Evaluation of Machine Learning Algorithms for Speech Prioritisation in Noisy Environments

Nikodem Bieniek

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

2025

Abstract

This dissertation takes the intersection of Machine Learning, Automatic Speech Recognition (ASR) and Hearing Aids (HA) to evaluate the effectiveness of machine learning algorithms in prioritising speech in noisy environments. The aim is to classify the environment based on the audio signal, and then apply heuristic speech enhancement techniques to prioritise speech in particular environments. The experiments are conducted on a novel dataset called HEAR-DS where we evaluate the baseline Convolutional Neural Network (CNN) classifier presented by the authors of the dataset. We then demonstrate techniques that speed up the training process of the CNN model, by using the Adam optimiser. It is also shown that the model can generalise better by applying data augmentation techniques. We then compare the performance of the CNN model to a Recurrent Neural Network (RNN) model, specifically the Long Short-Term Memory (LSTM). The results show that the LSTM model outperforms the CNN model.

Moreover, we develop a baseline speech enhancement model using a basic CNN model. We then take inspiration from the literatue and apply more layers to the model to improve the performance. We then compare the performance by comparing the STOI and PESQ scores of the models. The results show that the more complex model outperforms the baseline model.

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.

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 Aid (HA) 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 pair the proposed machine learning techniques with signal processing techniques to evaluate the effectiveness of speech prioritisation in noisy environments. The idea is to first perform acoustic scene analysis (ASA) to classify the environment based on the audio signal. Afterwards, speech enhancement techniques will be applied to the signal to prioritise speech. ASA can be done using a dataset of varying

environments, and using machine learning techniques to classify the environment.

In this project, we will be using a novel dataset proposed by Hüwel et al. [4] 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) is used to detect speech, however, the dataset helpfully contains labels which samples contain speech. 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 actually using the dataset and comparing various machine learning techniques to evaluate the effectiveness of the proposed machine learning techniques.

This project will investigate how recurrent neural networks (RNN) and their variants, such as LSTMs, can be used to classify the environment. Based on the environment classified, the system will then apply signal processing techniques to enhance the speech. The project however has to 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 [5]. 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)‘ [6] 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.

Delay constraints are also important in HA, as the user needs to hear the speech in real-time. For example, so long as the speech production is no higher than 30 milliseconds (ms), and ideally less than 20ms, the user is unaffected by the delay [7].

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.

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 [8], 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.

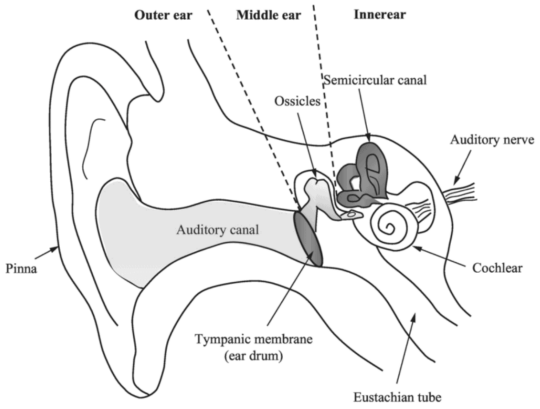

Figure 2.1: The external, middle and the inner ear from Wayland [8]

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.

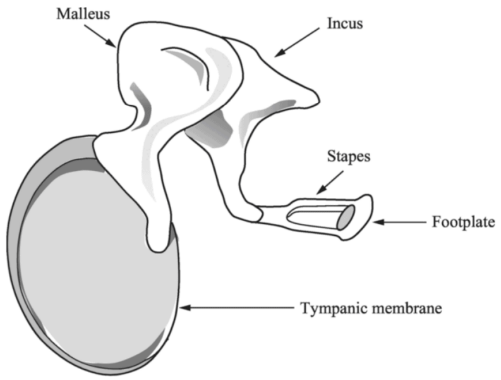

Figure 2.2: The components of the middle ear from Wayland [8]

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!

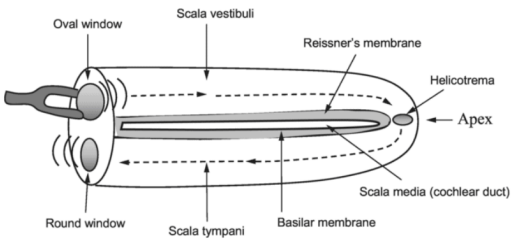

Figure 2.3: 'Unrolled' cohlear from Wayland [8]

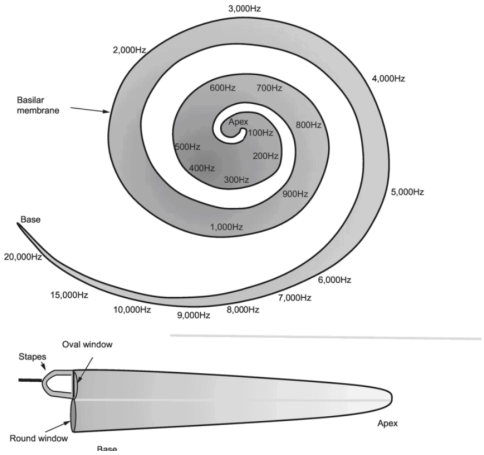

Figure 2.4: The Organ of Corti and the frequency responses. From Wayland [8]

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 [9]. Hearing loss is a spectrum, and varies from mild (26-40dB loss), moderate (41-60dB loss), severe (61-80dB loss) to profound($>$ 80dB loss) [10]. Moreover, hearing loss may occur in one ear (unilateral) or in both ears (binaural).

2.2 Hearing Loss Treatment

2.2.1 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. Another consideration is cost, the fitting of a hearing aid is a lower barrier of entry since there does not need to be any surgical operations done.

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. 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).

How do we choose these two values, you may ask? The sampling rate will vary between different speech tasks, and part of the decision is dictated by a theorem known as the Nyquist-Shannon Sampling Theorem, which states that the sampling rate should be at least twice the highest frequency range present in the signal. In other words, if your speech processing task involves working with frequencies of 8KHz, then the sampling rate should be at least 16KHz. As for the selection of an appropriate bit depth value, this depends on the need of representing a wide dynamic range of air pressure in a speech signal. Typically, it is common to use 16 bits as it can represent an amplitude range of 96dB.

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.

TODO: Will write this in more detail once I cover this in the Speech Processing course (Week 7-8).

2.4 Related Work

As this project will chain two algorithms—Acoustic Speech Analysis (ASA) and Speech Prioritization, which is also referred to as Blind Source Separation in some literature—we will cover each in separate sections. We will discuss the work that has been done and my plans for further development.

2.4.1 Acoustic Scene Analysis (ASA)

Acoustic Scene Analysis is the process of classifying the scene (environment) from an audio stream. For this project we will be using a supervised machine learning

model for ASA. So, we need a audio dataset that contains useful metadata on the recording such as the environment it is in. This project aims to use the HEAR-DS dataset presented by Hüwel et al. [4]. According to the researchers, the dataset came to fruition because existing acoustic scene classification databases are inappropriate for HA processing. We will go into more detail of the dataset in Chapter X [TODO: reference correclty once done]. For now it is enough to know that the dataset contains background samples But it is enough to know that the dataset is classified in an environment of which it is subcategorised as either containing speech only, background only or speech & background (both). To prevent machine learning models from overfitting and to increase the diversity of situations, each environment enforces a minimum of three recording situations (REC-SITs) i.e. there must be at least three different locations for the recordings. The researchers presented the applicability of this dataset by performing a series of classification experiments using various Convolutional Neural Networks (CNN) of differing sizes. This is where the paper could be more clear on the design decisions of the CNN, esepcially because the accuracy of the classification of the models can vary as high as 20%. This project hopes to reproduce the results, and explain the high variability in the accuracy and how those can be circumvented, because the high variability suggests that the random initialisation of weights during training is the primary factor influencing the model performance. [To Hao: This is what I understood from your interpretation in the discussion of this paper last Friday, did I understand you correctly?].

2.4.2 Speech Prioritisation

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] 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.

[5] 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

[6] 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.

[7] 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.

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

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

[10] 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.