Towards a State-Dependent Model for Speech Enhancement in Hearing Aids

Nikodem Bieniek

MInf Project (Part 1) Report
Master of Informatics
School of Informatics
University of Edinburgh

2025

Abstract

This dissertation explores the intersection of Machine Learning, Automatic Speech Recognition (ASR) and Hearing Aids (HA) through the lens of Acoustic Scene Analysis (ASA) and Speech Enhancement (SE). The primary objective of this dissertation is to develop machine learning models capable of prioritising speech in noisy environments whilst remaining computationally feasible for implementation in hearing aids. The vision is to chain the ASA and SE models together to form a state-dependent model, in essence, a heuristic approach to SE that is tailored to the environment.

Accordingly, each proposed model is evaluated not only in terms of performance but also on its feasibility for real-time deployment, as measured by metrics such as Floating Point Operations Per Second (FLOPS) and the number of model parameters.

A growing trend in the market is the incorporation of deep neural networks (DNNs) into hearing aids. Evidence from state of the art manufacturers suggests that a combination of ASA and SE techniques are employed, further motivating the approach adopted in this project.

Experiments are conducted primarily on a novel dataset called HEAR-DS. As a baseline, the Convolutional Neural Network (CNN) ASA model presented by the authors of the dataset is utilised. We will demonstrate improvements in both the training procedure and the generalisation capability of the baseline CNN through the implementation of data augmentation strategies and the adoption of an alternative optimiser.

Furthermore, the dataset is extended by demonstrating a proof-of-concept of a speech enhancement model using a shallow CNN model that works in the frequency-time domain. We evaluate the performance of the model using artificial intelligibility and quality metrics such as STOI and PESQ. While the performance is not on par with the state-of-the-art, it most certainly motivates further research in the next part of the dissertation.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Nikodem Bieniek*)

Acknowledgements

First and foremost, I would like to solemly thank my supervisor, Hao Tang for his enthusiasm in guiding me through this self-proposed dissertation. I am eternally grateful for his wisdom, guidance, and patience throughout the project.

I would like to also thank my friends, for their support and encouragement throughout the project.

And to my parents, without whose sacrifices I would not have been able to pursue this degree, let alone this project.

Table of Contents

Chapter 1

Introduction

1.1 Motivations

Hearing loss is a prevalent condition affecting as much as 430 million people - or 1 in 18 people. This is expected to rise to 1 in 10 by 2050 [1]. The most common treatment for hearing loss is the provision of hearing technology - such as Hearing Aids (HAs) or cochlear implants. There are many types of HAs, but the most common type is the behind-the-ear (BTE) hearing aid [2]. However, HA users often report that they struggle to hear speech in noisy environments. For example, in a study by Kochkin [2], 42% of HA users reported that wind noise was a significant issue for them. This project aims to evaluate the effectiveness of machine learning algorithms in prioritising the speech in various environments (such as windy environments).

Modern hearing aids now apply a wide range of techniques to achieve better speech prioritisation. For wind noise reduction, this can be achieved from mechanical solutions - product design to covers that reduce wind noise - to signal processing techniques to compensate for mechanical limitations. However, current techniques are still not perfect as shown by the study from Kochkin.

Wind noise reduction and indeed, noise reduction in general, is a challenging problem when paired with speech. This is because you have to strike a balance between reducing background noise and preserving speech. Korhen's paper [3] outlines various techniques that could be used to reduce the wind noise in hearing aids - from modulation-based noise reduction algorithms (Wiener filtering), adaptive filtering algorithms, to machine learning techniques. The paper mentions that the the proposed ML technique: Long Short-Term Memory (LSTM) neural networks provided modest improvements in wind noise reduction, however, it did highlight that ML techniques may still have utility through further research and careful algorithmic choices.

This project aims to investigate the effectiveness of machine learning techniques in prioritising speech in noisy environments. The idea is to first train a deep neural network (DNN) to perform acoustic scene analysis (ASA) to classify the environment. Afterwards, a speech enhancement model will be trained to enhance the speech in the environment. From now on, we will refer to the chaining of the ASA and SE models

together as a state-dependent model as inspired by Katagiri [4]'s categorisation of SE techniques (see **??** for more details).

In this project, we will be using a novel dataset proposed by Hüwel et al. [5] This dataset (called HEAR-DS) is unique because it is specially tailored for HA signal processing and contains various environments. Normally, voice activity detection (VAD) would be used to detect speech, however, we will take inspiration from Hüwel et al. [5] by mixing the samples with speech and label them as so. This can be used to implicitly train the machine learning model to classify the environment and whether speech is present. Additionally, the paper presents an elementary example of how the dataset can be used: to classify the environment - it showcases the use of a convolutional neural network (CNN) to classify the environment. This project will be extending the paper by looking at how HEAR-DS can be used to additionally train a speech enhancement model and how the training and generalisation capabilities of the model can be improved.

The project will also be mindful in its algorithmic choices - as the computational power required in HA is limited. It is difficult to pinpoint the exact computational power of a HA due to the proprietary nature of the devices. In August 2024, Phonak (Sonova Holding AG) released a new HA which is their first AI equipped HA. The device is said to be capable of handling 7,700 Million Operations Per Second to accommmodate its neural network with 4.5 million parameters [6]. Contrast this with a paper from 2021 investigating techniques in VAD for hearing aids quotes that it 'rarely exceeds 5 million instructions per second (MIPS)' [7] Moreover, Apple's release of a FDA approved hearing aid in which was previously a mainstream earphone wearable, Apple Airpods, also shows that there's more interest in this area. So suffice to say, the computational power of hearing aids is accelerating and is most likely going to continue to grow given the increasing demand for HAs.

Delay constraints also play a critical role in HA performance, as real-time speech perception is essential for user satisfaction. Prior research suggests that a delay of up to 30 milliseconds (ms), and ideally less than 20ms, the user is unaffected by the delay [8]. Therefore, the project will be mindful in its algorithmic choices to ensure that the model is computationally efficient and feasible for real-time deployment.

The main evaluation metrics we will consider in this part of the dissertation is short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ). These are both artificial metrics and give an indication of the quality of the speech enhancement model. Their limitations are discussed in chapter **??**, with subsequent sections outlining strategies to address these shortcomings. Additionally, informed by current state-of-the-art HA configurations, the project will also assess model complexity by analysing the number of weights utilised in our designs.

The project is predominantly aimed at the hearing aid industry. If successful, the project could advance the state-of-the-art techniques in hearing aids. Which in turn could improve the quality of life for hearing aid users. A secondary goal is to make hearing aid research more accessible to the computer science community. The project also hopes to benefit other fields that deal with audio signal processing, such as mainstream wearables like headphones or microphones.

1.2 Structure

Going forward, the dissertation will be structured as follows:

- **Chapter 2** - This chapter will give a high level overview of hearing loss, the auditory system, and the speech processing concepts needed to understand the project. We briefly also cover the two techniques that will be used in this project: Acoustic Scene Analysis (ASA) and Speech Enhancement and some of the related work done in the field. Lastly, I mention the criticism of the paper by Hüwel et al. [5] that I will be exploring in this project.
- **Chapter ??** - lorem ipsum TODO

Chapter 2

Background

2.1 Hearing Loss

2.1.1 Auditory System

For sound to be registered by humans, it has to travel through the ear to transform them into what is known as a neural impulse which is then trasmitted to the brain. We will give a high level overview of the process which was collated by the phonetics textbook from Wayland [9], but for further reading, refer to the textbook. This process of sound to be registered by humans, is all done in what's known as the Auditory System. Figure 2.1 shows a birds eye view of the auditory system, and it can be split into three segments that we will explore below.

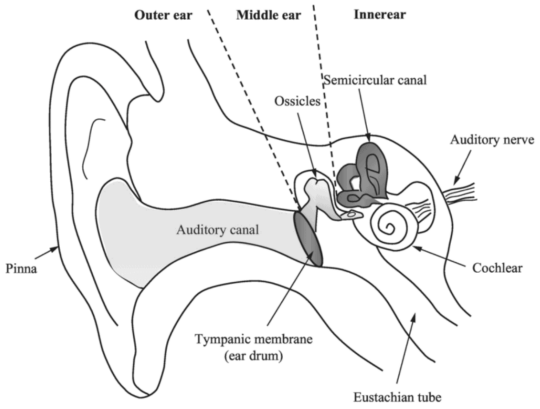

Figure 2.1: The external, middle and the inner ear from Wayland [9]

2.1.1.1 The Outer Ear

The outer ear (sometimes referred to as the External Ear) is responsible for channeling sound waves to the tympanic membrane (ear drum). Firstly, sound waves reach the Pinna which funnels it through the auditory canal, and at the end of the canal is the ear

drum, at which point the sound waves collide with the ear drum which causes the ear drum to vibrate. The vibrations are passed along to the middle ear.

2.1.1.2 The Middle Ear

Zooming into the middle ear (Figure 2.2), there are three bony structures: the Malleus, Incus, and Stapes, and together they make up a lever system. It is worth pointing out that when transmitting the energy from the ear drum to the middle ear, there is bound to be some energy loss [1]. The lever system's mechanical advantage allows for the sound energy to be amplified so the loss is compensated. The Stapes also has an additional function besides passing over the sound vibrations to the inner ear. The Stapes is connected by a muscle, the Stapedius muscle, and in response to loud noises, it temporarily increases the stifness of the bones in the middle ear, which temporarily prevents the acoustic energy to be amplified. This protects the inner ear from loud noises.

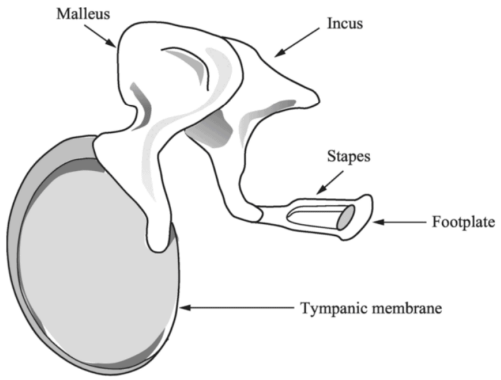

Figure 2.2: The components of the middle ear from Wayland [9]

2.1.1.3 The Inner Ear

One of the parts of the inner ear, is the Cohlea, and it is what gives humans the ability to hear sounds. The cohlea is a bony structure that resemblers that of a snail shell [2]. If the cohlea was unrolled its length would be about 3.5cm, and it is subdivided into various parts (see Figure 2.3). For our purposes, it suffices to know that in the Basilar membrane are a collection of cells, called the Organ of Corti, which contain hair cells, and that different hair cells register different frequencies. Figure 2.4 shows the frequency responses shows the various areas of the corti and its response to different frequencies. The Organ of Corti is linked to the audtiory nerves and so the oscillations that pass through it get coverted to neural impulses and transmitted to the brain.

[1] Depending on the frequency, this can be as high as 40%

[2] 'Cohlea' in Latin translates to snail shell!

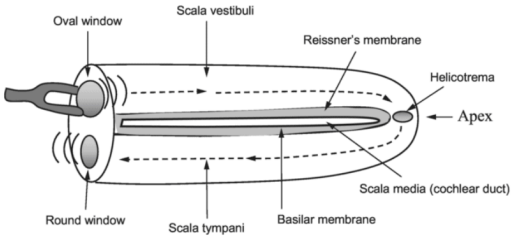

Figure 2.3: 'Unrolled' cohlear from Wayland [9]

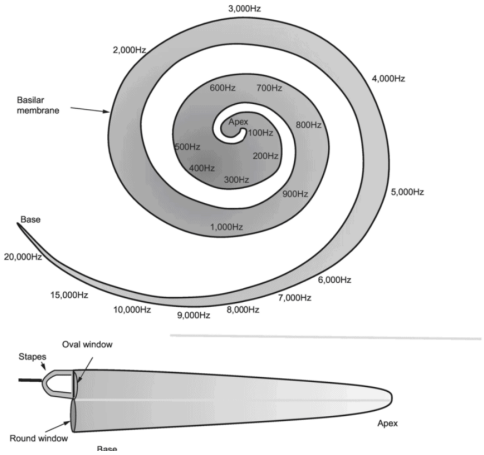

Figure 2.4: The Organ of Corti and the frequency responses. From Wayland [9]

2.1.2 Hair Cell Loss and Hearing Loss

As we have seen, the hair cells play a major role in allowing humans to register sound. Unfortunately, it is the major cause of hearing loss since those hair cells are suspectible to damage. The cause of hair cell loss is complex, since it could be due to many factors such as genetic abnormalities (congenital hearing loss), infection, diseases or extrinistic factors such as exposure to loud noises. It can also occur due to aging, which could be due to the reasons mentioned above, but also could be triggered due to age related hearing loss genes. To make matters worse, hair cell loss is unrecoverable in humans [10]. Hearing loss is a spectrum, and varies from mild (26-40dB loss), moderate (41-60dB loss), severe (61-80dB loss) to profound($>$ 80dB loss) [11]. Moreover, hearing loss may occur in one ear (unilateral) or in both ears (binaural).

2.2 Hearing Loss Treatment

2.2.1 History of Hearing Technology

TODO: Will see how much pages I have left after doing other chapters.

Users were fitted with what would have been considered a HA in this day and age as early as the 18th century.

2.2.2 Hearing Technology

As mentioned in the Introduction, the most common treatment for hearing loss is the provision of hearing technology. There are two common approaches to fitting of hearing technology: the equipment of a Hearing Aid (HA) or the fitting of a Cohlear Implant. There are similarities in the two, and both can be used to fit users that are experiencing profound hearing loss, though typically cohlear implants are only considered if the user is experiencing a severe hearing loss or further. While we focus on the HA, the concepts we explore can be applied to the CI. The components of a HA vary widely, but at the very least, it contains a microphone, amplifier, receiver, and battery [12]. A microphone is a device that converts acoustic sound waves into electrical signals. The electrical signals are then amplified by the amplifier, and the receiver converts the electrical signals back into the acoustic waves. The battery is used to power the device. Depending on the user, they may either be fitted with a hearing aid on each ear (binaural) or a single hearing aid on one ear (unilateral). Typically, the hearing aid contains multiple microphones, and the microphones are placed in different locations on the device. As you will see when we discuss the dataset we will be using, the samples are labelled with which microphone is being used.

2.2.3 Hearing Technology Now

Hearing aids are increasingly becoming more sophisticated and are starting to incorporate more powerful algorithms to improve the quality of life for hearing aid users. From Apple's FDA approved hearing aid in their Airpods, to big traditional players like Phonak, Oticon who are starting to incorporate DNN and AI into their devices. There is also a trend of rechargable hearing aids (Lithium Ion Batteries), and while it has a practical aspect i.e. the user does not need to change the batteries, it's also specifically used because of the ability to handle higher peak power consumption (which allows for powerful models to be used such as DNNs). Whereas non-rechargable hearing aids are restricted to what's allowed by the battery. There is however, a trade-off, as the battery life of the device is sacrificed. Battery powered hearing aids are most always using Zinc-air batteries [13] [14] and the most recent study by Thomas et al. [15] shows that the average battery life under high power conditions 50-80 hours. Contrast this with the average battery life of rechargable hearing aids which is typically in the region of 15-30 hours. In the case of Phonak's Speech Enhancer that kicks in when the user is in a loud environment, the maximum battery life is under 6 hours. This goes to show that using DNNs in HA can require careful design considerations but the potential benefits are worth it.

Due to the proprietary nature of the devices, it is unclear what sort of algorithms are used in the devices. Though, from Phonak's white paper on Spheric Speech Clarity [6] we can deduce that the devices use a combination of ASA and speech enhancement techniques. From the paper, before signal processing begins, the sound is fed into Autosense OS which performs 'scene classification' so in the context of this project, this can be considered as an acoustic scene analysis (ASA) algorithm. More specifically, if the signal is classified as 'Speech in Loud Noise', then the signal is fed into a deep neural network which outputs a mask that seperates the speech signal from the background

noise which is then applied to the signal. In the context of this project, we will aim to do something similar when we train a speech enhancement model on HEAR-DS.

2.3 Speech Processing Techniques

2.3.1 Digitisation

To be able to perform speech processing on a device containing a chip/processor such as a HA, requires us to have some sort of digital representation of the sound wave. A waveform is just that, a digital representation of the sound we hear. A microphone is a device that converts those sound waves into a representation of the waveform. Conceptually, this is done by measuring the relative air pressure at different points in time, and this is represented as a waveform or a time series of amplitudes. Digital devices can't have infinite precision, so the signal produced by the sound will be divided into discrete samples, and the rate at which this occurs is known as the sampling rate. Additionally, the precision of the amplitude of the signal will be quantized into discrete numbers, and the precision is dictated by the bit depth (or the quantization rate).

The selection of these two parameters is determined by the nature of the speech task and the Nyquist-Shannon Sampling Theorem. This theorem asserts that the sampling rate should be at least twice the highest frequency present in the signal. That is, for a signal containing frequencies up to 8KHz, the minimum required sampling rate would be $f_s \geq 16$KHz. Otherwise, aliasing will occur, which is the phenomenon where high frequency components could be misinterpreted as lower frequencies, which could be undesirable in both tasks we are trying to perform. Typically, datasets will have a sampling rate of 44KHz, and commonly for ASA tasks, a sampling rate of 16KHz is acceptable. Regarding the bit depth, it is selected based on the dynamic range required to capture speech signal variations. Commonly, 16-bit depth is used, which is capable of representing an amplitude range of approximately 96dB, thereby ensuring sufficient resolution in amplitude representation.

2.3.2 Engineered Acoustic Features

The prevalance of engineering acoustic features in previous work relating to my project and speech processing in general, warrants a high level overview of this concept to better understand its usage. Given some signal $x(t)$, an engineered feature $f(x(t))$ is a function that transforms the signal in some way. In this project we will be feeding the models the log mel spectrogram of the signal. To obtain the log mel spectrogram, we first need to convert the signal (waveform) into the frequency domain (spectrogram). The process of converting the signal into the frequency domain is done by utilising the Fourier Transform. We will technically be using the Fast Fourier Transform (FFT) due to efficiency, and the main hyperparameters for the FFT are the window length w and the hop length h. The window length is the length of the window that is used to split the signal into frames, and the hop length is the number of samples between the start of each frame. Additionally, there is a choice of a window function and we will be using the most common one, the Hann window. Once we apply the FFT, we will be left

with a complex valued vector of length w and the magnitude of this vector gives us the amplitude of the signal at different frequencies. We can then apply a Mel Filterbank - a filterbank that is used to mimic the human auditory system. Afterwards, we take the logarithm of the filterbank outputs and we are left with the log mel spectrogram which mimics the human perception of loudness. Diagram **??** shows the process to make it more clear.

The reason it is worthwhile to use an engineered feature is for many reasons, but one of the main ones is that it is a compact representation of the signal which reduces the amount of memory and computational resources required to process the signal which is especially important for embedded devices such as HAs. Additionally, the log mel spectrogram enhances the distinction between speech and background noise, which may make a model more effective at ASA and speech enhancement tasks.

2.4 Related Work

As this project will chain two algorithms—Acoustic Speech Analysis (ASA) and Speech Enhancement, to form a state-dependent model, we will cover each task in separate sections.

2.4.1 Acoustic Scene Analysis (ASA)

Acoustic Scene Analysis (ASA) refers to the task of classifying an environment based on an audio stream.

We will be using HEAR-DS dataset introduced by Hüwel et al. [5]. This dataset was developed in response to the inadequacies of existing acoustic scene classification databases for HA processing. A more detailed description of the dataset is provided in Chapter 4.1. The original paper on the HEAR-DS dataset illustrates its applicability by employing a Convolutional Neural Network (CNN) for environment classification. In Chapter 5, an attempt is made to reproduce the results presented in that study. The challenges encountered during this replication process are discussed comprehensively in the chapter. Additionally, to validate the model implementation, the same parameters were applied to the DCASE 2017 Acoustic Scenes Challenge dataset [16]. In particular, a comparison is drawn between the model proposed by Schindler et al. [17] and the CNN model used in the HEAR-DS paper, with the findings detailed in Section **??**. The two papers mentioned above utilise engineered features, and more specifically, the log mel spectrogram. This approach in some literature is called operating on the frequency domain. It is worth pointing out that there is research into doing ASA on the time domain i.e. operating on the waveform directly, such as the paper by [18] - [19]. While not explored in this part, it is something I will look into exploring in part 2 of this dissertation as there is a compelling reason to possibly consider the time domain approach. Namely, the overhead of the Fourier Transform and the Mel Filterbank in a hearing aid is potentially too high, and so working directly with the waveform may be a more feasible approach.

2.4.2 Speech Enhancement

On the other hand, Speech Enhancement aims to improve the perceptual quality of a speech signal that has been degraded by an additive noise [20]. This enhancement is crucial in applications where speech intelligibility is essential, such as telecommunications, hearing aids, and speech recognition systems. Noise sources may include environmental disturbances (e.g., wind noise) as well as interference from multiple speakers (e.g., babble noise).

There are many different approaches to speech enhancement, from more traditional methods such as spectral subtraction, or Wiener filtering to more modern methods that incorporate neural-networks. In this dissertation, we will be using a neural-network based approach. But even within the neural-network based approach, there are different techniques that can be used to enhance the speech signal. Neural network-based speech enhancement algorithms can be categorised into four primary types, as described in [4]:

- **Time-Domain Filtering** - Trains a neural network with noisy inputs (background + speech) and clean targets (speech only). This involves working with the waveform directly, however it is not without its challenges which is still an active area of research [21].
- **Transform-Domain Filtering** - Trains a neural network with noisy inputs (background + speech) and clean targets (speech only). The differnece between this and the Time-Domain Filtering approach is that the input is transformed into the frequency domain (whether that be a Fourier Transform, or a Mel Spectrogram, Mel-Frequency Cepstral Coefficients (MFCCs), LPC, etc). The advantage of this is that the dimensionality of the input is usually lower, and the representation is more robust to seperation of the speech and background noise. This is also a popular approach in the speech enhancement literature [22].
- **State-Dependent Model Switching** - The two mentioned approaches assumes that the speech signal and noise source are stationary. This is too strong of an assumption, and so state-dependent model switching is a paradigm that has been explored. The main idea is to have a class of models, each specialising in handling a different type of noise and to switch between them based on the characteristics of the input signal. This in a way is the chaining of ASA and Speech Enhancement, as the ASA model will first classify the input signal into a state, and then the appropriate speech enhancement model will be selected and applied to the signal.
- **Online Iterative Methods** - The focus is on incorporating adaptive techniques that do not rely on pre-existing training data. Key approaches include adaptive predictors, dual EKF algorithms, and noise-regularised adaptive filtering.

It is no accident I talked about Engineering Acoustic Features in the previous section, as this dissertation will be focusing on the Transform-Domain Filtering approach. One of the primary objects of part two of this dissertation will be investigating how ASA and Speech Enhancement can be integrated to form a state-dependent model. Though, exploring the time domain approach is something I will be looking into for the second part of this dissertation due to the potential overhead reduction of the time domain

approach as mentioned in the previous section.

2.5 Criticism of Previous Work

During the course of this project, I found out that the paper by Hüwel et al. [5] had some issues. Granted, the paper was more of an exploration of the applicability of the dataset rather than a thorough analysis of the performance of the model, it is still important to point out the issues I found. First of all, the paper was not clear in the details of the splitting strategy of the dataset so I had to make some assumptions. Additionally, it was not clear what loss function was used for training the model. As the dataset only contains raw cuts of the background noise, and the researchers opted to mix the speech signal with the background noise so that they have full control over the SNR of the signal, they used another dataset to mix the speech signal with the background noise. More specifically, they used the CHiME 3 [23] and CHiMe 5 [24] datasets. However, only the CHiME 3 development set was used for mixing the speech into the background noise. CHiME 5 was used for creating a new environment, 'Interferring Speakers', which are samples that contain speech from multiple speakers. I think using the CHiME 3 dataset and especially only the development set was an odd decision, as there is only roughly 10 hours of data from 4 speakers. So a model trained on this dataset may not generalise well to real world scenarios, due to the rich variability of speech from different speakers. Nevertheless, we continue with the paper's approach to set up a baseline. Another odd decision was in the use of a fixed learning rate scheduler, and the use of a high epoch count without what appears to be no early stopping. This will be discussed in more detail in Chapter 5. We still thank the authors for their work on the dataset as it is a promising dataset for the future of HA research. Additionally, given that the dataset has the potential to be used for HA devices, there is not any mention on the viability of the models when used in a HA device. So a significant aspect of this project is to investigate the feasibility of incorporating a DNN model into a HA device.

Chapter 3

Methodology

3.1 Acoustic Scene Analysis

Mathematically, ASA can be formalised as follows: Consider an audio signal $x \in \mathbb{R}^T$, where T denotes the number of samples in the signal, and let $\mathcal{E} = \{e_1, e_2, \ldots, e_N\}$ represent a set of N distinct environments. ASA is concerned with determining a mapping

$$f : \mathbb{R}^T \rightarrow \mathcal{E},$$

which assigns each audio signal to one of the environments in $\mathcal{E}$. In this project, f is implemented as a neural network and in this first part of the dissertation, we will be using a Convolutional Neural Network (CNN) to implement f.

3.1.1 Model

Model	CNN_1	CNN_2	FC
net-8	8	16	25
net-20	20	40	63
net-32	32	64	100

Table 3.1: Model parameters for the net-X models.

As our baseline model, we will be using the CNN model presented in the paper by Hüwel et al. [5]. The model has three hyperparameters: the number of output channels for the first and second convolutional layers (CNN_1 and CNN_2), and the number of neurons in the fully connected layer (FC). The researchers investigated the effect of different model parameters, and in the paper, they train the dataset on 7 different models, where they vary the number of output channels in intervals of 4, thereby training 7 different models. Due to the time constraints, we will be training 3 models, as we're mostly interested in seeing if we're getting similar results to the paper. Table 3.1 shows the parameters of the models we will be training.

The model architecture is shown in Figure 3.1. As mentioned earlier the models operate on the frequency domain, and we stuck with the paper's approach of using a Mel

Spectrogram as the frequency domain representation for reasons mentioned in the background section. As per the paper, all the convulaotional layers use a 7×7 kernel with a stride of 1×1 and padding of 3×3. After the convoluational operation, batch normalisation is applied which can help with the stability of forward propagation and can act as a regulariser which can help prevent overfitting [25]. The choice of activation function is a Rectified Linear Unit (ReLU) which is computionally efficient and reduces the risk of the vanishing gradient problem [25]. As the preceding operations can increase the dimensionality of the input, a form of down-sampling is applied by using a max-pooling operation which helps in achieving translation invariance and reducing computational complexity for the following layers. Something worth considering for part 2 of this dissertation is to explore the use of other down-sampling techniques, and in particular, Liu et al. [26] have proposed a family of simple pooling front-ends (SimPFs) which in some cases reduce the number of FLOPs as much as 75% in acoustic scene classification tasks. Lastly, dropout of 0.3 is applied after the max-pooling operation which can help with the model generalisation [25]. Applying convolutions is not enough for ASA, as we need to have a way to compare the output of the CNN with the set of environments $\mathcal{E}$. This is where the fully connected layer comes in, as it will output a vector of size N, where N is the number of environments. The values of the vectors are logits, and so during training and inference, we use a softmax function to convert them into probabilities. The softmax function is defined as follows:

$$\sigma(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^{N} \exp(x_j)},$$

where x_i is the i-th element of the vector. The environment with the highest probability is then chosen as the predicted environment i.e. $\hat{e} = \arg\max_{e_i \in \mathcal{E}} \sigma(x_i)$.

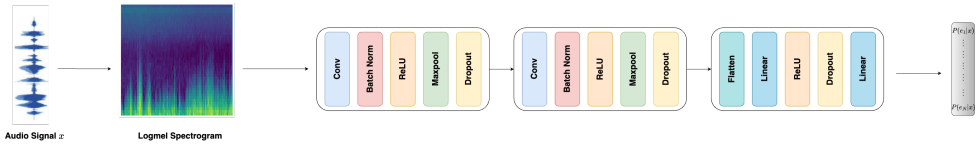

Figure 3.1: The baseline ASA model architecture.

3.1.2 Training

For ASA, we need to have a dataset that contains audio signals along with their corresponding environmental labels. Mathematically, we need to have a dataset of the form $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^T$ is the audio signal, and $y_i \in \mathcal{E}$ is the environmental label. Then a loss function is needed to steer the model towards the correct mapping. It is important to note that the dataset $\mathcal{D}$ is not necessarily balanced, meaning that some environments may have more samples than others. This is a problem, as the model may learn to classify all samples as the most frequent class. So to combat this, when using a loss function, we apply a weight to each sample based on the number of samples in the dataset. The weight is defined as follows:

$$w_i = \frac{n_{samples}}{n_{classes} \cdot n_i}$$

where $n_{samples}$ is the total number of samples, $n_{classes}$ is the number of unique classes, and n_i is the number of samples in the i-th environment. This balanced weighting scheme ensures that the weights sum up to $n_{samples}/n_{classes}$. So a vector of weights $\mathbf{w} \in \mathbb{R}^N$ is created, where N is the number of environments So we have a vector of weights $\mathbf{w} \in \mathbb{R}^N$ which we can use to weight the loss function accordingly. This is consistent with the paper by Hüwel et al. [5] where they did "weighted update steps for each target environment depending on the number of samples in each environment." On the topic of the loss function, the paper by Hüwel et al. [5] does not explicitly mention the loss function they used, but we assume that they used the weighted cross-entropy loss function. The weighted cross-entropy loss function is defined as follows:

$$L(y, \hat{y}) = -\sum_{i=1}^{N} w_i \cdot y_i \log(\hat{y}_i),$$

where y_i is the one-hot encoded environmental label, and $\hat{y}_i$ is the predicted probability of the model.

Once we have the loss function, we can use it to train the model. There are many approaches to training a model, and for reproducibility, we will be using the paper's approach of using Stochastic Gradient Descent (SGD) with a fixed learning rate. The learning rate was decreased every 40 epochs in a quasi-logarithmic manner: [0.05, 0.01, 0.001, 0.0005, 0.0002, 0.0001]. This leads to a total of 240 epochs of training. However one of objectives of this part of the dissertation was to explore if a different approach to learning rate scheduling could yield better results. So we will be exploring a different approach to learning rate scheduling in the experiments. It will be found that the Adam optimiser converges faster than the SGD optimiser and still results in a comparable performance.

3.1.3 Evaluation

The evaluation of the ASA model will be done using the accuracy metric. The accuracy is defined as follows:

$$Accuracy = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i = \hat{y}_i),$$

where N is the total number of samples, y_i is the one-hot encoded environmental label, and $\hat{y}_i$ is the predicted environmental label and $\mathbb{I}$ is the indicator function which is a piecewise function that returns 1 if the condition is true and 0 otherwise.

Evaluation from a feasibility of a hearing aid perspective will also be done. This will be done by looking at the following metrics:

- **Inference Time** - This will be done by looking at the time it takes for the model to process a sample.
- **Floating Point Operations (FLOPs)/Multiply Accumulate (MACs)** - This will be done by looking at the number of operations the model performs.
- **Number of Parameters** - This will be done by looking at the number of trainable parameters in the model.

3.2 Speech Enhancement

Mathematically, let a noisy speech signal be represented by $x \in \mathbb{R}^T$ and the corresponding clean speech signal by $y \in \mathbb{R}^T$, where T denotes the total number of samples. Speech enhancement is therefore concerned with determining a mapping

$$f : \mathbb{R}^T \to \mathbb{R}^T,$$

which takes a noisy speech signal x and produces an enhanced version $f(x)$ that is as close as possible to the clean speech signal y i.e.

$$f(x) \approx y.$$

Just as in the previous section, we will be using a Convolutional Neural Network (CNN) to implement f.

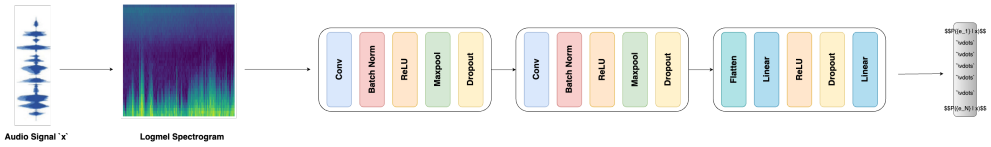

Figure 3.2: The speech enhancement model architecture.

3.2.1 Model

As this is more of an exploration of the viability of using a DNN model for speech enhancement and the establishment of a baseline which can be used to improve upon, we will start with a simple 5-layer CNN model. The model is shown in Figure 3.2. As we want to preserve the dimensionality of the input, we have removed batch normalisation and pooling layers. For similar reasons, we have set the padding to 1×1 which ensures that the output has the same dimensionality as the input and the kernel size is set to 3×3. We kept the activation function as a ReLU. Since this model has 5 layers, it has 5 hyperparameters: the number of output channels for each layer $(C_1, C_2, C_3, C_4, C_5)$. Table 3.2 shows the parameters of the model.

Layer	Output Channels
C_1	64
C_2	128
C_3	64
C_4	32
C_5	40

Table 3.2: Model parameters for the speech enhancement model.

It is worth noting that the output of the model is a complex spectrogram, which would not be suitable for a HA device. So we need to convert the output back to the time domain. There are two approaches that we have seen in the literature:

- **Inverse Short-Time Fourier Transform (ISTFT)** - As the name suggests, this approach takes the inverse of the Short-Time Fourier Transform (STFT) of the input signal. So the input is a magnitude and phase spectrogram, and the output is a time domain signal.
- **Griffin-Lim** - This is an iterative algorithm that estimates the phase of the signal just from the magnitude spectrogram alone.

For the purposes of this dissertation, the Griffin-Lim approach was adopted, as the initial implementation did not track phase information however, it is something that I will be exploring in the future.

3.2.2 Training

We will adopt the Adam optimiser instead of SGD as one of the experiments I will be doing is to validate if the claim made in Hüwel et al. [5] that a fixed learning rate SGD approach is better than an Adam optimiser approach. As will be discussed in Chapter **??** we found that the Adam optimiser converged faster than the SGD optimiser and still resulted in a comparable performance. Additionally, from the ASA experiments, we also found that adopting an early stopping criterion once the training loss plateaued, sped up the training process at no loss of performance. So we will be adopting this approach for the speech enhancement experiments as well. More specifically, we will be using a learning rate of 0.001 and early stopping once the training loss stops improving after 5 epochs.

The way the model is trained is by comparing the ground truth speech signal with the enhanced speech signal. As such, we have used the Mean Squared Error (MSE) loss function to train the model. The loss function is defined as follows:

$$L(y,\hat{y}) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2,$$

where y_t is the ground truth speech signal and $\hat{y}_t$ is the enhanced speech signal.

3.2.3 Evaluation

The evaluation of the speech enhancement model in this part of the dissertation will be done by assessing two artificial metrics. Firstly, the Short-Time Objective Intelligibility (STOI) metric will be used to assess the intelligibility of the enhanced speech signal. Mathematically, the STOI metric is defined as follows:

$$STOI = \frac{\sum_{t=1}^{T}\mathbb{I}(y_t = \hat{y}_t)}{\sum_{t=1}^{T}\mathbb{I}(y_t \neq 0)},$$

where y_t is the ground truth speech signal and $\hat{y}_t$ is the enhanced speech signal. TODO: Verify this formula Secondly, the Perceptual Evaluation of Speech Quality (PESQ) metric will be used to assess the perceptual quality of the enhanced speech signal. Mathematically, the PESQ metric is defined as follows:

$$PESQ = \frac{1}{T}\sum_{t=1}^{T}\frac{y_t\hat{y}_t}{\|y_t\|_2\|\hat{y}_t\|_2},$$

where y_t is the ground truth speech signal and $\hat{y}_t$ is the enhanced speech signal. TODO: Verify this formula

3.3 Laying the Groundwork for State-Dependent Models

In the preceding sections, we have established the groundwork for the two tasks that we will be doing in this part of the dissertation. The reason we are devising these two tasks in the first place is to set the groundwork for the next part of the dissertation which will be focused on the implementation of a state-dependent model. We hypothesise that a feasible state-dependent model can be created by chaining the two models together. Figure **??** shows the vision we have for the state-dependent model of which we aim to implement in the next part of the dissertation. TODO: Explain the diagram We thought that adding this section would give you the reader a better understanding of the motivation of this part of the dissertation.

Chapter 4

Experimental Setup

4.1 Datasets

4.1.1 HEAR-DS

Given how much this dissertation revolves around the HEAR-DS dataset, we dedicate a section solely to the dataset. HEAR-DS is a dataset created by Hüwel et al. [5] which is specially tailored for Hearing Aid (HA) research. It consists of recordings captured on a dummy head equipped with Pinnae models and three microphones per side: one In-The-Canal (ITC) [in the ear] and two Behind-The-Ear (BTE) microphones (front and rear). Each sample is recorded in either the left or right channel and decomposed into its respective microphone channels (BTE_front, BTE_rear, ITC), available in 48kHz/32bit format. In the paper, the researchers used only the ITC samples, so for the scope of this project, we will only use the ITC samples as well. As the researchers were based in Germany, any recordings that they took were situated in Germany. The dataset provides 7 environments:

- **Cocktail Party** - Covers situations such as multiple speakers in a noisy environment or 'babble' noise. The paper mentions that this was mimicked by recording them in a university cafeteria or at a senior citizens' meeting.
- **Wind Turbulence** - Concerns the sound that is produced when wind passes through a microphone.
- **In-Traffic** - Samples where traffic noise is dominant such as bus stations or sidewalks. The researchers controlled for this environment to not be similar to the *Wind Turbulence* environment by choosing 'calm' days for the recordings.
- **In-Vehicle** - The sounds produced when seated inside a car. The researchers controlled for this environment to not be similar to the *In-Traffic* or *Wind Turbulence* environments by ensuring the windows were closed.
- **Quiet Indoors** - Mimic noises produced in a typical household such as washing dishes, clock ticking, etc. The researchers controlled for variability by recording the samples in several flats (rural, city centre, etc).

- **Reveberant** - Samples in highly reverberant environments such as a railway station hall, staircases or a church.
- **Music** - The data is actually taken from the GTZAN dataset [27] and is used to test the model's ability to classify music.

Additionally, the researchers artificially created a new environment, *Interfering Speakers*, which are samples that contain speech from multiple speakers. This is done by taking speech samples from the CHiME 5 dataset [24]. This environment is supposed to mimic the typical conversational speech that occurs in a real-world scenario. We however, decided to use the CHiME 6 dataset which supercedes the CHiME 5 dataset [24] by correcting the alignment of audio channels and the researchers recommend the use of the CHiME 6 dataset instead of CHiME 5 for any new research so we will use that instead. This could be an important factor as the methodology of creating the *Interfering Speakers* environment is by finding 10s segments of speech that contain multiple speakers, so if the alignment of audio channels is not correct, the recordings may have a delay between the left and right channels which could have potentialy affected the results of the paper. We will discuss the dataset in more detail in section 4.1.3.

To prevent the model from overfitting and increase the diversity of the data, the researchers provided multiple REC-SITs (Recording Situations) for each environment. In the case of the *Music* environment, each REC-SIT represents a different genre of music. This has some potential to cause issues and we discuss this in more detail once we aim to reproduce the results of the paper (Chapter 5). As for the artificially created *Interfering Speakers* environment, it is unclear how they define REC-SITs for this environment. What I did was to extract the session ID from the CHiME 6 dataset and use that as the REC-SIT, which gives us a total of 16 REC-SITs for this environment.

At the REC-SIT level, the raw dataset comprises multiple cuts, with each cut not necessarily of uniform duration. In accordance with the methodology of the original study, each cut was segmented into 10s snippets. Table 4.1 presents the distribution of samples across the various environments after segmentation. Overall, our statistics are largely consistent with the original report, albeit with minor discrepancies in the *Interfering Speakers* and *Cocktail Party* environments. Specifically, our version of the *Interfering Speakers* environment comprises 1,364 samples compared to the 1,481 reported originally, a difference that may be attributable to the aforementioned alignment issues in the CHiME 5 dataset. Similarly, the *Cocktail Party* environment contains 716 samples in our dataset as opposed to 667 in the original report—a discrepancy that could potentially result from a typographical error in the original document.

As noted in the Introduction, the model can be made to act as a Voice Activity Detector (VAD) by incorporating environments with and without speech. The raw dataset however, only contains samples without speech, so we will be following the paper's approach to create the mixed environments. Although recording environments with and without speech was something the researchers did consider, the researchers prioritised full control over the signal-to-noise ratio (SNR) during the speech-background mixing process. One deviation from the original methodology is that, in this dissertation, the creation of mixed environments will not incorporate a head-related transfer function

Environment	Total No of Samples
Cocktail Party	716
Wind Turbulence	1364
In-Traffic	1000
In-Vehicle	1094
Quiet Indoors	951
Reveberant	1007
Music	2991
Interfering Speakers	1364

Table 4.1: HEAR-DS dataset statistics after cutting each environment's cuts into 10s snippets.

(HRTF) due to time constraints and the scarcity of readily usable libraries. Otherwise, the approach remains the same: using the CHiME 3 dataset (see Section **??**) to blend the speech signal with background noise. We ensure in a similar manner to the paper that each mixed 10-second snippet contains at least 7.5 seconds of speech. It was not clear how the researchers split the speakers for the mixed environments, so we will be splitting the speakers into two groups—one for training and one for testing. Table **??** shows that the CHiME 3 dataset includes four speakers with an equal gender split, thus we ensured for generalisability that each set contained a male and female speaker.

4.1.2 CHiME3

As mentioned earlier, the researchers used the CHiME 3 dataset to mix the speech signal with the background noise. CHiME 3 is a dataset published in 2015 for ASR tasks, and was dedicated to improving the performance of mobile devices in everyday, noisy environments. The vocabulary of the speech samples is taken from a subset of the Wall Street Journal (WSJ0) corupus [**?**]. We won't go into the details of the environments, because while not exactly clear from the paper, it is mentioned that a development set was used for the mixing. We assume that they used the isolated development set recorded in a booth, as it would not make sense otherwise since the other sets contain noisy samples. The use of the development set for mixing is not ideal, as it is not representative of the real-world data. In particular, as can be seen from Table **??**, it contains 10 hours of data from 4 speakers. This could be a problem in both tasks (ASA and Speech Enhancements), in the former we could be overfitting to the data due to the Pigeon-hole principle i.e. we have much more background samples than speech samples, and in the latter, the enhancement model may not generalise well to real-world scenarios due to the small size of the dataset. A better approach would of perhaps been to use the cleanm speech samples Nevertheless, we will continue with the paper's approach to set up a baseline. The data was already in 16KHz so we did not need to resample it.

4.1.3 CHiME5/CHiME6

The CHiME 5 dataset was originally used in the paper to create the *Interfering Speakers* environment for the HEAR-DS dataset. However, since the CHiME 6 dataset supersedes CHiME 5 with improved audio channel alignment, we opted to use CHiME 6 for our implementation. CHiME 6 (and CHiME 5) was a dataset created for the Speech Seperation and Recognition Challenge in 2020. It focuses on conversational speech in everyday home environments and particular emphasis was placed on eliciting a 'dinner party' scenario i.e. a mixture of speech from multiple speakers. It was not clear from the paper what sets were used for the creation of the *Interfering Speakers* environment however, they did provide a table of number of samples in each environment. From the table, we were able to deduce that they must have used the training, development and evaluation sets for the creation of the *Interfering Speakers* environment. We talk more about the creation of this environment when we discuss the baseline model in Chapter 5. From Table **??**, we can see that the dataset contains almost 50 hours of data from 48 speakers. TODO Table

4.1.4 TUT Acoustic Scenes 2016

As mentioned in the previous chapter, I came across better results than the baseline model in the paper by Hüwel et al. [5] so as a means of validation, I decided to use the TUT Acoustic Scenes 2016 dataset [28] to test the performance of the model. In particular, I will comparing the results by Schindler et al. [17]. The paper presents a multi-temporal approach to ASA, and the authors have used the development set of the TUT Acoustic Scenes 2016 dataset for their experiments. It is worth stressing that this dataset contains background noise only, and the paper does not evaluate it with speech, so when we use it to validate the model, we will maintain the same approach as the paper.

4.1.5 VOICEBANK + DEMAND

TODO

4.2 Hypotheses

1. **Using Fixed Learning Rate Stochastic Gradient Descent (SGD) will result in a longer training time than using Adam Optimiser.** The paper by Hüwel et al. [5] used the fixed learning rate SGD approach. I hypothesise that using Adam Optimiser will result in a shorter training time.
2. **Data Augmentation will result in a higher accuracy of the model.** It is unclear from the paper whether the authors used data augmentation...
3. **Using an LSTM model will result in a higher accuracy of the model.**

Chapter 5

Acoustic Scene Classification Experiments

5.1 Data Preparation

TODO

5.2 Fixed Learning Rate SGD Approach

TODO

5.3 Adam Optimiser Approach

TODO

5.4 Validating against TUT Acoustic Scenes 2016

Chapter 6

Speech Enhancement Experiments

6.1 Data Preparation

TODO

6.2 One Model Approach

TODO

6.3 Per Environment Model Approach

TODO

6.4 Validating against VOICEBANK + DEMAND

Chapter 7

Conclusions

7.1 Discussion

TODO

7.2 Future Work

TODO

Bibliography

[1] World Health Organization. Deafness and hearing loss, 2024. URL https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

[2] Sergei Kochkin. Marketrak VIII: Consumer satisfaction with hearing aids is slowly increasing. *The Hearing Journal*, 63(1), 2010. doi: 10.1097/01.HJ.0000366912.40173.76.

[3] Petri Korhonen. Wind Noise Management in Hearing Aids. *Seminars in Hearing*, 42(3), 2021. doi: 10.1055/s-0041-1735133.

[4] Shigeru Katagiri. *Handbook of Neural Networks for Speech Processing*. Artech House, Inc., USA, 1st edition, August 2000. ISBN 978-0-89006-954-7.

[5] Andreas Hüwel, Kamil Adiloğlu, and Jörg-Hendrik Bach. Hearing aid research data set for acoustic environment recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. doi: 10.1109/ICASSP40776.2020.9053611.

[6] Krylova Alena Hasemann Henning. Revolutionary speech understanding with spheric speech clarity. Technical report, Phonak, 2024. URL https://www.phonak.com/content/dam/phonak/en/evidence-library/white-paper/techn

[7] Joaquín García-Gómez, Roberto Gil-Pita, Miguel Aguilar-Ortega, Manuel Utrilla-Manso, Manuel Rosa-Zurera, and Inma Mohino-Herranz. Linear detector and neural networks in cascade for voice activity detection in hearing aids. *Applied Acoustics*, 175, 2021. doi: 10.1016/j.apacoust.2020.107832.

[8] Michael A. Stone and Brian C. J. Moore. Tolerable Hearing Aid Delays. II. estimation of limits imposed during speech production. *Ear and Hearing*, 23(4), 2002. doi: 10.1097/00003446-200208000-00008.

[9] Ratree Wayland. *Phonetics: A Practical Introduction*. Cambridge University Press, 2018.

[10] David N. Furness. Molecular basis of hair cell loss. *Cell Tissue Research*, 361: 387–399, 2015. doi: 10.1007/s00441-015-2113-z.

[11] Carrie L. Nieman and Esther S. Oh. Hearing loss. *Annals of Internal Medicine*, 173(11):ITC81–ITC96, 2020. doi: 10.7326/AITC202012010. URL https://www.acpjournals.org/doi/abs/10.7326/AITC202012010.

[12] James Schuster-Bruce and Emilee Gosnell. Conventional Hearing Aid Indications and Selection. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2025. URL http://www.ncbi.nlm.nih.gov/books/NBK567712/.

[13] C. Sparkes and N. K. Lacey. A study of mercuric oxide and zinc-air battery life in hearing aids. *The Journal of Laryngology and Otology*, 111(9):814–819, September 1997. ISSN 0022-2151. doi: 10.1017/s002221510013871x.

[14] Shaila Mir, Sunali Vij, and Nikhil Dhawan. Evaluation of end-of-life zinc-air hearing aid batteries for zinc recovery. *Minerals Engineering*, 198: 108082, July 2023. ISSN 0892-6875. doi: 10.1016/j.mineng.2023.108082. URL https://www.sciencedirect.com/science/article/pii/S0892687523000961.

[15] James Thomas, Barry Bardsley, Jane Wild, and Michael William Owen Penman. Zinc–Air Hearing Aid Batteries: An Analysis of Functional Performance. *Audiology Research*, 14(4):659–673, August 2024. ISSN 2039-4349. doi: 10.3390/audiolres14040056. URL https://www.mdpi.com/2039-4349/14/4/56. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[16] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen. DCASE 2017 challenge setup: Tasks, datasets and baseline system. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pages 85–92, November 2017.

[17] Alexander Schindler, Thomas Lidy, and Andreas Rauber. Multi-Temporal Resolution Convolutional Neural Networks for Acoustic Scene Classification, November 2018. URL http://arxiv.org/abs/1811.04419. arXiv:1811.04419 [cs].

[18] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. Very Deep Convolutional Neural Networks for Raw Waveforms, October 2016. URL http://arxiv.org/abs/1610.00087. arXiv:1610.00087 [cs].

[19] T. Vijaya Kumar, R. Shunmuga Sundar, Tilak Purohit, and V. Ramasubramanian. End-to-end audio-scene classification from raw audio: Multi time-frequency resolution CNN architecture for efficient representation learning. In *2020 International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5, Bangalore, India, July 2020. IEEE. doi: 10.1109/SPCOM50965.2020.9179600. URL https://ieeexplore.ieee.org/abstract/document/9179600.

[20] Philipos C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, Boca Raton, June 2007. ISBN 978-0-429-13373-2. doi: 10.1201/9781420015836.

[21] Nasir Saleem, Teddy Surya Gunawan, Sami Dhahbi, and Sami Bourouis. Time domain speech enhancement with CNN and time-attention transformer. *Digital Signal Processing*, 147:104408, April 2024. ISSN 1051-2004. doi: 10.1016/j.dsp.2024.104408. URL https://www.sciencedirect.com/science/article/pii/S1051200424000332.

[22] Zhongshu Hou, Qinwen Hu, Kai Chen, Zhanzhong Cao, and Jing Lu. Local spectral attention for full-band speech enhancement. *JASA Express Letters*, 3(11):

115201, November 2023. ISSN 2691-1191. doi: 10.1121/10.0022268. URL https://pubs.aip.org/jel/article/3/11/115201/2919537/Local-spectral-attention-

[23] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 504–511, December 2015. doi: 10.1109/ASRU.2015.7404837. URL https://ieeexplore.ieee.org/abstract/document/7404837.

[24] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines. In *Proc. Interspeech 2018*, pages 1561–1565, 2018. doi: 10.21437/Interspeech.2018-1768.

[25] Simon J.D. Prince. *Understanding Deep Learning*. The MIT Press, 2023. URL http://udlbook.com.

[26] Xubo Liu, Haohe Liu, Qiuqiang Kong, Xinhao Mei, Mark D. Plumbley, and Wenwu Wang. Simple Pooling Front-Ends for Efficient Audio Classification. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, June 2023. doi: 10.1109/ICASSP49357.2023.10096211. URL https://ieeexplore.ieee.org/document/10096211. ISSN: 2379-190X.

[27] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, July 2002. ISSN 1558-2353. doi: 10.1109/TSA.2002.800560. URL https://ieeexplore.ieee.org/document/1021072/?arnumber=1021072. Conference Name: IEEE Transactions on Speech and Audio Processing.

[28] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. TUT database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1128–1132, August 2016. doi: 10.1109/EUSIPCO.2016.7760424. URL https://ieeexplore.ieee.org/document/7760424/?arnumber=7760424. ISSN: 2076-1465.