Investigating Machine Learning Techniques for Hearing Aids

Nikodem Bieniek

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 solemnly 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.

1.2 Structure

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.
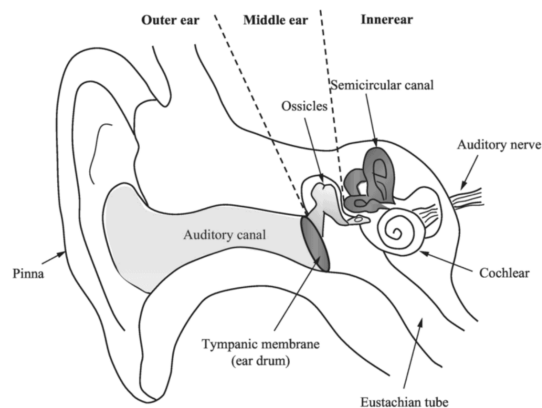

Figure 2.1: The external, middle and the inner ear from Wayland [8]

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.
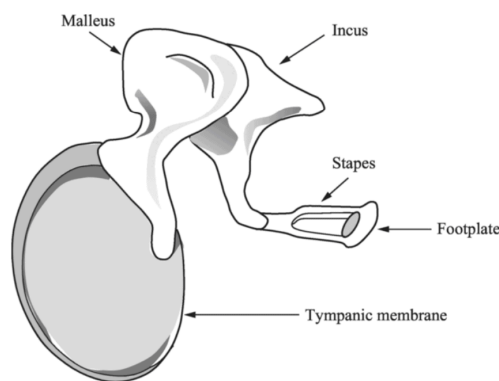

Figure 2.2: The components of the middle ear from Wayland [8]

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%

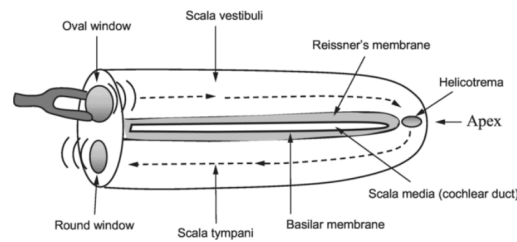[2]'Cohlea' in Latin translates to snail shell!
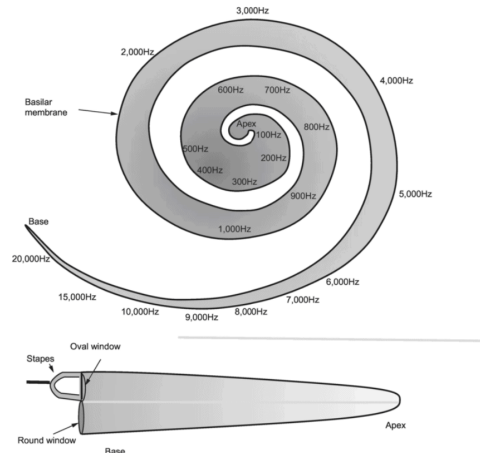
Figure 2.3: 'Unrolled' cohlear from Wayland [8]

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 History of Hearing Technology

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. 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.

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 [5] 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. 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. 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. The other reason is that it is much more clearer to distinguish between speech and background noise from this representation than from the waveform, so it is easier for the model to learn.

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. As such, we need a sufficiently large dataset that contains useful metadata on the recordings 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 3.1. The paper where the dataset was first introduced, presents the applicability of the dataset by showcasing the use of a Convolutional Neural Network (CNN) to classify the environment. In Chapter 4, we will be attempting to reproduce the results of the paper. As you will see, I struggled to reproduce the results of the paper and I discuss my findings in the chapter too. So as a validation of the model implementation, I decided to see if the same model parameters can be used to classify the DCASE 2017 Acoustic Scenes Challenge dataset [11]. More specifically, I compare Schindler et al. [12] model to the CNN model used in the paper. I discuss my findings in **??**.

2.4.2 Speech Enhancement

2.5 Criticism of Previous Work

During the course of this project, I found out that the paper by Hüwel et al. [4] 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 [13] and CHiMe 5 [14] 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 4. We still thank the authors for their work on the dataset as it is a promising dataset for the future of HA research.

Chapter 3

Experimental Setup

3.1 Datasets

3.1.1 HEAR-DS

HEAR-DS is a dataset created by Hüwel et al. [4] 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) 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, their recordings 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 [15] 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 [14]. 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 [14] 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 3.1.3.

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 4). 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.

The model presented in the paper assumed a sampling rate of 16KHz, so we resampled the HEAR-DS dataset to 16KHz. Additionally, the model expected a log-mel spectrogram of 10s segments, and the HEAR-DS dataset provided longer segments. We therefore split the HEAR-DS segments into 10s segments by traversing the dataset and extracting 10s segments. There is potential for this to have discontinuities in the beginning and end of the segments, but from the paper it is unclear how this was addressed if at all. We went with a naive approach of just cutting the segments.

3.1.2 CHiME3

As mentioned in the previous chapter, 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.

3.1.3 CHiME6

The CHiME 6 dataset is a newer dataset that supersedes the CHiME 5 dataset. It contains 100 hours TODO

3.1.4 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. [4] so as a means of validation, I decided to use the TUT Acoustic Scenes 2016 dataset [16] to test the performance of the model. In particular, I will comparing the results by Schindler et al. [12].

3.2 Baseline Models

3.3 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. [4] 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 4

Acoustic Scene Classification Experiments

4.1 Data Preparation

TODO

4.1.1 Fixed Learning Rate SGD Approach

TODO

4.1.2 Adam Optimiser Approach

TODO

4.1.3 LSTM Model

TODO

Chapter 5

Speech Enhancement Experiments

5.1   Data Preparation

TODO

5.1.1   CNN Model

TODO

Chapter 6

Conclusions

6.1 Discussion

TODO

6.2 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] 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/tech

[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.

[11] 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.

[12] 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].

[13] 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`.

[14] 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.

[15] 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.

[16] 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.