternal and neonatal mortality in the 21st century: a worldwide population-based ecological study with longitudinal data. *BJOG: An International Journal of Obstetrics & Gynaecology*, 123(5):745–753. 1

Yoshida, S., Martines, J., Lawn, J. E., Wall, S., Souza, J. P., Rudan, I., Cousens, S., Aaby, P., Adam, I., Adhikari, R. K., et al. (2016). Setting research priorities to improve global newborn health and prevent stillbirths by 2025. *Journal of global health*, 6(1). 65

Zhang, L., Zhou, Y.-X., Chang, L.-W., and Luo, X.-P. (2011). Diagnostic value of amplitude-integrated electroencephalogram in neonatal seizures. *Neuroscience bulletin*, 27(4). 13, 65