Machine Learning in Audiology: Applications and Implications

By

François Charih

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

Master of Applied Science
in Electrical and Computer Engineering
(with Data Science Specialization)

Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario
August, 2018

À mon grand-papa, Jean-Guy Deslongchamps.
Ta curiosité est pérenne; elle te survit en moi.

Abstract

Recent mobile and automated audiometry technologies have allowed for the democratization of hearing healthcare and enables non-experts to deliver hearing tests. The problem remains that a large number of such users are not trained to interpret audiograms. In this work, we outline the development of an intelligent audiogram classification system. More specifically, we present how a training dataset was collected, the development of the classification system relying on supervised learning, as well as other tools designed for the analysis of audiograms in large databases. Using a dedicated annotation tool developed specifically for this study, the Rapid Audiogram Annotation Environment, we collected hundreds of audiogram annotations from three licensed audiologists. Our analysis demonstrates that inter-rater reliability is substantial or better for classification of hearing loss configuration, symmetry, and severity, in spite of the subjective nature of the classification task. Furthermore, our results suggest that the agreement nonexistent for the identification of audiometric notches or potentially unreliable thresholds. The system proposed here achieves a performance comparable to the state of the art, but is significantly more flexible. Finally, we demonstrate qualitatively that a method based on density estimation with Gaussian mixture models is useful for the detection of potential reliability issues in audiograms.

Résumé

De nouvelles technologies d'audiométrie mobile et automatisée démocratisent l'accès aux tests auditifs et permettent à des non-experts d'administrer des tests auditifs. Reste-t-il qu'une grande partie des utilisateurs ne sont pas formés dans l'interprétation d'audiogrammes. Dans ce travail, nous présentons l'élaboration d'un système de classification d'audiogrammes intelligent. Plus particulièrement, ce travail détaille la collection de données d'entraînement, le dévelopement d'un système de classification intelligent basé sur des techniques d'apprentissage supervisé, ainsi que le dévelopement d'outils pour l'analyse rétrospective d'audiogrammes dans de grandes banques de données. À l'aide d'un outil d'annotation développé spécifiquement pour cette étude, soit le Rapid Audiogram Annotation Environment, trois audiologistes professionnels ont annoté plusieurs centaines d'audiogrammes. Notre analyse démontre que la fiabilité entre évaluateurs et d'un même évaluateur est au moins substantielle pour la classification de perte auditive par configuration, symmétrie et sévérité, et ce, malgré la nature subjective de la classification d'audiogrammes. Nos résultats indiquent une absence de fiabilité entre évaluateurs pour l'identification d'encoches audiométriques ou de problèmes de fiabilité potentiels dans l'audiogramme. L'exactitude de la classification du système proposé est similaire à l'état de l'art, mais est plus flexible. Finalement, nous démontrons de façon qualitative qu'une méthode basée sur la modélisation de la densité par mélange de gaussiennes est utile pour la détection de problèmes de fiabilité potentiels dans des audiogrammes.

Acknowledgements

First and foremost, I would like to express my gratitude to my research advisor, Prof. James Green, for his guidance, his support during the difficult times, and for his unwavering patience. I thank him for continuously pushing me to take an extra step out of my comfort zone so that I can see a little farther and a little clearer. I very much look forward to working with him as a doctoral student.

Special thanks to the Clearwater Clinical Limited staff who provided expertise and helped in shaping the research in this thesis: Dr. Matthew Bromwich, Dr. Amy Fraser, Ms. Renée Lefrançois, Mr. Siegurd Weber and Mr. Julian Bromwich.

I would also like to thank my fellow lab mates, especially Kevin Dick and Yasmina Souley Dosso, for the brainstorming sessions, the late-night dinners at *Ollie's*, the moral support during the hard times, and their listening ear. They truly are like family to me. Thank you, Pratyush Singh and Ashlynn Steeves, for your invaluable contributions to this project, but also for helping me to become a better mentor. Thank you, Mr. T. Hortons and Mr. S. Cup; your magic touch helped me power through some very difficult days.

I am grateful towards the Natural Sciences and Engineering Research Council of Canada, the Ontario Centres of Excellence and Clearwater Clinical Limited who financially supported this project.

Finally, I would like to thank you, mom and dad. You inspire me in everything that I do.

Statement of Originality

The work presented here was completed by the author, under the supervision of Prof. James R. Green. The research was completed at Carleton University as part of the requirements for the degree Master of Applied Science in Electrical and Computer Engineering with specialization in Data Science.

Portions of the work presented in Chapters 3 and 4 were published in the proceedings of the IEEE Medical Measurements and Applications (MeMeA) conference held in Rome, Italy in June 2018:

> François Charih, Matthew Bromwich, Renée Lefrançois, Amy E. Mark, and James R. Green. Mining Audiograms to Improve the Interpretability of Automated Audiometry Measurements. In *Proceedings of the 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, Rome, Italy, 2018

Another conference publication resulted from the work done as part of this thesis. The work on missing audiometric data imputation and quality assurance presented in Chapter 6 was presented at the IEEE Life Science Conference (LSC) held in Montréal, Québec in October 2018:

> François Charih, Ashlynn Steeves, Matthew Bromwich, Amy E. Mark, Renée Lefrançois, and James R. Green. Applications of Machine Learning Methods in Retrospective Studies on Hearing. In *Proceedings of the IEEE Life Sciences Conference 2018*, Montréal, Canada, 2018

Finally, the work presented in Chapter 5 will form the basis of a journal paper, and a patent disclosure, both of which are currently in preparation.

Contents

List of Tables

List of Figures

List of Abbreviations

ASHA	American Speech-Language-Hearing Association
AWS	Amazon Web Services
BIC	Bayesian information criterion
CDC	Center for Disease Control
dB	decibel
dB HL	decibel hearing level
dB SPL	decibel sound pressure level
DDAE	Data-Driven Annotation Engine
EHF	extended high frequencies
GMM	Gaussian mixture model
k-NN	*k*-Nearest Neighbors
NHANES	National Health and Nutrition Examination Survey
NIHL	noise-induced hearing loss
PTA	pure tone average
RAAE	Rapid Audiogram Annotation Environment
RDS	Relational Database Service
UI	user interface
WHO	World Health Organization

1

Introduction

Scientific research is one of the most exciting and rewarding of occupations. It is like a voyage of discovery into unknown lands, seeking not for new territory but for new knowledge. It should appeal to those with a good sense of adventure.
-- Frederick Sanger, *Nobel Banquet Speech, 1980*

1.1 Introduction

Hearing loss is an often underrated and overlooked condition that global health authorities such as the World Health Organization (WHO) have called a "silent epidemic" [3]. As a matter of fact, it is estimated that 350 million people cope with some form of hearing disorder worldwide [4]. According to recent projections, this figure could reach 630 million by 2030 [5]. The consequences of this condition are functional, social and economic in nature [6].

Individuals affected by hearing loss cope with a variety of personal issues. It is widely acknowledged that hearing disorders in children can impair language acquisition [7, 8, 9] as well as academic performance, which can, in turn, reduce earning potential [6]. Quality of life can be severely impaired in adults as well. In fact, adults living with untreated hearing loss are at higher risk of experiencing isolation and limiting social activities, leading to heightened feelings of isolation and depression [10]. In addition, adults with hearing loss have higher odds of being underemployed or unemployed [11].

Hearing loss also has implications for families, friends and co-workers, among others,

and as such, is experienced socially. Families and friends often experience frustration as a result of the difficulties they face in communicating with the affected individual and report increased dependence [12]. It has also been shown that noise-induced hearing loss is associated with higher risks of accidents in the workplace [13].

On a global scale, hearing loss is associated with costs estimated US$750B per year to cover costs related to health care, education and productivity losses [6]. This figure is roughly equivalent to the 2016 nominal GDP of the Netherlands [14].

Surprisingly few people with hearing loss are aware of their condition. For example, about 77% percent of Canadian adults with some form of hearing loss were undiagnosed between 2012 and 2015 [15]. In the U.S., only 20% of individuals with hearing loss seek treatment and the average wait time between diagnosis and treatment is 10 years [16]. This is due partly to the insidious nature of the condition, but also to the limited availability of hearing healthcare providers in developing and developed countries alike. Swanepoel *et al.* eloquently pointed out the lack of alignment between the need and availability of hearing healthcare [3]:

> "Nowhere is the irony of global inequality more striking than in hearing health care, with more than 80% of people with hearing loss residing in developing countries where services are either totally absent or very limited."

Goulios and Patuzzi found, in a comprehensive study [17], that this shortage of qualified professionals stems from various factors including insufficient government funding, prioritization of other health issues, lack of awareness from the broader public, and deficiencies in audiology training – particularly the absence of audiology programs and clinical placements. This problem plagues not only developing countries, but also developed countries [18].

In face of the net increase of demand for hearing healthcare specialists caused by the aging population, the U.S. President's Council of Advisors on Science and Technology has become a proponent of the "direct-to-consumer approach" [18]. In other words,

it proposes that consumers forego a medical evaluation and, guided by a hearing aid retailer, acquire hearing aids. Clearly, this approach can be problematic, as in some instances, the cause of hearing loss will be serious and require a medical intervention. As such, there is a need for technologies that will enable individuals to get their hearing tested, and to determine whether their condition calls for a referral to a physician.

Fortunately, modern technologies and frameworks, namely mobile health and telemedicine, gave birth to a new era where hearing tests can be inexpensively delivered in areas where audiological services are limited. Mobile phones and tablets are now used for disease management, counselling, coaching, and even diagnostics, ultimately helping millions worldwide to improve their health at costs that are considerably lower than was previously possible [19]. Audiological care did not escape this trend, and mobile audiometers eventually made their way to the market. Today, a handful of companies developed tools capable of delivering hearing tests using minimal equipment, among which Clearwater Clinical Limited.

Clearwater Clinical Limited is an Ottawa-based company specialized in medical instrumentation and markets the SHOEBOX Audiometry™ system. The system, which we shall henceforth refer to as the SHOEBOX, consists of an Apple iPad tablet and a pair of commercially available headphones specifically calibrated for use with the tablet. The SHOEBOX is used to perform pure-tone audiometry outside of a sound booth, allowing users to deliver hearing tests in quiet clinics, schools, hospital, homes and offices. Furthermore, the system exists in two distinct versions: a standard version and a pro version. The full-featured pro version was designed specifically for audiology experts, while the standard version was designed for used by non-expert healthcare providers such as primary physicians, nurses, occupational health technicians, etc. Both versions can deliver automated hearing tests by means of an intuitive user interface (Figure 1.1) that requires the user to indicate if the presented tones are heard.

The SHOEBOX is currently deployed worldwide by more the 1,500 users, in and

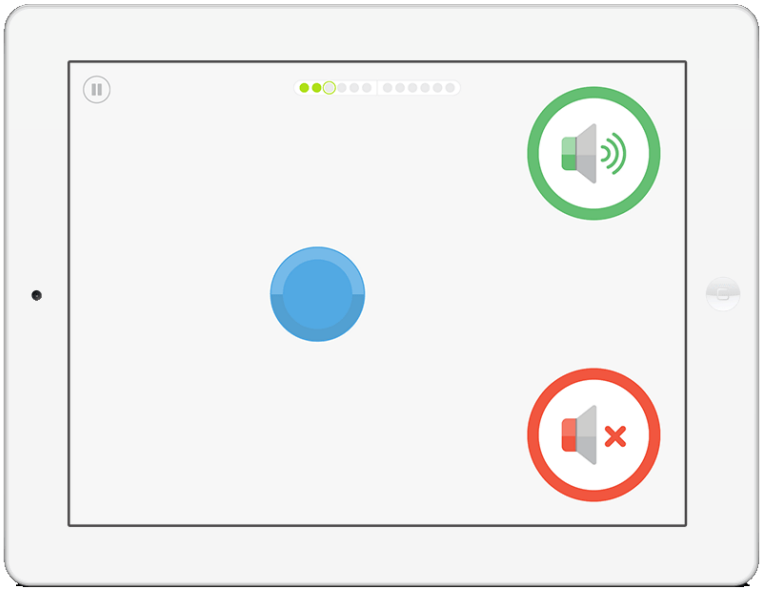


Figure 1.1: The SHOEBOX Audiometry™ automated test interface
The user is instructed to drag and drop the blue puck into the green circle if the sound is heard or into the red circle otherwise.

outside of the clinic. Users include researchers on the prevalence of hearing loss and ototoxicity, among others. Furthermore, the system has been used to deliver hearing tests to thousands of individuals in developing countries. While the SHOEBOX is now in use all around the world, the problem remains that many users lack the training necessary to properly interpret an audiogram (Figure 1.2), the output generated by the pure-tone audiometry procedure. Ultimately, non-expert users would benefit tremendously from a clinical decision support system capable of providing a differential diagnosis and recommending the optimal treatment or referral options.

1.2 Rationale

A first step towards this goal of building a fully automated diagnostic tool involves developing a classification system capable of summarizing the audiogram using a standardized set of descriptors such that it can easily be understood by expert and non-expert users. These descriptors would summarize the audiogram using terminology related to its

Figure 1.2: A typical audiogram output by SHOEBOX Audiometry™
The SHOEBOX audiometer produces a standard audiogram plotting the hearing thresholds as a function of frequency.

shape, symmetry, severity and reliability. Expert users could benefit from such a system, letting the automated classification system annotate the audiogram, allowing them to devote more time to engage in discussions with the patient about their condition and treatment options.

Machine learning and data-driven approaches have had much success in a wide variety of applications in healthcare. For example, these approaches have been used to diagnose diabetic retinopathy from retinal images [20], to predict response of cancer cells to drugs [21], and to predict the outcome of cardiovascular events [22], just to name a few. This serves as evidence that supervised learning, the variant of machine learning concerned with classification and regression with models trained using labeled data, can potentially be applied to the field of audiology. More specifically, supervised learning could be used to classify audiograms such that they can be easily be interpreted by non-experts and to help the individual delivering the test in identifying an adequate follow-up procedure if necessary. At the time of writing, machine learning has not been used for these purposes, although it has found other usages in audiology [23, 24, 25, 26].

One group has developed a rule-based system for audiogram classification [27], but this system is rigid and complex. As evidenced by the current shift of clinical support systems towards data-driven approaches made possible by machine learning, there is significant value in investigating the potential of this technology for audiogram classification.

1.3 Problem Statement

In this work, we mainly seek an answer to the following research questions:

1. What is the extent of the agreement between audiologists when annotating audiograms, an intrinsically subjective process? Are the audiologists self-consistent?

2. Can supervised learning be applied to audiogram classification? How does the

performance of such a system compare with a traditional rule-based system?

3. Is it possible to estimate the reliability of an audiogram using only the audiogram itself, and no additional sources of data?

4. Can data in large audiogram databases be leveraged to impute missing values in incomplete audiograms?

1.4 Thesis Organization

This thesis present a cohesive series of steps leading ultimately to an answer to each of the questions formulated in the previous section, roughly in chronological order. The organization of this thesis is illustrated in Figure 1.3.

Chapter 2 will provide the context and background necessary for the reader to understand and appreciate this scientific endeavour. Fundamental topics in audiology will be introduced. A treatment of audiogram classification and a discussion of the state of the art will follow. Finally, the mathematical tools used throughout this thesis are succinctly described.

Equipped with the necessary knowledge of audiology and the context in which this research project is situated, we shall discuss how we developed the Rapid Audiogram Annotation Environment (RAAE), a web-based software to faciliate audiogram annotation (Chapter 3), and assembled a high-quality dataset of audiogram annotations (Chapter 4). More specifically, we shall describe how audiograms were selected, sampled and annotated by expert audiologists before proceeding to an in-depth analysis of these data.

In Chapter 5, the core of this thesis, we present the methodology used to build the Data-Driven Annotation Engine (DDAE), an automated audiogram classification system, using the training set assembled in Chapter 4. We assess the performance of the system and compare it with the current state of the art.

In Chapter 6, we present additional applications of machine learning in audiology. In particular, we present the development of an approach suitable for the identification of problematic audiograms. This approach can be used to flag suspicious or unusual audiograms in large databases or to provide real-time feedback during testing. In addition, we present a novel data-driven method for missing threshold imputation in audiograms.

Ultimately, this thesis culminates in Chapter 7 wherein the contributions made as part of this thesis are summarized in light of the findings and avenues for future research are provided.

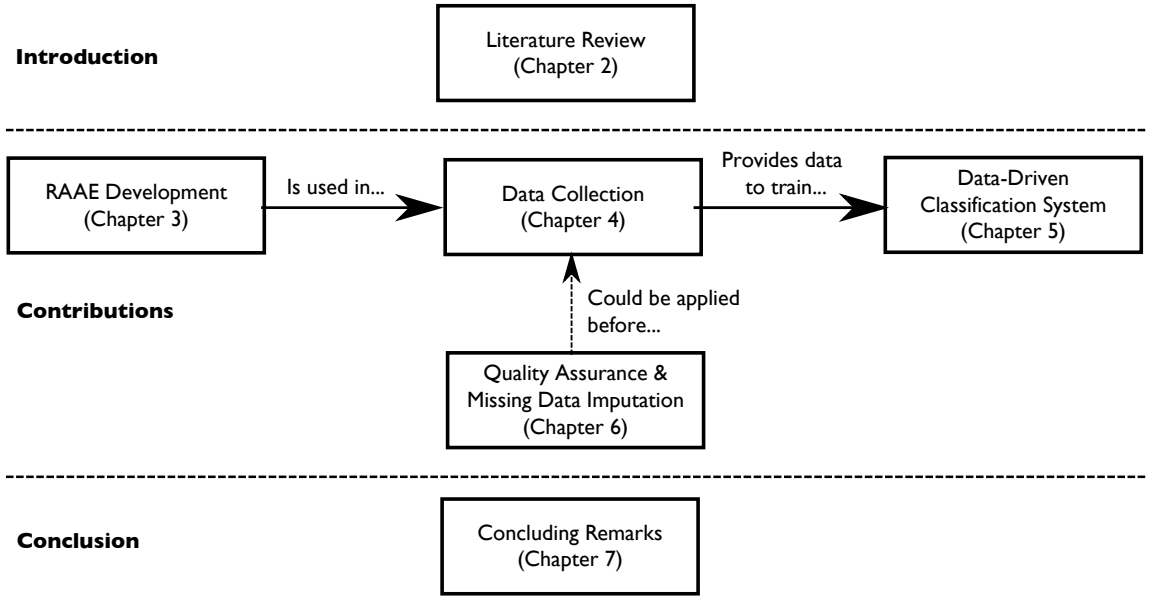


Figure 1.3: Logical organization of this thesis
This thesis comprises 6 chapters, excluding the current chapter. The relationship between these 6 chapters is illustrated here.

2

Background

It takes considerable knowledge just to realize the extent of your own ignorance.
-- Thomas Sowell

In this chapter, the reader will be introduced to fundamental concepts in audiology. It is assumed that the reader has limited familiarity with the field, and that laying out these concepts will help position this work within the audiological research landscape.

First, we introduce the audiogram and discuss recent advances in audiometry, in particular mobile and automated audiometry. A treatment of audiogram classification and the state of the art in the field follows. Finally, the mathematical tools upon which this work builds are described succinctly.

2.1 The Audiogram

Audiologists are trained to deliver a large range of tests to assess the hearing of a patient. Most people will be familiar with otoscopy, wherein an otoscope is inserted in the ear canal to look for obstruction by ear wax, and inspect the tympanic membrane, because it is often conducted as part of a routine medical examination. Fewer people are familiar with pure tone audiometry, the gold standard procedure for measuring a subject's hearing acuity. In pure tone audiometry, the subject is presented with pure

tones of fixed frequency and amplitude. The minimum amplitude required to elicit a response at a given frequency at least half the time, termed a *threshold*, is recorded and plotted on an *audiogram* [28]. The reader is invited to read Appendix B for a more thorough treatment of pure tone audiometry.

The audiogram plots the threshold of hearing of the patient as a function of frequency, as illustrated in Figure 2.1.

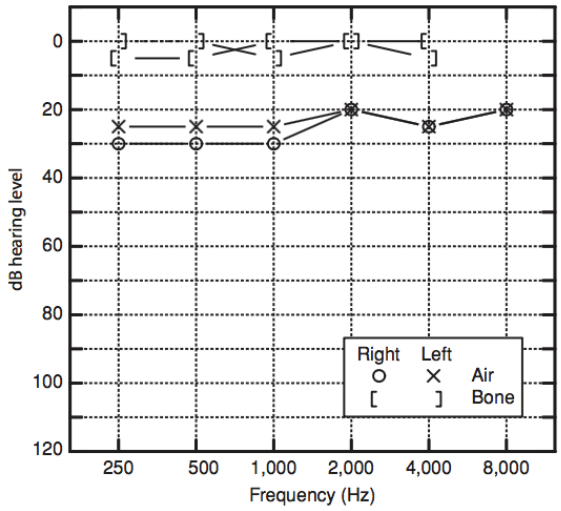


Figure 2.1: A standard pure tone audiogram
This figure (reproduced from [28]) represents a complete standard audiogram. Audiograms symbols follow a set of conventions, indicating to the type of measurement made to determine the threshold. Air conduction thresholds are denoted by circles and crosses for the right and left ear, respectively. Bone conduction thresholds (with masking) are represented with square brackets. Notice the logarithmic scale along the x-axis and the inverted linear decibel scale along the y-axis.

While the audiogram itself only provides a partial picture of the patients hearing health, it provides invaluable clues that can guide the rest of the examination. A complete audiogram can be used to determine whether hearing loss is present, the site of the lesion, and, in conjunction with other tests, the underlying cause, or *etiology*, of the loss.

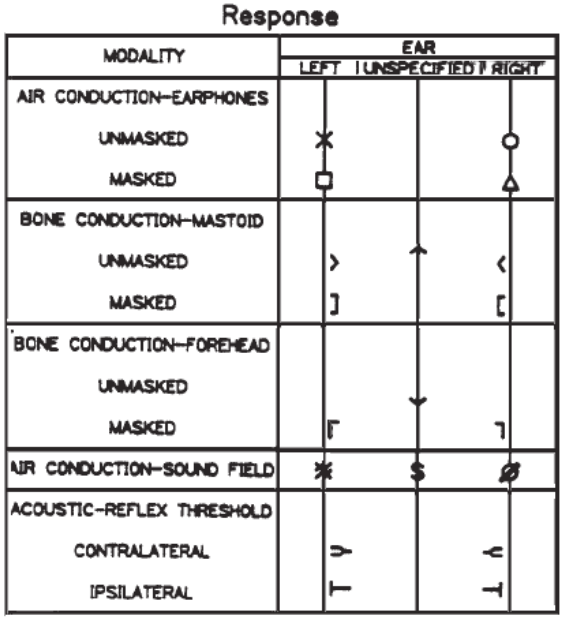
Many audiologists follow the ANSI S3.21-1978 standard for plotting audiograms [28]. For historical reasons [29] that are beyond the scope of this thesis, the vertical axis of

the audiogram is intentionally inverted. One should remember, however, that higher threshold values are worse, as they imply that sounds of higher amplitudes must be presented for the patient to respond. Therefore, the lower the threshold is on the audiogram (the higher the value), the worse the hearing at that frequency.

The American Speech-Language-Hearing Association (ASHA) recommends different symbols (listed in Table 2.1) which specify, among others, the ear (left or right), whether masking (see Section B.5) was applied or not, whether the measurement could be completed within the audiometers limits, and how the stimulus was presented: through air or through bone.

Table 2.1: Audiogram symbols recommended by ASHA

This table (reproduced from [30]) provides a non-exhaustive list of recommended symbols for use in clinical audiograms. Mobile audiometers such as SHOEBOX mostly produce measurements whose symbols are presented in the two upper sections of this table, namely air conduction thresholds with headphones and bone conduction thresholds through the mastoid bone.



The units along the vertical axis are in decibel hearing level (dB HL). This unit of measurement is related to the more common decibel sound pressure level (dB SPL). The dB HL is a normalized version of the dB SPL, and corresponds to the difference

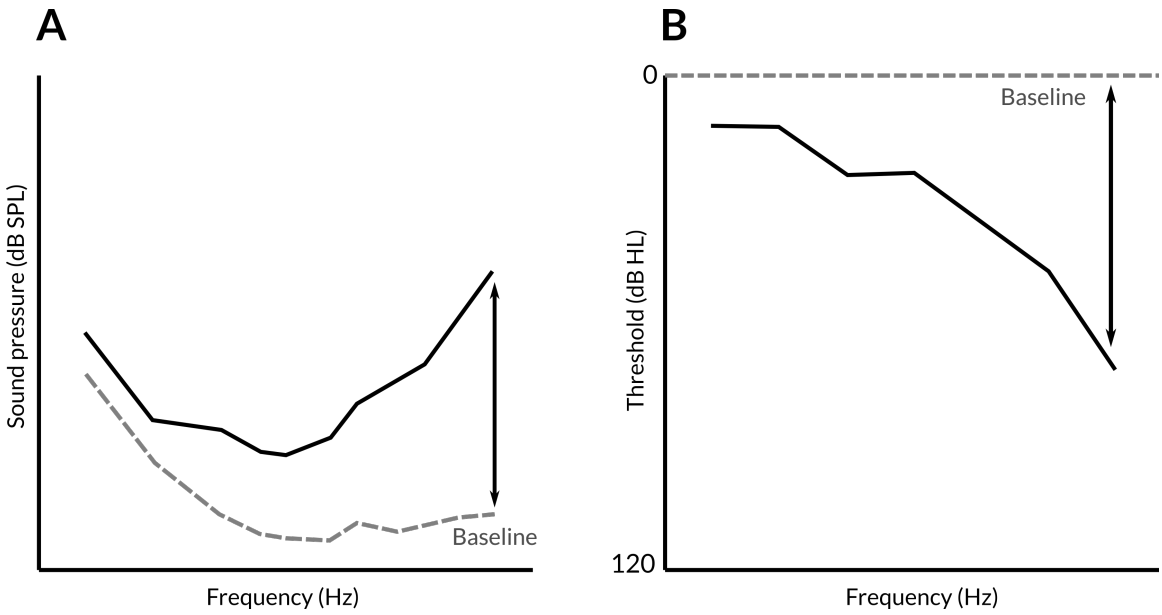


Figure 2.2: Definition of the dB hearing level (dB HL)
This figure (adapted from [28]) illustrates the same fictitious audiogram plotted using a dB SPL scale (A) and the more common dB HL scale (B) which is typically used to plot audiograms. The double-headed arrows represent the same distance.

between the threshold of hearing of the patient and the "median threshold of a young adult with no history of ear problems" [28].

2.2 Mobile and Automated Audiometry

There is an important shortage of professionals with the training necessary to diagnose and treat hearing disorders [31, 17, 3, 32]. This shortage of audiology expertise is illustrated in Figure 2.3. Windmill and Freeman determined that to keep up with the demand in the U.S., the number of audiologists trained would have to increase by 50% and the attrition rate would need to be reduced by at least 20% [32]. Not only is the raw number of trained professionals an alarming, their geographical distribution is largely biased as they tend to work in large centers, leaving rural communities underserved [3].

Fortunately, modern developments aligned with the new paradigm centered around

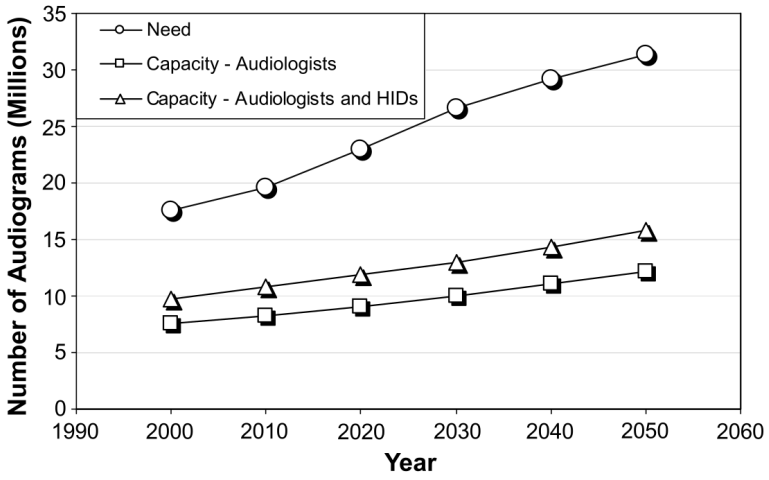


Figure 2.3: Projected gap between the need and supply for audiological expertise
This plot (reproduced from [31]) clearly depicts the growing gap between the supply and demand for audiological expertise and makes a solid case for the development of automated audiometry.

mobile health (mHealth) and telemedicine, have led to the development of new technologies that allow for the delivery of hearing tests with a setup that is orders of magnitude lower than the conventional setup consisting of a large -- and quite expensive -- audiometer and a soundproof booth.

The telemedicine framework divides into two distinct models: *synchronous* and *asynchronous* [3]. The synchronous model facilitates remotes interactions between the healthcare provider and the patient, often through specialized software that transmits test results over the internet in near real-time. In this framework, immediate interactions between the healthcare provider and the patient are possible, and there may or may not be a trained facilitator on-site. In contrast, asynchronous protocols tend to have a higher degree of automation and to only transmit results to the healthcare provider in a *post-hoc* fashion, either over the internet or by phone, so that a professional can interpret the results. In this asynchronous model, a facilitator is typically on site to ensure that the measurement procedure is correctly executed. It goes without saying

that the asynchronous model is more scalable, as results may only be transmitted from the non-expert facilitator to the provider, a physician or audiologist, for example, if a follow-up is indicated.

The asynchronous model has gained a lot of traction for automated pure tone audiometry. A handful of mobile devices are now deployed and used by experts and non-experts to deliver hearing tests with minimal supervision. Diagnostic-grade mobile systems currently on the market include KUDUwave™ (Emoyo, Johannesburg, South Africa) [33], GSI AMTAS™ (GSI Grason-Stadler, Eden Prairie, U.S.A.) [34] and SHOEBOX Audiometry™ (Clearwater Clinical Limited, Ottawa, Canada) [35]. The most important advantage associated with these devices is that they can be used outside of a sound proof booth and still produce accurate measurements. Thanks to their minimal form factor, these tablet-based audiometers are very portable, making them deployable in a variety of environments from the hospital to clinics, schools and other moderately quiet environments in urban and remote environments.

Multiple studies have concluded that the quality of the audiograms generated by these devices is similar to that of audiograms generated in a soundproof booth by professionals. In fact, Thompson *et al.* [35] determined the error to be below 10 dB for 97% of measurements made with the SHOEBOX audiometer in ''moderately noisy environments'' such as consultation or waiting rooms. Saliba *et al.* [36] determined, using a setup combining a noisy soundbooth and noise cancellation (active and passive), that self-administered tests generated measurements within 10 dB of the measurement obtained through conventional audiometry 91% of the time. They also found that a patients tended to prefer mobile audiometry to conventional audiometry.

Mobile audiometers such as SHOEBOX Audiometry™ and GSI AMCLASS™ integrate varying degrees of automation, allowing users with minimal expertise to administer the pure tone audiometry procedure. The SHOEBOX audiometer for example, provides instructions for the user, administers the test using a modified Hughson-Westlake

threshold search algorithm, can detect potentially unreliable measurement conditions due to excessive ambiant noise, and generate an audiogram. Automated audiometry reduces reliance on trained professionals to administer the test and thus increases access to hearing screening, as the test can be administered by non-experts such as family physicians, hearing instrumentation specialists, nurses, occupational health professionals, etc.

Automated audiometry is now deployed in a variety of environments. Tools like the SHOEBOX audiometer are used in clinicals and hospitals, but also in remote areas for humanitarian projects [37] and for research [38, 39, 40]. Many studies have resorted to automated audiometry to study the prevalence of hearing loss. For example, the SHOEBOX audiometer has been used to study the prevalence of hearing loss in children living in Northern Canada [38], HIV-positive children in Uganda [39], and in singing teachers and students [41] for example.

So far, the intention was to build a solid case in favour of mobile and automated audiometry. The reader may still wonder, however, how exactly this relates to the main objective of this thesis, namely to develop an intelligent audiogram classification system. This will become clear in the upcoming section.

2.3 Classification of Hearing Loss

It is common practice for audiologists and for researchers in the field of audiology to assign descriptors to audiograms. Audiologists use descriptors to summarize different aspects of the hearing loss. The different components making up an audiogram annotation are presented in Table 2.2.

One should appreciate the fact that a given audiogram may be summarized using different words by two audiologists. In fact, the specific words employed to summarize a given aspect of the audiogram often depends on the training received by the audiologist

Table 2.2: Descriptors used to summarize the hearing loss

Audiologists use different types of descriptors to succinctly summarize different aspects of the hearing loss.

Descriptor type	Meaning	Descriptor values
Symmetry	Whether both ear are equally impaired	Symmetrical or asymmetrical
Severity	The extent of the hearing loss	Normal, mild, profound, ...
Configuration	The shape of the audiogram	Flat, sloping, rising, cookie bite, ...
Site of lesion	The nature of the hearing loss	Sensorineural, conductive, mixed, ...

reviewing the audiogram, and may vary from one clinic to another. These inconsistencies are present throughout audiological literature, particularly with respect to audiogram configuration.

The terminology used to refer to this practice varies. The terms "audiogram classification" and "audiogram annotation" are used interchangeably throughout this thesis.

2.3.1 Rationale

The main function of these descriptors is to distil the essence of the audiogram down to a small set of simple terms. Audiogram classification is done so that the hearing loss can be summarized to non-expert patients or other professionals, in addition to facilitating the identification of the etiology behind the hearing loss [28].

Audiogram classification and interpretation requires expertise, which, as we have seen in Section 2.2, is becoming scarcer every year. Consequently, there exists a need for methods capable of automatically describing or summarizing audiograms in terms that are easily interpretable by non-expert users of automated audiometry systems. Such support decision systems can help the tester in determining whether a referral is in order.

This practice is also highly important in audiological research, and this application is far from new. Raymond Carhart, widely acknowledged as the father of modern audiology,

expressed its importance in audiology research as early as 1945:

> "Whenever large numbers of pure tone audiograms are to be grouped for study and analysis, it is essential that be classified in some simple and definite manner." [42]

Researchers in the field of audiology often classify audiograms with descriptors in order to study the prevalence of different hearing losses. This knowledge can then be used by governments and public health authorities to influence policy.

2.3.2 Audiogram Descriptors

Symmetry

Symmetry is arguably one of the most revealing aspects of hearing loss. Highly asymmetrical audiogram are often seen in patients with acoustic tumors, and therefore, it is a key factor in determining whether patient should be referred for medical evaluation [43]. Hearing loss is considered to be symmetrical if the thresholds are roughly the equal bilaterally [28].

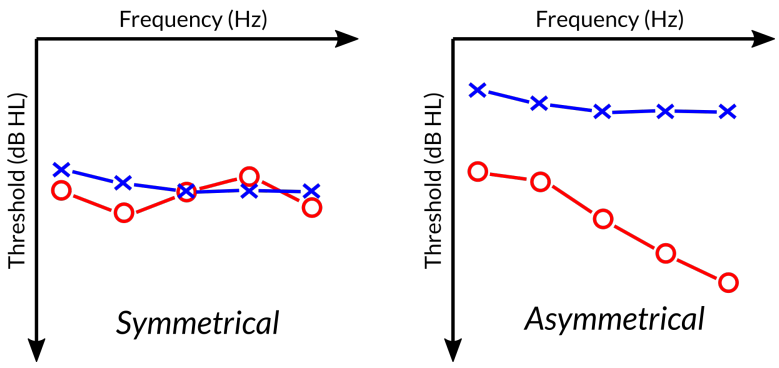


Figure 2.4: Audiogram symmetry
This figure presents two audiograms where the symmetry is unambiguous. The audiogram on the left would be considered symmetrical, while that on the right would be considered asymmetrical.

Configuration

The configuration of the audiogram can be useful in establishing a differential diagnosis for the hearing loss. For example, a sloping audiogram in older patients is likely to awaken suspicions of age-related hearing loss. Audiograms with a notch between 3,000 and 6,000 Hz, in contrast, are often observed in patients with noise-induced hearing loss (NIHL).

Although the specific words -- which we shall occasionally refer to as *labels* -- used to describe configuration vary from practice to practice, in this work, the audiogram configurations descriptors used throughout this thesis are those used at Clearwater Clinical Ltd. These configurations are shown in Figure 2.5.

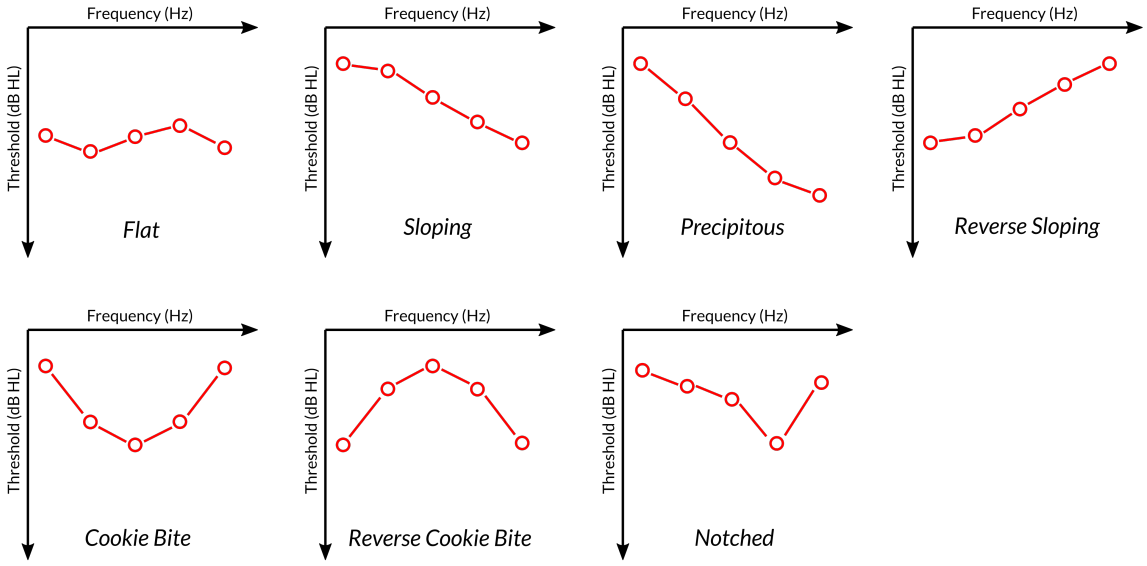


Figure 2.5: Common audiogram configurations
Different descriptors are used to describe the configuration (shape) of the audiogram. This figure depicts ideal cases of the configurations used in this thesis.

Severity

Many professionals use the pure tone average (PTA) to summarize the extent of the loss. In this method, the average of a subset of three or four frequencies, often selected from 500 Hz, 1,000 Hz, 2,000 Hz, and 4,000 Hz is computed and the severity label(s) are determined through a reference table lookup. Table 2.3 presents reference values for three severity classification systems.

Table 2.3: Reference values for severity description

This table (reproduced from [28]) presents reference values, in dB HL, from three systems for PTAs computed from 500 Hz, 1,000 HZ and 2,000 Hz thresholds. This thesis uses Goodman's reference values.

Degree of Loss	Northern and Downs [44]	Goodman [45]	Jerger and Jerger [46]
Within normal limits	<16	<26	<21
Slight	16-25		
Mild	26-30	26-40	21-40
Moderate	30-50	41-55	41-60
Moderately severe		56-70	
Severe	51-70	71-90	61-80
Profound	>70	>90	>80

Often, however, more than one severity descriptor may be necessary, as using a single descriptor may fail to capture the severity over the entire frequency spectrum. For example, a flat audiogram and a sloping audiogram may have identical pure tone averages, but significantly different hearing acuities in extreme frequencies beyond those used to compute the PTA. As a result, the severity for a sloping audiogram will typically be described using a descriptor for the low frequencies (best hearing for a sloping loss) and high frequencies (worst hearing for a sloping loss). For instance, the description for a sloping audiogram may look as follow:

"**Mild** hearing loss sloping **to severe**"

Site of Lesion

There are three major types of hearing loss, namely: sensorineural, conductive and mixed. These words themselves are considered to be descriptors of hearing loss. Unfortunately, the site of lesion cannot be identified uniquely from air conduction thresholds and requires bone conduction thresholds and/or other tests. The reader is invited to read Section B.4 of Appendix B for a more detailed explanation.

2.3.3 State of the Art

The audiology literature concerned with different aspects of audiogram classification is extensive, but the subset focusing on its application in automated audiometry is rather limited. One can distinguish between at least two large families of approaches for audiogram classification: *rule-based methods* and *cluster-based methods*. While the goals remain the same, the assumptions and implications of these two families of methods differ.

Rule-based Methods

The first validated rule-based system dates back from the 1940s with Carhart's audiogram classification system [42]. His system classifies audiograms based on their configuration and severity. This system first assigns the audiogram to a ''major'' configuration category by measuring the residuals between the audiogram and the curve of best fit, as well as the slope of the ''curve of best fit''. Then, it proceeds to describe deviations from this major category (audiometric notch, fluctuation around a fixed threshold value, etc.) using a more complex set of rules and descriptors. The severity of the audiogram is then also determined with another set of rules, manually refined to maximize the agreement with a set of observers. Ultimately, the audiogram description is coded as a chain of characters and numbers summarizing different aspects of the audiogram,

which makes it essentially meaningless to people unfamiliar with the system, as shown in Figure 2.6. This system was criticized for being unwieldy and unnecessarily complex [27]. One could argue that Carhart was unwilling to sacrifice completeness for the sake of simplicity and interpretability. Furthermore, it is unclear whether the "curve of best fit" is approximated visually or whether it is to be calculated exactly. The first option is more likely given that the system was developed before the large-scale adoption of computers.

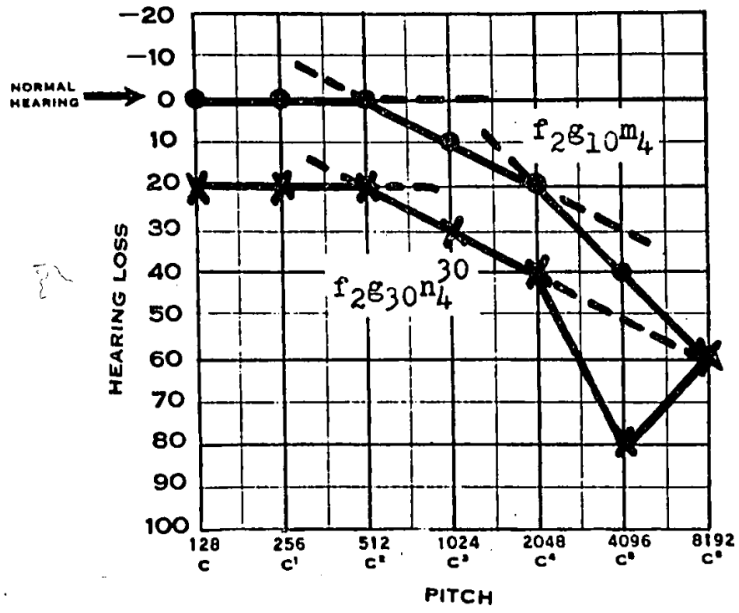


Figure 2.6: Audiogram descriptions generated by Carhart's audiogram classification system
This figure, taken from Carhart's landmark paper on audiogram classification [42], shows how difficult to interpret the descriptions generated with his system are.

Multiple rule-based systems [47, 48, 49, 50, 27] relying on hand-crafted rule were introduced since. One notable and recent example is AMCLASS™, developed by Margolis and Saly [27]. AMCLASS™ consists of over 160 rules manually derived and refined to maximize the agreement between the rule set and a panel of 5 annotators over a set of audiogram annotations. AMCLASS™ was designed partly with the intent of determining

whether there is a standardized definition for the different types of hearing loss. Margolis and Saly found that there was substantial inter-rater agreement between the annotators in terms of how to classify audiogram by configuration, severity, and site of lesion. They did not, however, assess intra-rater agreement. Nor did they assess agreement for audiogram symmetry, identification of audiometric notches and evaluation of audiogram reliability.

Rule-based systems are associated with certain advantages. For instance, they are easy to interpret, because the rules are often derived by researchers that are audiologists themselves. On the other hand, most rule-based audiogram classification systems rely on an assumption which, in practice, may not hold: audiograms belong to a single configuration class. Because of this assumption, these systems are not particularly robust when dealing with edge cases such as audiograms where two configurations are present over different frequency ranges. In addition, current rule-based systems are purely qualitative, and do not offer an estimate of confidence in the classification. Finally, rule-based systems are rigid and cannot automatically evolve as more data becomes available.

Cluster-based Methods

Cluster-based methods differ from rule-based methods in that this family of approaches is data-driven in nature. In this approach, a large collection of audiogram from the general population or a subpopulation is assembled. An unsupervised clustering algorithm is subsequently applied in order to cluster individual ears. These cluster-based approaches work on an ear-by-ear basis, and simply operate using the audiogram thresholds as features. This approach implies that audiograms cannot be clustered in terms of symmetry across ears. The number of cluster can be determined arbitrarily using judgement [51], or using quantitative metrics such as the F-statistic and the mean squared error, such as done in [52]. The clusters are then reviewed and labeled. A new

audiogram can then be classified according to the label of the nearest cluster.

In contrast with rule-based methods, cluster-based approaches are quantitative. As such, they introduce the quantitative notion of similarity (or distance) between audiograms. Cluster-based method, while further flexible than traditional rule-based methods, still require a lot of effort to refine, due to the problem of determining the optimal number of clusters, which may vary with the dataset size and nature.

2.4 Machine Learning and Statistics

This thesis makes use of concepts from machine learning and statistics. Supervised learning techniques are used here to train and validate a model whose task is to classify audiograms. Clustering and density estimation, two techniques belonging to the family of unsupervised learning applications, are also used in this thesis to sample the audiograms that eventually compose the training set, and to estimate audiogram reliability, respectively. Statistics, particularly ones related to inter- and intra-rater reliability estimation are also utilized in this work. Each of these topics are reviewed in the following sections.

2.4.1 Supervised Learning

Learning Problem

Supervised learning algorithms solve the learning problem. This problem consists in finding the hypothesis (model) in the hypothesis space $h \in \mathcal{H}$:

$$h : \mathbf{x} \rightarrow t \tag{2.1}$$

that best approximates the true process f, by mapping a vector of *features* (or *attributes*) $\mathbf{x}$ to a target value t [53]. In other words, the objective of the learning problem is to find

the optimal parameter set for h. The model is said to perform *classification* if the value to predict is discrete, and *regression* if that value is continuous. As such, models whose task is to classify and to perform regression will here forth be referred to as *classifiers* and *regressors*, respectively.

To learn, the algorithm relies invariably on a training dataset $\mathcal{D}_{train}$, consisting of instances whose features and target values are both known (labeled):

$$\mathcal{D}_{train} = \{(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2), ..., (\mathbf{x}_N, t_N)\} \tag{2.2}$$

Ultimately, the goal is to find a model that generalizes well and that is capable of correctly mapping a new unseen instance $\mathbf{x}_i$, to the correct target value t_i.

Model Training

The process through which the model parameters are tuned is termed *training*, and relies often on the application of some optimization technique such as gradient descent. Often, parameters are optimized in order to minimize a loss function, L, which is reflective of the error committed by the model in classifying instances in $\mathcal{D}_{train}$. For many supervised learning algorithms, the loss function L is a function of some distance between the actual target value y_i and the predicted value $\hat{y}_i$ over all instances in $\mathcal{D}_{train}$.

One may choose one of many loss functions such as the 0-1 loss, cross-entropy and hinge loss for classification, or the mean squared error or the $L2$ loss for regression problems. It is also possible for one to define their own loss function, which may be appropriate for some specific problems where different types of errors are to be weighed differently.

During the optimization procedure, it may be desirable to adjust *hyperparameters*. These parameters can be considered to be external to the model, because they are not learned, even though they do control *how the model is built* and *how the model*

learns from training data. Specific examples of hyperparameters will be presented in Section 2.4.1. Hyperparameters are often adjusted to minimize the prediction error on a validation dataset $\mathcal{D}_{validation}$ which, ideally, does not overlap significantly with $\mathcal{D}_{train}$.

At the end of the training procedure, the performance of the model and its ability to generalize to new, unseen data is often assessed using a third labeled dataset $\mathcal{D}_{test}$. In well-designed experiments, this dataset is composed of independent instances that were not used to train the model or adjust its hyperparameters.

Performance Metrics

A variety of metrics are used to quantify the performance of a model. It is not unusual to report the mean value of the loss function for regression problems or multiclass classification problems. In binary classification problems, where the target variable belongs to one of two classes $t \in \{positive, negative\}$, many people report the following performance metrics:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.3}$$

$$Sn = \frac{TP}{TP + FN} \tag{2.4}$$

$$Pr = \frac{TP}{TP + FP} \tag{2.5}$$

where accuracy (Acc), sensitivity (Sn) and precision (Pr) are functions of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

Some metrics, such as the F_1 score, represent an average of multiple metrics:

$$F_1 = 2\frac{Pr \times Sn}{Pr + Sn} \tag{2.6}$$

Cross-Validation

Cross-validation is a very powerful tool, particularly when limited labeled data is available. This technique, illustrated in Figure 2.7 allows for data to serve three purposes simultaneously: training and testing.

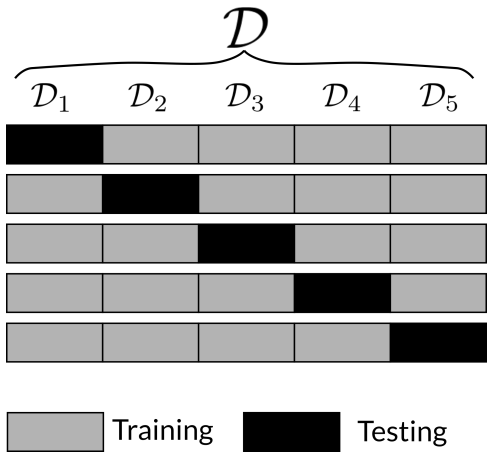


Figure 2.7: Example of a 5-fold cross-validation scheme
This figure shows how the entire dataset $\mathcal{D}$ is randomly split into 5 exclusive folds and cycled through, using, at every iteration, one of the folds for testing, and the remaining folds for training.

This technique randomly splits the dataset into N folds of equal size $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N\}$, and cycles through every fold $\mathcal{D}_i \in \mathcal{D}$ using $\mathcal{D} \setminus \mathcal{D}_i$ as a training set and $\mathcal{D}_i$ as a test set. When the data is imbalanced, i.e. there are many more instances of one class (positive or negative) than the other, it may be useful to use stratified cross-validation, wherein the folds are created such that the imbalance in every fold is representative of the overall dataset.

Cross-validation also reduces the odds of *overfitting*, the process in which the model learns the data instead of the patterns in the data. By assessing the performance of the model over different tests sets during the hyperparameter tuning procedure, the model is less likely to be tuned specifically for a small subset of the data, because all instances in the dataset are used to develop the model. It goes without saying that

cross-validation may be computationally expensive for very large datasets.

Decision Tree Learning

Decision trees, constitute one of many possible representations of a classification or regression model. Decision tree learning is particularly appreciated for biomedical applications, as decision trees are highly expressive, and as such, the predictions made by a decision can easily be understood.

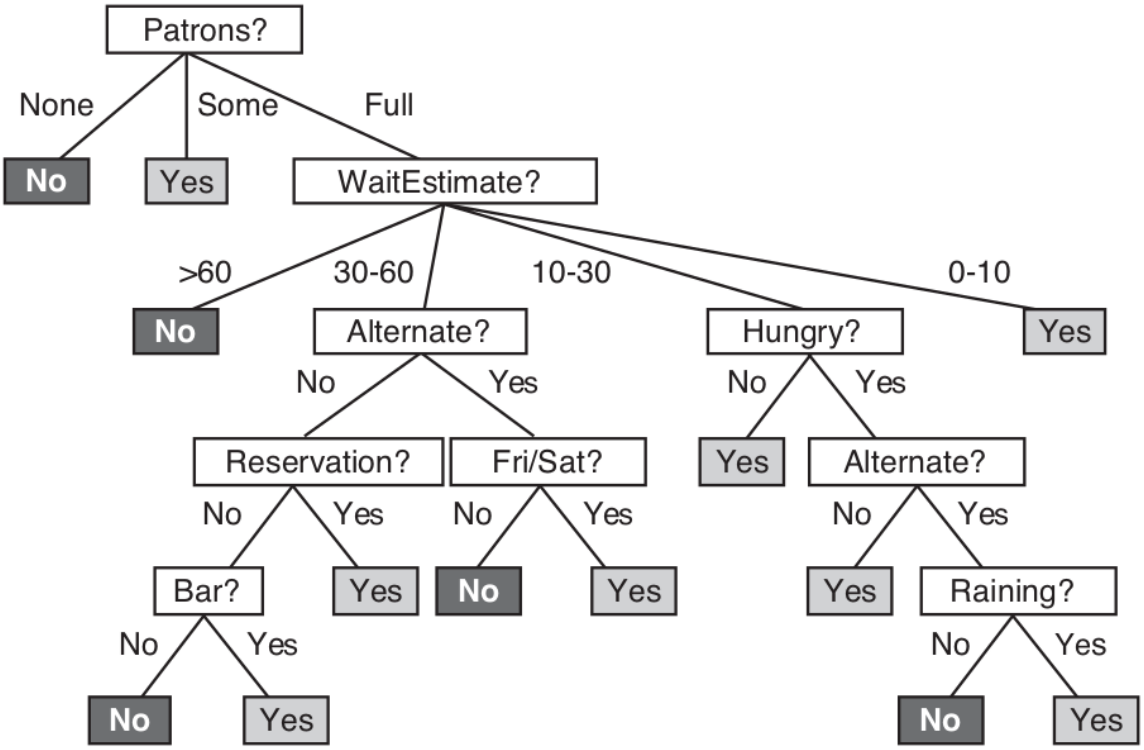


Figure 2.8: Example of a decision tree
This decision tree, reproduced from [53], determines whether a client should wait for a table. One can reach a decision for an instance by selecting the appropriate edge for the feature value at every non-leaf node until a leaf node is reached.

Popular variants of decision tree learning algorithms include Classification and Regression Tree (CART) and C4.5 [54]. The model training process consists in determining the optimal structure of tree, but how exactly this is done depends on the particular

variant of the algorithm used. In many cases, the differences between the algorithms are rather limited.

Some of these algorithms rely on the notion of information entropy, H, which represents the uncertainty associated with a dataset $\mathcal{D}$:

$$H(\mathcal{D}) = -\sum_{c \in C} P(c) \log_2 P(c) \tag{2.7}$$

where C is the set of classes in $\mathcal{D}$, and $P(c)$ is the probability that an instance belongs to class c. Other algorithms use the Gini impurity [53].

Decision trees are built recursively. Here, we describe one iteration, taking as an example the tree presented in Figure 2.8:

1. The algorithm seeks to determines which feature, and split points best segregates the instances into the positive and negative classes. In the ideal case, a single feature value would have one split point that segregates the instances into two sets consisting exclusively of positives instances in one, and negative instances in the other. Since this rarely happens in real world problems, the most valuable feature is often taken to be the one that provides the highest *information gain*, IG, or reduction in entropy associated with the split:

$$IG(f; \mathcal{D}) = H(\mathcal{D}) - \sum_{s \in S} P(s) H(s) \tag{2.8}$$

   where f is a feature, s is a set in the sets generated by the split on f, and $P(s)$ is the probability belongs to the set s. In the example, the algorithm determined that the `Patrons` feature provides the highest information gain at the root node. Notice how a restaurant with no patrons always leads to a "No" decision, and restaurants with some patrons always leads to a "Yes" decision.

2. A node in the tree is created with outgoing edges taking the possible values defined

by the split points. For example, the `WaitEstimate` feature puts split points at 10, 30, and 60.

3. If a node is no longer worth splitting, it becomes a leaf node. Leaf nodes correspond to a decision. The class assigned to an instance whose feature values lead to a leaf node is taken to be that of the majority of training instances that reached that node. A node may become a leaf node if it has sufficiently low entropy, or if the number of instances it contains is below a certain threshold, for example.

This process is repeated recursively to create sub-trees at every non-leaf node, using only feature not used to split the dataset previously. The algorithm will stop splitting nodes and create leaf nodes under certain conditions (base cases), for example when the instances of the two classes are perfectly split, or when all instances where classified in leaf nodes higher up in the tree.

Decision tree learning algorithms are associated with multiple hyperparameters, which control primarily the complexity of the tree. A non-exhaustive list is presented in Table 2.4. Selecting appropriate values is important, as decision trees are reputed for being prone to overfitting. Very deep and complex decisions trees are more likely to only memorize the training set. Putting bounds on the depth of the tree or the maximum number of leaf nodes, for example, reduces the odds that the model becomes so complex that it generalizes poorly to new data. Another technique used to control the complexity of the model is tree pruning, where subtrees are eliminated in a bottom-up fashion. The elimination proceeds iteratively up to the point where the error on the training set becomes unacceptable.

2.4.2 Unsupervised Learning

Unsupervised learning, in contrast with supervised learning, does not make use of labels for the data. In other words, any instance $\mathbf{x}$ in the dataset need not to be paired with

Table 2.4: Hyperparameters used in decision tree learning

This table present some hyperparameters that can be specified and tuned when training a decision tree.

Hyperparameter	Description
Maximum tree depth	Controls the depth of the tree
Minimum number of instances per leaf node	Prevents the creation of leaf nodes with very few instances
Split criterion	Whether information gain or the Gini index should be used
Maximum number of features	The number of features that can be used by the tree; controls model capacity (complexity)
Minimum decrease in entropy (or impurity) to split	Prevents the creation of split that only modestly decrease the entropy or impurity, to limit the complexity of the tree

a target value t. Unsupervised learning algorithms are often used to cluster, or group, instances based on some similarity measure. Another application of such algorithms is to model the distribution of instances in the feature space (probability density function), a process referred to as density estimation.

Hierarchical Clustering

Hierarchical clustering, reviewed in [55], is one of multiple clustering algorithms. The clustering procedure can be done in an agglomerative fashion (bottom-up), where the number of clusters is taken to be equal to the number of instances in the dataset, and iteratively reduced by merging nearby clusters according to some merging strategy. In the divisive approach (up-down), it is initially assumed that there is one cluster, and this cluster is iteratively divided to generate separate clusters.

In one variant of agglomerative hierarchical clustering, the merging strategy consists in defining the distance between two clusters as the maximum distance (Euclidean, Minkowski or other) between any of their two instances, and to merge the two clusters having the smallest distance. In another variant based on Ward's criterion, the two

clusters whose merging increases the merged cluster variance the least are merged. Other merging strategies include the Unweighted Pair Group Method with Arithmetic Mean and average linkage.

Determining the Optimal Number of Clusters

Determining the optimal number of clusters is not as straightforward as tuning parameters in the supervised learning framework. In unsupervised learning, the target value t of the instances in the data set is normally unknown, so it is impossible to determine whether the elements in a cluster truly are truly similar (belong to the same class).

While it is possible to assess the quality of clustering manually, objective criteria have been developed to assess the quality of the clustering. The mean silhouette index [56] is often used for that purpose. This index computes, for every instance in the dataset, the "extent of its membership" to the cluster to which it was assigned and compares it to the extent of its membership to the nearest non-encompassing cluster. This quantity is defined as follows

$$s_{i,c} = \frac{b_{i,c} - a_{i,c}}{\max(a_{i,c}, b_{i,c})} \tag{2.9}$$

where $s_{i,c}$ is the silhouette index for the instance i belonging to the cluster c, a_i is the mean of distances d between the instance and all other instances in the same cluster $a_i = mean\{d(i,j) \mid i \neq j \text{ and } i,j \in c\}$, and b_i is the distance between the instance i and the closest instance from any other cluster $b_i = min\{d(i,j) \mid j \notin c\}$. Clearly, we want this value to be maximized, which can be done by changing the distance metrics used or by changing the target number of clusters to form. The optimal number of clusters can be taken to be that at which the silhouette index is maximal.

Gaussian Mixture Models

Gaussian mixture models can be summarized as a distribution resulting from the linear combination of multiple Gaussians distributions [53]. GMMs are useful, as they can be used to model more complex (e.g. multimodal) distributions than a simple Gaussian distribution. A mixture of Gaussians has the form

$$P(\mathbf{x}) = \sum_{c=1}^{C} \pi_c \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \mathbf{S}_c) \tag{2.10}$$

which is a sum of Gaussian distributions (or *components*) c.

The parameters of a GMM are the mixing weights π_c, the mean $\boldsymbol{\mu}_c$ and the covariance matrix $\mathbf{S}_c$ of the Gaussians in the mixture. These parameters are initially unknown, and must be determined somehow. The training problem for GMMs consists in determining the parameters that best model the distribution density. The expectation-maximimization algorithm [57] assigns initial values to the parameters (sometimes using k-means clustering for initialization). An expression for the expected value of the log-likelihood function of the mixture is formulated using the initial parameters (expectation step). The parameters which maximize this expression of the expected log-likelihood of the model are computed (maximization step). The expression corresponding to the log-likelihood function of the mixture is then modified using the updated parameters, and the cycle is repeated until the parameters converge.

The number of components, a tuneable parameters, controls the complexity of the model. It can be optimized using some form of information criterion such as the Aikake Information Criterion (AIC) [58] or the Bayesian Information Criterion (BIC) [59]. Both of these values provide a measure of the compromise between the complexity of the model and its ability to fit the data. The mathematical details of these methods are beyond the scope of this thesis, and the reader may refer to the original publications for additional details.

2.4.3 Rater Reliability

One concern inherent to the development of any measurement or diagnostic tool is the reproducibility of the results. To be useful, the tool should produce reproducible results, or results within some acceptable error margin, under the same experimental conditions [60]. The results should be reproducible when performed by the same tester, in which case, the tool is said to have high *intra-rater reliability*. It is also desirable for the tool to produce the same results from one tester to another, or in other words, to have high *inter-rater reliability*.

Multiple metrics exist to compute the inter- and intra-rater reliability for tasks, such as the assignment of descriptors to audiograms. One could simply count the number of times that measurements agree and report the *percentage of agreement*. For example, one could say that audiologists assign the same symmetry descriptor to an audiogram 80% of the time. The *kappa statistic* (κ) is often preferred, however, because it takes into consideration the odds of measurements agreeing by pure change, unlike the percentage agreement. There are multiple variants of this statistic, such as Cohen's κ [61], which is applicable when the experiment has at most two raters, and Fleiss' κ [62], a generalization for experiments with an arbitrary number of raters. The value of κ ranges from -1 to 1, and is often interpreted using the scheme defined by Landis and Koch (Table 2.5).

Table 2.5: Landis and Koch's kappa statistic interpretation scheme

The interpretation scheme defined by Landis and Koch [63] is often used to interpret the results of a rater experiment.

κ statistic	Agreement
< 0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.80-1	Almost perfect

To evaluate intra- and inter-rater reliability, it is necessary to present the same data more than once. In intra-rater reliability experiments, a single rater is presented with duplicates or triplicates and asked to rate the same data. In inter-rater reliability experiments, multiple raters rate the same dataset.

3

Rapid Audiogram Annotation Environment

Software is a great combination between artistry and engineering.

-- Bill Gates

As implied by its name, the supervised learning framework invariably requires labeled data to "supervise" the learning process. In other words, the model is trained from labeled training data. The quality of the data is pivotal in training a model that can generalize well and accurately predict the label for new unseen data.

This thesis, whose first objective is to automate the classification of audiograms using supervised learning, does not escape this requirement. As such, it calls for the collection of a set of high-quality audiogram annotations.

In this chapter, we present the RAAE, a software developed specifically developed to collect audiogram annotations with the intent of building a training set for learning algorithms. More specifically, we discuss of the design process from the listing of requirements to implementation and deployment at a high level.

3.1 Software Requirements

A significant effort was deployed to understand the requirements that should underpin the RAAE's development. In this section, we discuss, although admittedly in an informal fashion, some considerations that steered the development of the software and guided design decisions.

These software requirements were developed with the intent of deploying the software on a medium scale, i.e. to run a study with a couple dozen annotators. As we will see later, in Chapter 4, the software was only used by three annotators in the work presented in this thesis. That being said, we nonetheless intended to build a scalable piece of software, so that it may be deployed on a larger scale, should there be a need to collect additional annotations to improve the performance of the learning algorithm.

3.1.1 Ease of use

The user-friendliness of the software was an important consideration when developing the RAAE. Given that it may not be possible to interact with the users, who might be audiologists working overseas, for example, we believed that the software should be intuitive and require minimal instructions. This consideration was particularly important when developing the user interface (UI) for the software. The software should make it clear with on-screen indication what is expected and when. Ideally, the software would walk the user through the annotation process.

3.1.2 Availability

The software should be easily accessible by multiple users simultaneously. The installation process, if necessary, should be straightforward. The software should run on all major computer platforms and on a variety of devices such as personal computers, tablets and mobile phones, even though optimal ergonomy on all platforms is not a priority.

Increasing the accessibility to the RAAE facilitates the collection of larger quantities of annotations.

3.1.3 Data Quality and Consistency

Data annotation collection software should be built such that the data collected is of high quality and consistent. One way to achieve this is to ensure that the software allows the adminstrators to screen the users, such that only qualified individual can contribute audiogram annotations. Furthermore, the annotations input in the software should be standardized and consistent as much as possible. This can be done through the use of UI elements such as dropdown menus, radio buttons, and checkboxes, for example. Free text forms should be used only when absolutely necessary. Furthermore, the software should prevent the submission of incomplete, inconsistent, or invalid audiogram annotations. Finally, under no circumstance should it be possible for a user to remove or temper with an annotation which has already been submitted. It should be possible to modify an annotation, but only by providing a second separate instance which will take precedence over the previous one.

3.1.4 Robustness

In order to ensure that no data is unduly lost, the software should contain mechanisms that prevent the user from entering data if the annotations cannot be stored in the database for reasons which may include the loss of the internet connection or server failure, for example. Furthermore, the software should handle situations such as server failure through regular database backups.

3.1.5 Scalability

The software should be scalable. It should be built using an infrastructure that makes it equally easy to use and administer when there are a couple or hundreds of users. The level of maintenance efforts from developers should ideally be independent of the number of users.

3.1.6 Security

It is crucial that the software be secure. It should be as difficult as possible for unauthorized individuals to access the data in the software. Furthermore, users should only have access to their own data (annotations, account information, etc.), and this data should only be accessible following adequate authentication.

3.2 Technologies

Considering the requirements mentioned in the previous sections, we opted to develop a cloud-based software that runs in the browser. Here, we describe the principal technologies used to implement the RAAE.

3.2.1 React.js

React.js [64] is a cutting-edge Javascript library maintained by Facebook that facilitates the development of so-called "reactive" user interfaces. While the inner workings of React.js are outside of the scope of this work, the "reactive" essence of React.js can be succinctly summarized: React.js constantly "watches" for changes in the state of the application (data), and reacts by updating the UI to reflect these changes. The architecture of a typical React.js application is presented in Figure 3.1.

The user interacts with the UI as it inputs data or requests data from a server, both

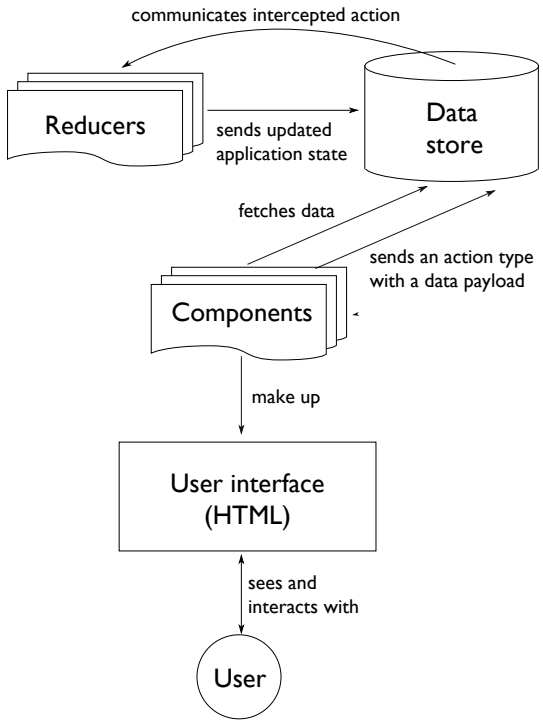


Figure 3.1: High-level architecture of a standard React.js application
This diagram presents a simplifed architecture of a standard React.js application. The user can modify the state in the data store by dispatching "actions" to the store through events such as mouse clicks or key strokes. The actions are intercepted by the store and sent to units of code called "reducers" that update the state in the data store. Updates in state trigger component changes and automatic re-rendering of the user interface.

of which trigger changes in the state of the application, ultimately resulting in changes in the UI. These changes can include, for example, the addition of text on the web page, color changes, insertion of images or icons, etc. The UI is completely separate from other business logic which is executed on a remote server. React.js is concerned only with the front-end of the application; i.e. rendering data.

React.js is *component-based*, meaning that all elements in the application are individual components living in a hierarchical organization of components. Examples of components include text boxes, buttons, dropdown lists, images, *etc.* The purpose of components is to organize and reflect the state of the application. Data from the

store must, by consequence, be explicitly passed to the components in the hierarchy. Typically, it is passed to components higher up in the hierarchy, and cascades down to components in lower levels.

In many non-trivial web applications, the state of the application is maintained in a "data store". Components may then interact -- i.e. receive or modify the state -- directly with the store instead of getting data from components higher up the hierarchy. The use of stores greatly facilitate the development of applications with a complex state containing a wide range of variables.

3.2.2 Node.js

Node.js is a runtime that allows for rapid development of server-side code using the Javascript language which was once exclusive to the internet browser. The platform has gained a lot of popularity for server development in recent years, and a wide range of frameworks for web server development are now available for Node.js, among which Sails.js, Koa.js, Loopback and Express.js.

Express.js is one of the most popular frameworks for building RESTful APIs, a type of server that accepts requests from some client and responds following some sort of processing, e.g. data retrieval from a database, image processing, *etc*. These requests may communicate the intent to receive data (GET request) or to send data (PUT or POST requests). RESTful APIs differ from other web servers in that they expose an interface for interacting with the server through URL endpoints. The exact implementation that runs on the server is unknown, and often irrelevant to the client. RESTful APIs operate in the context of a client-server architecture with request-response cycles.

3.2.3 Amazon Web Services

Amazon Web Services (AWS) now dominates the market of cloud computing [65], and provides a comprehensive set of services for building personal or commercial-grade

applications, all in the cloud. AWS's tooling supports an unfathomably large set of architectures, and is used by the likes of Adobe, Comcast, Kelloggs, *etc*.

For a reasonable price, AWS provides the tooling (e.g. web servers, databases, processing pipelines, machine learning toolkits, etc.) necessary to develop scalable web applications. Applications developed using these tools are, by default, automatically managed by AWS. Using services like AWS Beanstalk, it is possible to upload a web application that will be deployed, monitored and protected by Amazon without any intervention from the developer. Using the auto-scaling functionality, server allocation can dynamically respond to increase in traffic and spin up servers to handle increased load on servers. Very little work is required from developers to maintain an infrastructure on AWS.

3.3 High-Level Architecture

As mentioned previously, we elected to build the RAAE as a cloud-based web application. The only requirement to access the software is a reasonably fast internet connection. This decision also makes the application cross-platform by default, as it can be accessed in any browser, be it on a laptop, a tablet or a mobile phone.

The application divides into three major components: the user interface (or front-end), the API server and the database. All of these components are hosted in the cloud AWS. Once the user visits the RAAE's URL, the application's user interface implemented using the React.js library is delivered to the browser. All subsequent requests made directly from the user's browser are directed to an Express.js API server living in AWS Beanstalk. This API server acts as an intermediary between the user and the database, and implements the logic necessary to authenticate users, to retrieve the audiograms and annotations, *etc*.

The database, hosted in AWS's Relational Database Service (RDS), is only accessible

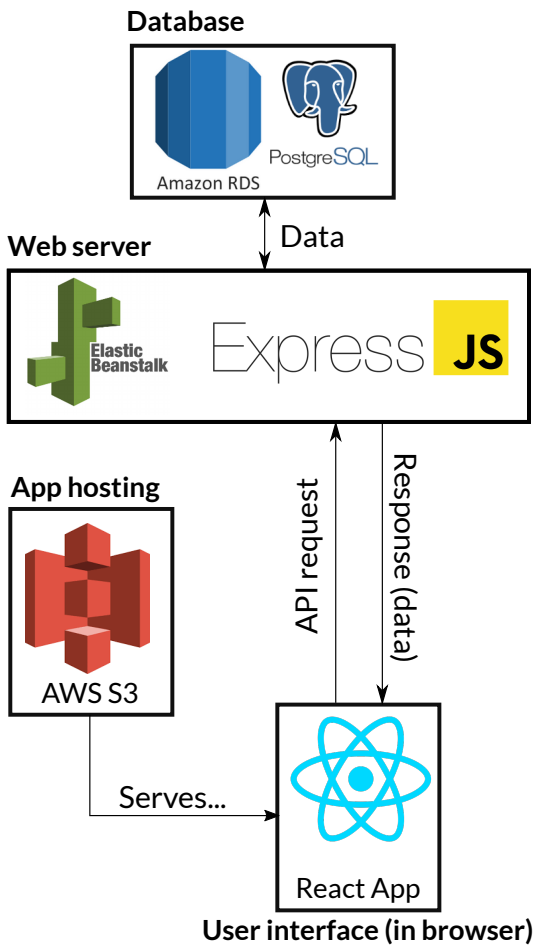


Figure 3.2: High-level architecture of the RAAE
All components of the RAAE (user interface, API server, and database) are hosted in the cloud, more specifically on Amazon Web Services' infrastructures.

by the web server, mitigating risks of intrusion in the database.

3.4 Database

To store the data collected as part of this study, we designed a database that we implemented using a PostgreSQL database engine hosted in Amazon's RDS. An entity-relationship diagram for the database is presented in Figure 3.3.

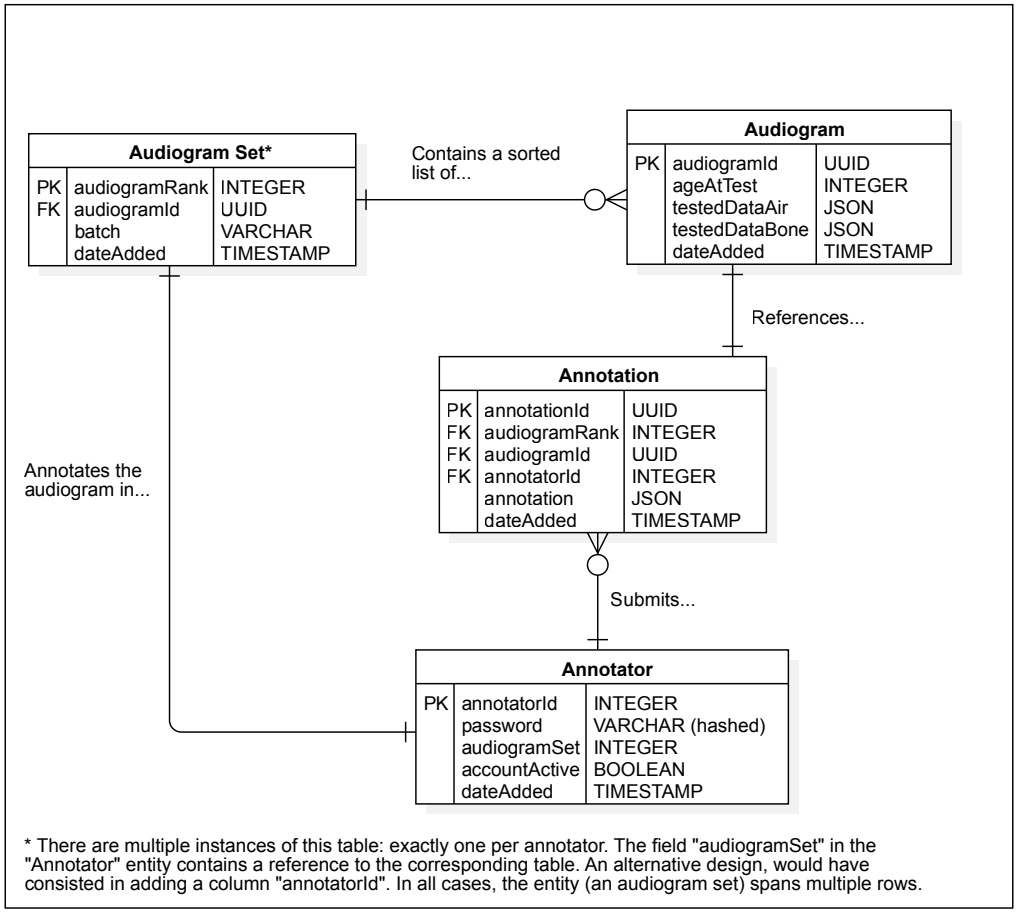


Figure 3.3: Entity-relationship diagram summarizing the RAAE's database architecture

The relationship and cardinality linking the different tables in the database developed for the RAAE are shown here. The boxes correspond to ''entities'' -- or tables -- in the database. All entities are described by fields (second column) of some type (third column). The abbreviations ''PK'' and ''FK'' indicate that the fields are primary keys and foreign keys, respectively.

Using standard SQL statements such as ''SELECT'', ''JOIN'', ''INSERT'', ''MERGE'' and ''ORDER BY'', data from the different tables can be retrieved and organized arbitrarily using complex queries. The exact queries developed here are not necessary for understanding the work done here, but suffice it to say that SQL queries are performed to retrieve and insert annotators, audiograms and annotations in the database.

Using the *pg* driver for Node.js, our Node.js web server can query the database and

obtain the data returned by the query in the JSON format. This is particularly convenient, because JSON is the *de facto* standard for communication with RESTful APIs.

3.5 User Interface Design

We introduced, in Section 3.1.1, the importance of ensuring that the application be easy to use. The RAAE consists of a total of 4 different pages: a log in/sign up page, a home page, an annotation zone, and a page with a minimal set of instructions for annotators.

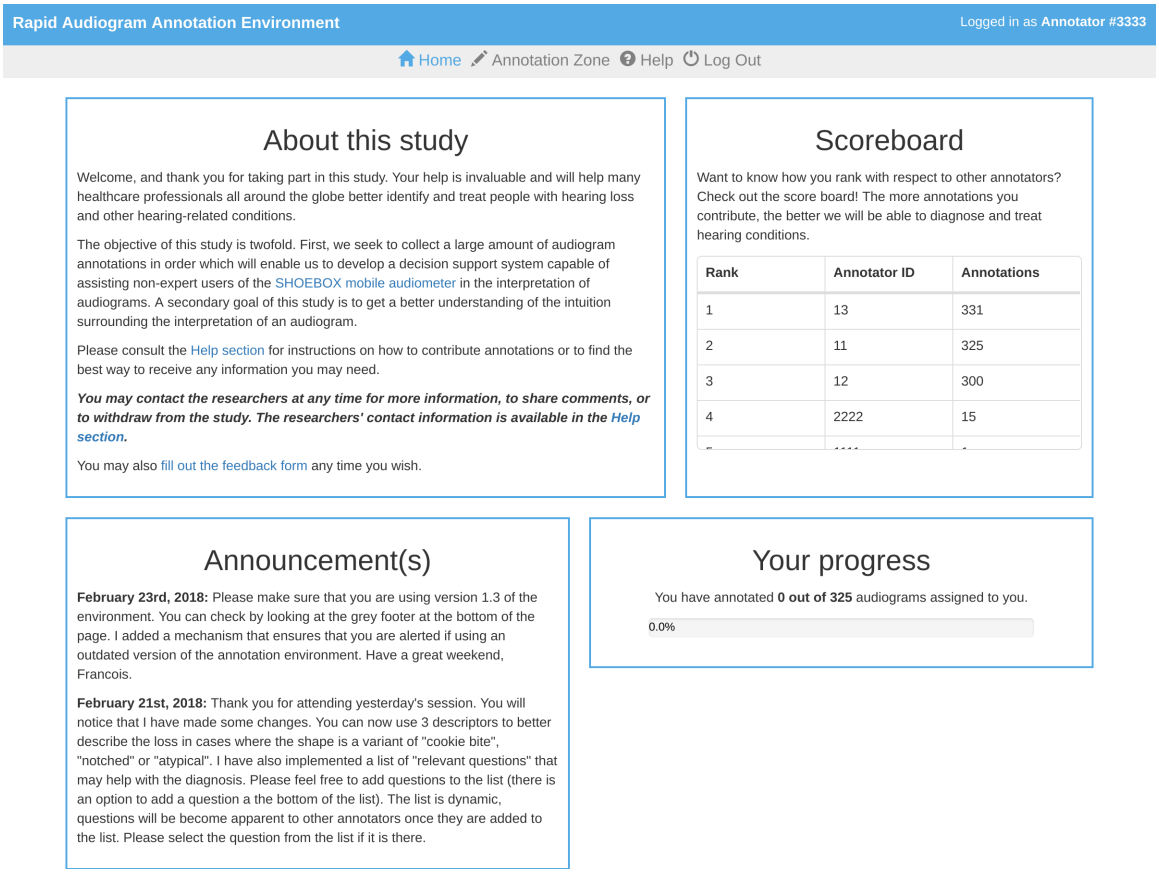


Figure 3.4: The RAAE's Home Page
The Home Page allows annotators to read announcements, track their progress, and see how they rank on a scoreboard. The announcements are retrieved from the server via an API call, making them version-independent.

Once users log in, they are redirected to the Home Page (Figure 3.4). On this page,

they may read announcements related to the project, and get updates. Annotators may also consult the "score board" which displays the number of annotations provided by all annotators. Some annotators may be motivated to provide more annotations upon comparison with other annotators, particularly if some sort of reward is offered to the top ranking annotators.

Users are encouraged to read the instructions in the Help Page which contains basic instructions regarding the annotation task as well as contact information to reach the researchers.

All annotations are done in the "Annotation Zone" whose interface is presented in Figure 3.5. Audiogram annotations are collected in a controlled, systematic fashion that leaves minimal room for inconsistency across annotations and annotators. For every audiogram, a standardized series of questions is presented. The audiogram is displayed at the top of the screen at all times, while the questions are presented within a "panel" in the bottom half. This allows the audiologist to refer to the audiogram with no scrolling or clicking while answering the questions. The questions are presented sequentially, and the annotator may only proceed to the next question after having provided a satisfactory answer to the current question. All questions for which the answer must be standardized are answered via buttons or dropdown menus, or through mouse clicks on the audiogram markers. Other data in natural language, such as comments on the audiogram's reliability, are collected through text fields. For every audiogram, a summary of the annotation is displayed prior to submission to allow the annotator to go back and make corrections, if necessary.

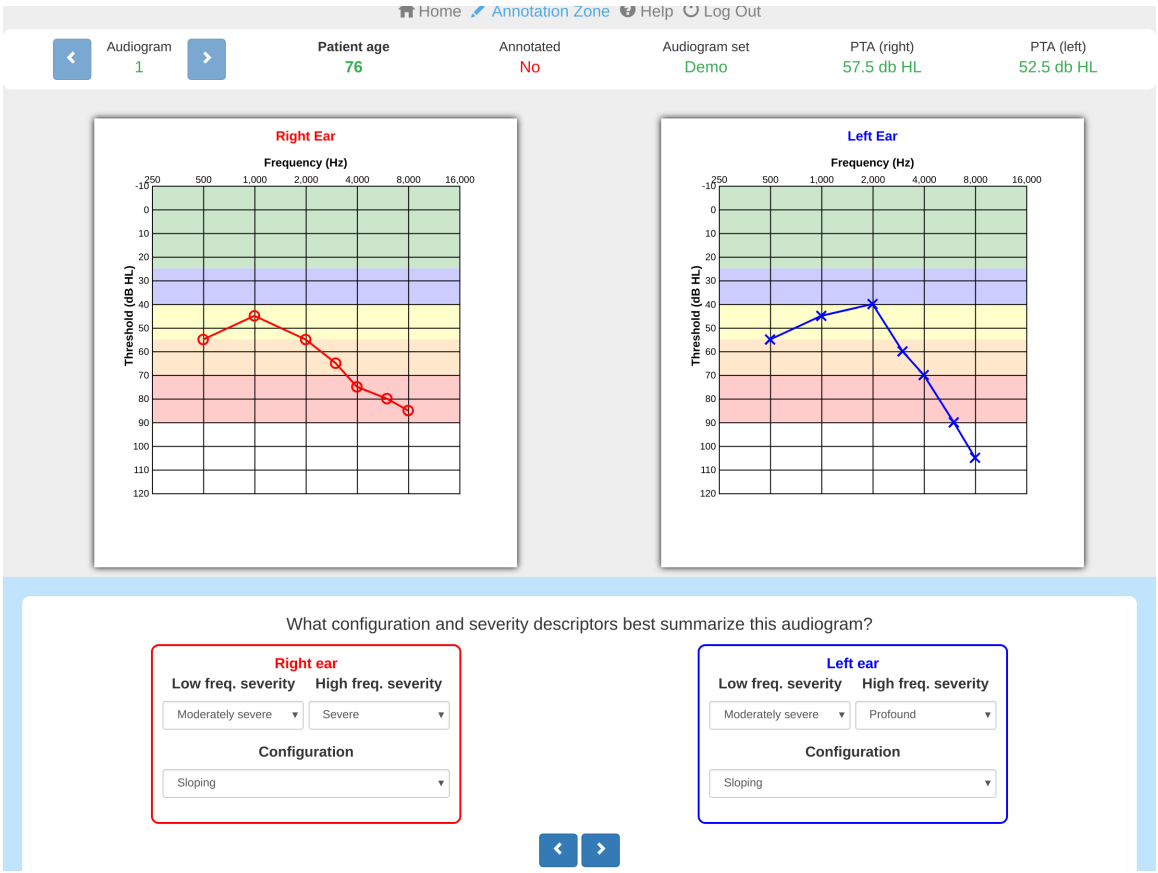


Figure 3.5: The RAAE's Annotation Zone
Annotations are provided through the "Annotation Zone" which displays the audiogram and asks a series of questions about the audiogram in a systematic, consistent fashion.

3.6 Other Considerations

3.6.1 Security

We took several measure to ensure that the RAAE is secure. This concern is particularly important because audiograms are a form of medical data which is protected by privacy and confidentiality laws.

First, a password-based authentication workflow relying on the JWT token technology [66] was implemented. Upon successful identification through a login form, the user is assigned a "signed token" (string of characters) which is then stored in the user's

browser. Any subsequent use of the application through requests requires the inclusion of that token in the header of HTTP request headers. The web server only serves requests with a valid, non-expired token, and returns an error code for unauthorized users. We elected to assign a lifetime of 2 hours to the tokens, such that an annotation session expires after that period. Upon token expiration, the user is prompted to log back in and assigned a new token.

Second, we ensured that all communications between the user's browser and the web server were completed through the HTTPS protocol. The HTTPS protocol ensures that communications between the client and server are encrypted and only meaningful to these two parties. Obtaining an HTTPS certificate, required to establish the authenticity of the web server, was made trivial by AWS.

3.6.2 Administrative Privileges

The RAAE's administrator must be able to do the following actions, among others:

- Approve an annotator registration request
- Assign an audiogram set to an annotator
- Add/order audiograms in an audiogram set
- Post a message/notification to be read by the annotators on the landing page

Due to time limitations, all of these operations were made possible by a set of Python scripts. Ideally, these tasks could be performed from a browser-based dashboard available only to users with administrative privileges.

3.6.3 Robustness

The last, but not the least, consideration regarded the robustness of the application. In particular, we aimed to ensure that no data would be lost if the software were to fail.

Two measures were implemented to achieve this. First, we set up the database such that a backup (“snapshot”, in AWS terminology) is performed daily. Second, we implemented a mechanism that logs the user out and notifies him or her if the annotation cannot be inserted in the database.

3.7 Deployment

The RAAE was developed according to the principles outlined previously, and was deployed on AWS in mid-February 2018. As described in the upcoming chapter, three audiologists provided hundreds of audiogram annotations with the RAAE with no major issues reported.

Audiologists had the opportunity to provide feedback throughout the annotation procedure, as well as in a group discussion that followed. The feedback was largely positive with only a few suggestions on how to improve the UI. An example of suggestion includes the ability to copy the labels for one ear to the other in cases of symmetrical loss. Future releases will address these suggestions.

4

Dataset Assembly and Analysis

Knowledge is of no value unless you put it into practice.

-- Anton Chekhov

In the previous chapter, we outlined the development of the RAAE, a web-based annotation environment built specifically for the purpose of collecting audiogram annotations.

This chapter concerns itself with the assembly of a high quality dataset of audiogram annotations. First, we present the strategy developed to optimally sample audiograms from a very large dataset to maximize the unique information collected using the minimum number of audiogram annotations. Next, we summarize the experimental design which leverages the RAAE, before proceeding to an analysis of the resulting collection of annotations.

4.1 The NHANES Dataset

The National Health and Nutrition Examination Survey (NHANES) is a national health survey conducted on a continuous basis in the United States. The purpose of this initiative is to provide researchers with health data representative of the general population. Subjects are carefully sampled and recruited, to prevent bias. The Center for Disease Control

(CDC) publishes and grants access to an anonymized version of the data collected as part of the survey.

A portion of the survey assesses the hearing status of subjects through pure tone audiometry. As such, the NHANES dataset contains a large collection of pure tone audiograms. In this work, we retrieved the audiograms acquired between 1999 and 2012, and aggregated them to build a dataset of 15,498 audiograms from participants aged between 12 and 85 years (mean: 39 $\pm$ 21 years).

The audiograms were obtained using a standard pure tone audiometry protocol, using a conventional audiometer, and supra-aural or insert headphones, although it is unknown which kind was used for any specific audiogram. Air conduction thresholds were measured at 7 test frequencies: 500 Hz, 1,000 Hz, 2,000 Hz, 3,000 Hz, 4,000 Hz, 6,000 Hz and 8,000 Hz, without masking in the non-test ear (please refer to Section B.5 for more information about masking). Bone conduction thresholds were not recorded in the survey.

4.2 Pre-selection

We applied a series of pre-selection filter to the initial NHANES dataset in order to eliminate incomplete audiograms or uninformative ones. The pre-selection pipeline is summarized in Figure 4.1.

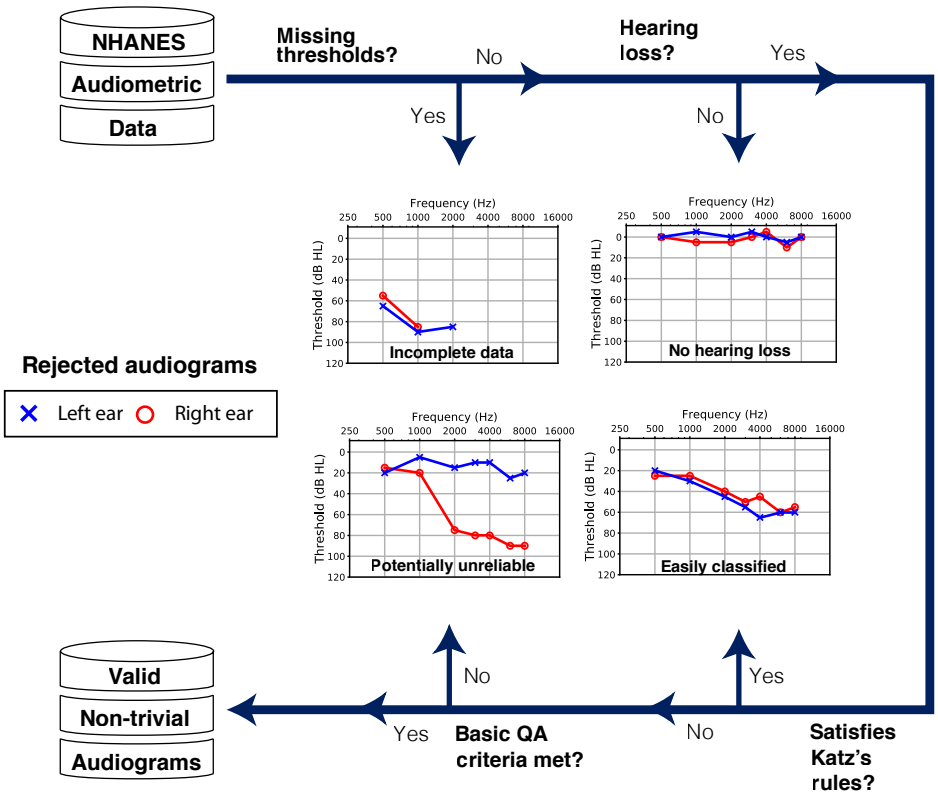


Figure 4.1: Audiogram pre-selection pipeline
A series of filters were applied to ensure that the audiograms used in sampling are complete, valid and informative.

First, to ensure consistency of the data, we discarded audiograms, i.e. both of the ears, where any one of the 14 thresholds from both ears were missing.

Next, we discarded audiograms where all thresholds are within normal limits of hearing, defined as $\leq$15 dB for children and $\leq$25 dB for adults, as specified by the ASHA's guideline on hearing loss [67].

Thirdly, we applied the rules defined in [28] which are adapted from Carhart's orginal classification system. More specifically, we fit a line through the thresholds from every ear, and eliminated audiograms where all thresholds fell within $\pm$10 dB from the line of best fit for both ears, as these audiograms can easily be classified according to the rules in Table 4.1.

Table 4.1: Classification rules used to identify trivial audiogram configurations

A modified rule set inspired by Katz's classification rules [28] was applied to all the ears where all thresholds fell within $\pm$ 10 dB of the line of best fit, to assign a configuration descriptor.

Shape	Slope
Flat	Between -5 and 5 dB/octave
Sloping	Between 5 and 14 dB/octave
Precipitous	Between 15 and 20 dB/octave
Rising	Below -5 dB/octave

All thresholds were obtained in absence of any masking noise in the contralateral ear. It is likely that if the non-test ear has better hearing than the test ear and that the tone is loud enough, the subject would respond due to crosshearing (see Section B.5). As such, it would be unlikely to observe a threshold gap greater than inter-aural attenuation, which we took to be 50 dB, although this value is highly dependent on the individual and their head physionomy. Gaps greater than 50 dB at one frequency between ears were thus considering to be potential indicators of unreliability. We therefore ended the audiogram pre-selection pipeline with the elimination of audiograms where such a gap was present for 2 or more thresholds.

4.3 Audiogram Sampling

The number of possible audiograms is unimaginably large. Taking only physiologically plausible audiograms with six air conduction measurements per ear into consideration, it was determined that, in theory, 3.62 million unique audiograms could be observed [27]. This number skyrockets and reaches 376 billion, when bone conduction measurements are considered. The vast number of possible audiograms is easily observed in the NHANES dataset, where the audiograms are remarkably heterogeneous. Nevertheless, one quickly notices that many audiograms have similar patterns. It is a reasonable assumption that the the anatomy and physiology of the ear (summarized in Appendix A)

and the different hearing conditions, constrain the audiogram space to a set of clusters.

In order to minimize redundancy in our collection of annotation and maximize the information that it will hold, we developed a sampling strategy that combines clustering and greedy sampling.

4.3.1 Feature Engineering

Before performing clustering, it was necessary to engineer the features that span the audiogram space. In many previous studies which focused on the clustering of configuration, audiograms were clustered on an ear-by-ear basis (eg. [52, 68, 69]). As such, previous studies used thresholds as features, with one study [52] using normalized thresholds so that only configuration, and not severity, influences the cluster formation. In this study, we wish to select heterogenous audiograms on the basis of configuration, symmetry, and severity. It follows that the audiogram representation, in terms of features, should capture these various aspects. Given that symmetry was an aspect, our feature space is defined for an entire audiogram, and not for a single ear, a significant departure from previous audiogram representations.

The feature vector for the audiogram is obtained through concatenation of the features presented in Table 4.2.

Table 4.2: Features used in audiogram clustering

We used distinct features that clearly represent each aspect of an audiogram: the shape, severity or symmetry.

Feature	Description
$\mathbf{t}'_{better}$	Air conduction thresholds of the better ear normalized at 1,000 Hz
$\mathbf{t}'_{worse}$	Air conduction thresholds of the worse ear normalized at 1,000 Hz
PTA_{better}	Pure tone average of the better ear (calculated with 500 Hz, 1,000 Hz, 2,000 Hz and 4,000 Hz)
PTA_{worse}	Pure tone average of the worse ear (calculated with 500 Hz, 1,000 Hz, 2,000 Hz and 4,000 Hz)
NI_{better}	The notch index [70] of the better ear
NI_{worse}	The notch index of the worse ear
$\Delta_{0.5,1}$	Root mean square difference between the thresholds of the best and worse ear for 500 Hz and 1,000 Hz
$\Delta_{2,3,4}$	Root mean square difference between the thresholds of the best and worse ear for 2,000 Hz, 3,000 Hz, and 1,000 Hz
$\Delta_{6,8}$	Root mean square difference between the thresholds of the best and worse ear for 6,000 Hz and 8,000 Hz

The features were defined such that their contribution to the audiogram representation is unambiguous. For instance, the vectors of normalized thresholds t' account solely for the configuration of the ear, as does the notch index NI. The severity of the hearing loss is captured by the PTA feature, while the symmetry over three frequency ranges is captured with the Δ features. This is in contrast with the traditional approach of using simply the unormalized thresholds as features where the contributions of the features are nebulous.

In order to further limit redundancy, we defined the features to be independent of the side (left or right). Had this not been done, two audiograms with identical thresholds in opposite ears would yield considerably different feature vectors. To ensure this does not happen, we used the PTA of the ears to relabel them as "better" and "worse".

4.3.2 Clustering

Following our assumption that audiograms are located in clusters constrained both by the hearing status of the individual and the physiology of the ear, we performed agglomerative hierarchical clustering to group similar audiograms together. By only selecting one representative from the clusters, we ensure that the audiograms selected for annotations are unique.

In reality, experts have found that most hearing loss patterns can be represented by a dozen configurations [51]. In practice, however, if one was to assume that there were 12 clusters or so, the clustering would result in the formation of highly heterogeneous clusters. In this work, we made no prior assumption about the number of clusters to be obtained, and rather sought to uncover a natural number of clusters. To do so, we used the silhouette index, introduced in Section 2.4.2 and repeated the clustering experiment such that the maximal value for the index is obtained. The clusters which yield the highest silhouette index were assumed to be the correct ones.

Following the clustering procedure, we selected the audiogram that is closest to the cluster center, as we assumed it to be a good representative for the cluster. We refrained from using the cluster centroid, because the resulting audiograms would have been ''synthetic''; resulting possibly in unlikely configurations, and/or thresholds between the conventional 5 dB increments used to report audiometric data. When the cluster had only two instances, we picked the audiogram that appears first in the NHANES dataset (sorted by a sequence identifier assigned by NHANES).

4.3.3 Sampling

To further reduce the size of the collection of audiograms to be annotated, we developed a greedy audiogram sampling Algorithm 1. Throughout the sampling procedure, we aim to balance the two following objectives:

1. **Exploring as much of the space as possible:** We seek to sample audiograms that are dissimilar one from another, as heterogenous dataset will be much more informative to a learning algorithm.

2. **Sampling common audiograms:** We seek to sample audiograms, which are more likely to be encountered when the classification system will be deployed, and minimize the likelihood of sampling outliers.

The sampling algorithm is greedy, in that it attempts, at every iteration to select the audiogram that has the highest sampling score s which ranges between 0 and 1. We defined this score as follows:

$$s_i = \alpha \underbrace{\frac{d_i - d_{min}}{d_{max} - d_{min}}}_{Objective\ 1} + (1 - \alpha) \underbrace{\frac{c_i - c_{min}}{c_{max} - c_{min}}}_{Objective\ 2} \tag{4.1}$$

where d_i is the Euclidean distance between the audiogram i and the closest sampled audiogram, c_i is the size of the cluster the audiogram i was sampled from, and α is a coefficient used to tune the relative importance of increasing space coverage as opposed to sample audiograms from populous clusters.

The algorithm, being exhaustive, has a runtime of $O(nk)$ where n is the total number of audiograms and k is the number of audiograms to sample, which was acceptable given that only a couple hundred audiograms we to be sampled, given the limited annotation resources available for the study.

Algorithm 1 The Greedy Audiogram Sampling Algorithm

Input: $\mathcal{A}$, the set of audiograms to sample from
Input: N, the number of audiograms to sample
Input: α, the relative importance of normalized distance in the score function
Output: $\mathcal{A}_{sampled}$, the set of sampled audiograms
procedure BuildAudiogramSample($\mathcal{A}$, N, α)
    $\mathcal{A}_{sampled} \leftarrow \{a_{seed}\}$
    $\mathcal{A} \leftarrow \mathcal{A} \setminus \mathcal{A}_{sampled}$
    repeat
        $s_{best} \leftarrow -\infty$ ▷ Initialize the best score
        $a_{best} \leftarrow$ None ▷ Initialize the audiogram with highest score
        for all $a \in \mathcal{A}$ **do**
            $d'_a \leftarrow$ computeNormalizedMinDistance(a, $\mathcal{A}$)
            $c'_a \leftarrow$ computeNormalizedClusterSize(a, $\mathcal{A}$)
            $s_a \leftarrow \alpha d'_a + (1 - \alpha) c'_a$
            if $s_a > s_{best}$ **then**
                $a_{best} \leftarrow a$
                $s_{best} \leftarrow s_a$
            end if
        end for
        $\mathcal{A}_{sampled} \leftarrow \mathcal{A}_{sampled} \cup \{a_{best}\}$ ▷ Sample the audiogram with highest score
        $\mathcal{A} \leftarrow \mathcal{A} \setminus \{a_{best}\}$ ▷ Remove it from the pool
    until $|\mathcal{A}_{sampled}| = N$
    return $\mathcal{A}_{sampled}$
end procedure

We ran the algorithm for multiple values of α, and selected the value that offered the best compromise between heterogeneity and frequency. The final value selected was $\alpha = 0.6$, and the resulting dataset was confirmed to be sufficiently heterogeneous by the Clearwater Clinical Ltd.'s in-house audiologist.

4.4 Experimental Design

The experiments outlined here were approved by the Carleton University Research Ethics Board (CUREB-B, #108278). The documentation related to ethics is available in Appendix C.

4.4.1 Rater-Reliability Experiment

Part of the work done here, involved quantifying the intra- and inter-rater agreement on audiogram annotation tasks. To measure intra-rater reliability, it is necessary to present audiologists with audiogram replicates. For inter-rater reliability, one must present a set of audiograms to multiple audiologists. One outstanding question related to experimental design is: *how many replicates should be presented to obtain an accurate estimate?*

In this work, we use the approach proposed by Gwet [60] to determine a reasonable number of replicates for our intra-rater reliability assessement. This approach, however, requires that an approximation of the expected value of the κ statistic. It is then possible to use this estimate and to determine the expected uncertainty associated with different experiment designs (e.g. many duplicates vs. fewer triplicates). We therefore presented the in-house audiologist at Clearwater Clinical Ltd. with 20 audiograms sampled at random and requested that the audiograms be annotated twice, leaving approximately three days between each annotation session. Audiograms were shuffled between sessions. We computed the κ statistic for the descriptions of symmetry, severity and configuration, and determined an appropriate number of duplicates.

To simplify the experiment aiming to evaluate inter-rater reliability, we opted to set up the experiment so that all annotators review the same audiograms. This also comes with the advantage that multiple ground truth values can be obtained for a descriptor.

4.4.2 Building the Annotation Set

Using the greedy sampling algorithm, we sampled from the top-ranking audiograms in terms of sampling score. Knowing that a total of 320 annotations per annotator could be collected, due to budgetary and time restrictions, we designed a common set of audiograms. The set consisted of 270 different audiograms, 40 of which were repeated

to yield the final 320 audiograms in the annotation set. We had determined, from our initial annotation experiment, that 40 duplicates should be sufficient to predict intra-rater reliability with an error of approximately 0.1 on the κ statistic [60]. Increasing the number of duplicates to 50 only marginally improved the expected error in κ, reducing it to approximately 0.075.

We included a number of ''trivial'' audiograms whose two ears could easily be classified using the rules in [28]. The aim for this was to determine whether the audiologists tend to reproduce the configuration annotations produced with these rules, which they were trained with. Of course, some ''non-trivial'' audiograms may contain an ear that is easily classified, as long as the other ear is not.

We selected audiograms from both the ''trivial'' and ''non-trivial'' types for duplication, in order to measure the agreement for these two types of audiograms.

The final audiogram set is summarized in Table 4.3.

Table 4.3: Audiogram set composition

This table provides the breakdown of audiograms used to build the audiogram set presented to professional audiologists.

Number of presentations	Non-trivial	Trivial	Total
Once	200	20	220
Twice	40	10	100
Total	280	40	320

The order of the audiograms was shuffled, but we ensured that no audiogram duplicates were consecutively presented to the audiologist.

4.4.3 Annotation Process

We recruited three licensed professional audiologists trained at different institutions, and with varying years of experience. The audiologists were selected so as to impart the variability that exists across audiologists to the final annotations set.

We organized an initial meeting with the audiologists and presented the study, along with its objectives, before introducing them to the RAAE and the annotation procedure. They were also informed that their assigned set would contain a quantity of duplicates.

We conducted a short annotation session on 5 audiograms in order to allow the audiologists to familiarize with the environment, and to determine whether modifications should be made to the classification procedure before launching the data collection.

Every audiologist was assigned their own set of audiograms which differed only in the order in which the audiograms were presented, and instructed to annotate the audiograms using the RAAE. More particularly, he audiologists were asked to provide answers to the questions summarized in Table 4.4.

Table 4.4: List of components in an audiogram annotation

Audiologists were presented with a series of questions for every audiogram through the RAAE.

Question	Possible answers
Is the audiogram symmetrical?	*Yes, no, indeterminate*
What is the configuration?[1,2]	*Flat, sloping, precipitous, reverse sloping, notched, cookie bite, reverse cookie bite, atypical,*
How severe is the loss?[1,3]	*Normal, mild, moderate, moderately severe, severe, profound*
How was this audiogram?	*Easy, challenging*
Are there potentially unreliable thresholds?[1]	Possibility to click on unreliable thresholds
Are there notches?[1]	Possibility to click on thresholds in a notch

[1] On a per ear basis
[2] Only required for ears where there is hearing loss
[3] The number of descriptors varies between 2 and 3, depending on the configuration provided

We avoided being overly specific in the instruction as to influence the annotations, although we did specify that Goodman's scale (see Table 2.3) should be used to describe severity. This ensures some basic consistency, while leaving some discretion as to what frequencies should be considered to describe the extent of the loss in the low, mid, and high frequencies, for example.

The identification of thresholds belonging to an audiometric notch was effectively a

binary classification question for every single threshold in the dataset. The question amounted to *"Does this threshold occur within an audiometric notch?"*. If the annotator believed that the threshold occured within a notch, they could indicate it by clicking on the threshold. The same idea applied for the identification of questionable thresholds, in which case, the question amounted to *"Is this threshold questionable in quality?"*.

Given that the audiologists had two occasions to indicate the presence of a notch within the audiogram, we instructed them to assign the ''notched'' configuration to the audiogram if the audiogram has a notch and to select the thresholds in the notch, but is otherwise normal. In other words, the ''notched'' configuration was to be selected if the *overall* configuration of the audiogram was notched. If a notch occured within an audiogram which shows a sloping trend for example, then the annotators were instructed to label the audiogram as sloping and to select the thresholds that belong to an audiometric notch.

The annotations were collected over a period of approximately two months.

4.5 Retrospective Audiogram Review

Once the annotation deadline expired, the annotators were convoked to discuss the annotation process and to review certain annotations. Audiograms which had obvious mistakes associated with data entry were corrected during this session. In fact, we found that in several cases (5) instances where an annotator had made an obvious mistake by selecting ''reverse sloping'' instead of ''sloping'' from the configuration dropdown menu, for example. That was likely due to the fatigue that may arise in longer sittings. The instances that were caught represent a negligible portion of the audiograms, but given the small size of the dataset, these few mistakes could have proven to be costly in ways of accuracy.

The audiologists were asked to provided feedback regarding the difficulty of the

task, the RAAE software and the overall experience. Some of the insights gained as part of this session were very useful in interpreting the annotation results and will inform future revisions of the RAAE.

4.6 Results

4.6.1 Audiogram Sampling

We found that agglomerative hierarchical clustering on the 8,564 pre-selected audiograms maximized the silhouette index when the number of clusters is set to below 9 (see Figure 4.2). Such a small number of clusters however, appears largely unrealistic, especially when both ears are represented in the audiogram. The presence of both ears allows for a very large variety of configuration combinations across ears, severity combinations and symmetries. The next best silhouette index was obtained for 3000 clusters; a figure that is far more realistic.

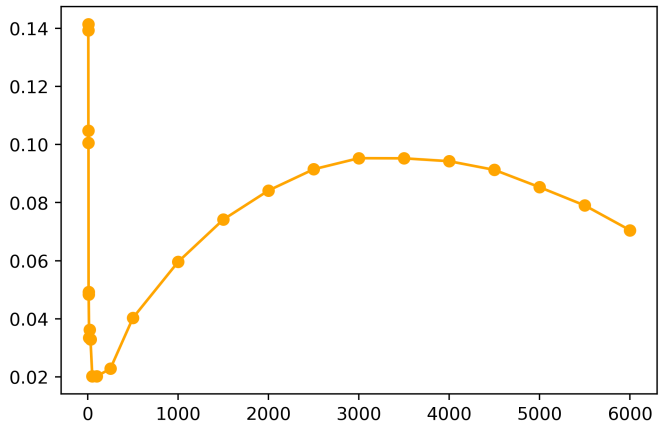


Figure 4.2: Silhouette index as a function of the number of cluster
The silhouette index was maximized by varying the number of clusters to determine an objectively appropriate number of natural clusters in the audiogram space defined in this study.

As expected with such a high number of clusters, most clusters are small, with fewer than 4 members (Figure 4.3). The average cluster size is approximately 3. We estimated that this would not be problematic if audiograms are intelligently sampled from these numerous clusters.

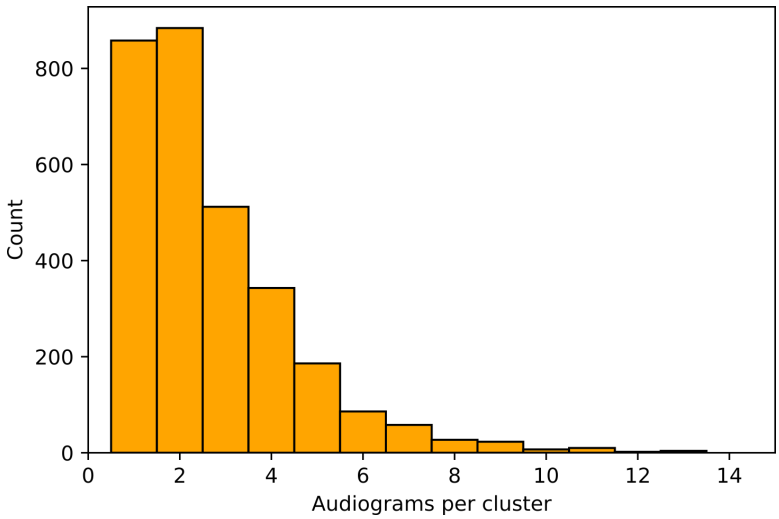


Figure 4.3: Distribution of the cluster sizes
Using 3,000 clusters as a parameter for agglomerative hierarchical clustering, a right-skewed distribution of cluster sizes is observed.

Having such a high number of clusters does imply that the clusters have higher homogeneity, at the cost of a lesser redundancy reduction. We were willing to accept a lesser redundancy reduction as clustering was only an intermediary step in the sampling pipeline. Larger clusters tended to contain common audiograms such as those with only mild hearing loss, and smaller clusters, more interesting ones, consistent with our assumptions, as illustrated in Figure 4.4.

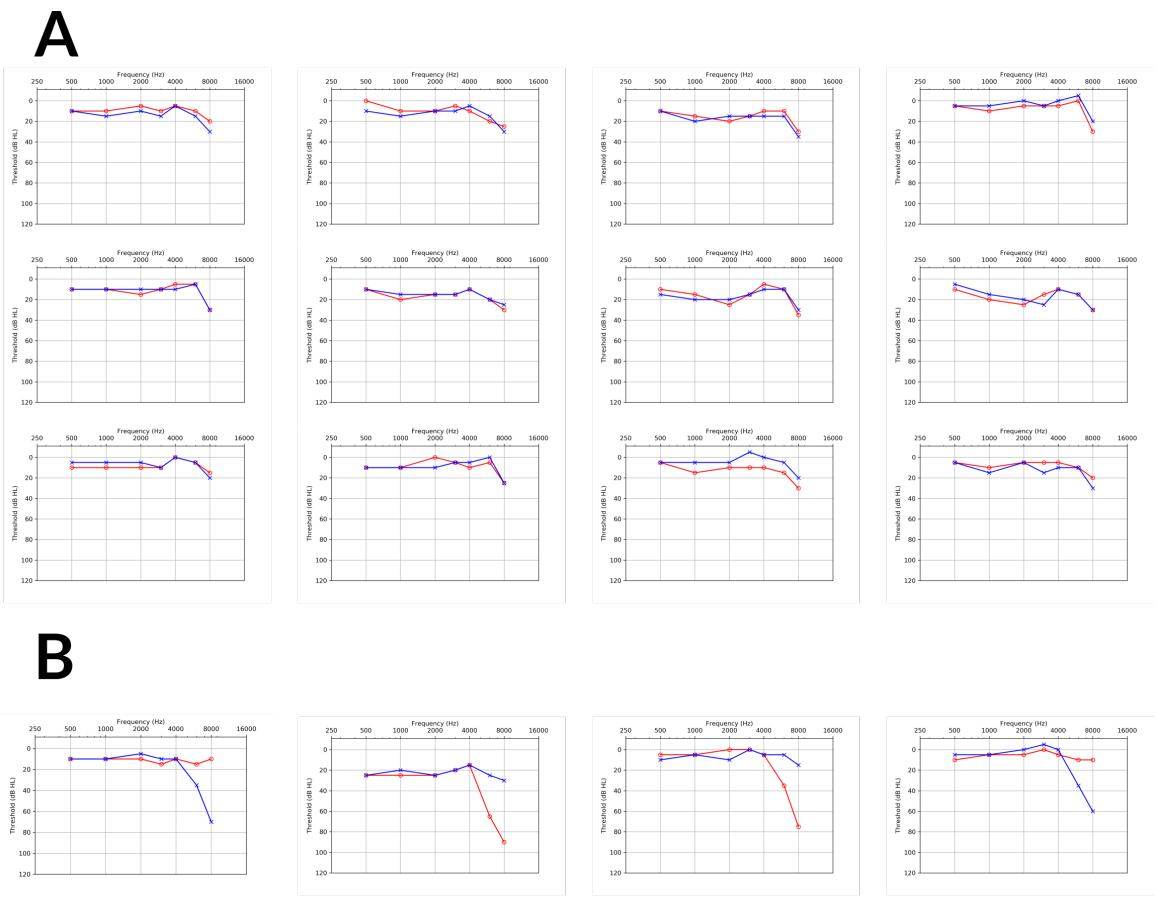


Figure 4.4: Examples of cluster generated by hierarchical clustering
Audiograms from a large cluster of size 12 are shown in (A), while audiograms from a small cluster of size 4 are shown in (B). In both cases, the audiograms within the cluster are very similar on all aspects: configuration, severity and symmetry.

The 240 non-trivial audiograms that were sampled using the greedy sampling: algorithm represent a total of 1712 audiograms, based on the size of their cluster of provenance. This corresponds to a 20% coverage of the space that we defined.

4.6.2 Rater Reliability

As can be seen in Figure 4.5, all three audiologists were self-consistent in classifying the symmetry, configuration and severity of the hearing loss. Only one of the three annotators had *almost perfect* intra-rater agreement in classifying thresholds as belonging

to an audiometric notch or not, while the other two had *substantial* intra-rater agreement. Intra-rater reliability was, in general, poorer for the identification of thresholds with potential reliability issues.

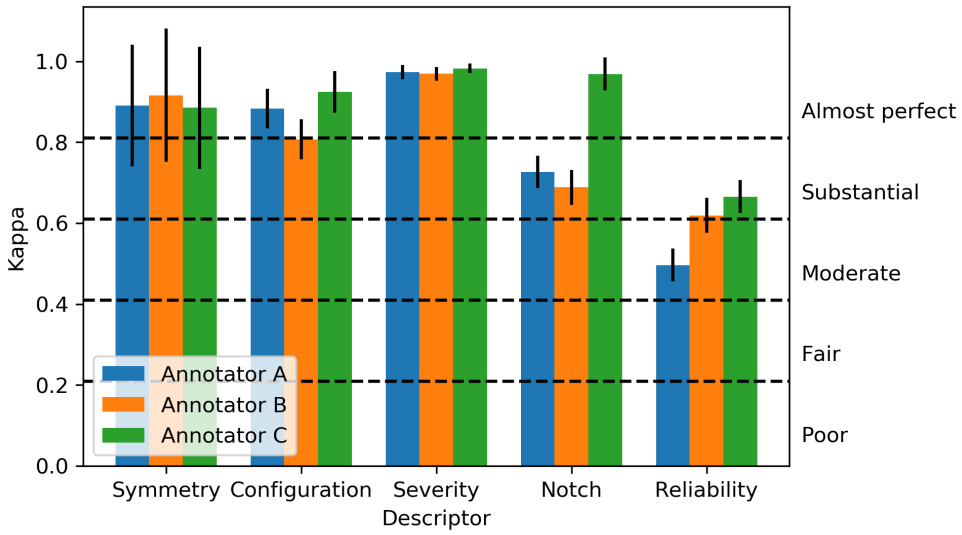


Figure 4.5: Intra-rater reliability over the 5 audiogram annotation tasks
The intra-rater reliability, assessed using Fleiss' κ, was computed using the duplicates presented to the audiologists. The error was computed for all measurements using bootstrap sampling.

While the audiologists had high intra-rater reliability over the assessed tasks, the inter-rater reliability was generally worse. This was especially the case for the identification of audiometric notches ($\kappa = 0.00 \pm 0.06$) and potentially unreliable thresholds ($\kappa = 0.00 \pm 0.06$) in audiograms. For these two tasks, the inter-rater agreement can be considered to be poor. The inter-rater agreement was highest -- almost perfect -- for severity ($\kappa = 0.96 \pm 0.03$). The annotators also agreed often on the symmetry of the hearing loss ($\kappa = 0.82 \pm 0.24$). Unsurprisingly, due to the complexity of the task, configuration classification was associated with a lesser level of agreement than severity and symmetry ($\kappa = 0.55 \pm 0.10$).

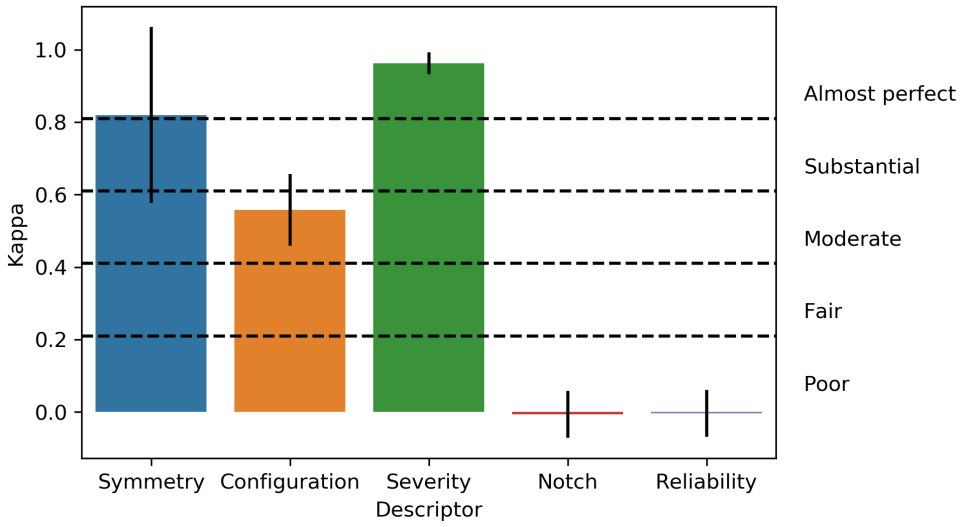


Figure 4.6: Inter-rater reliability over the 5 audiogram annotation tasks
The inter-rater reliability, assessed using Fleiss' κ, was computed over all audiograms presented. For audiogram duplicates, the second annotation was considered to be correct, under the assumption that the second annotation was more "accurate" due to increased familiarity with the annotation process. The error was computed for all measurements using bootstrap sampling.

Unsurprisingly, we found that inter-rater reliability on the audiogram configuration classification task was lower on "challenging" ears that could not be classified with Katz's rules (410 ears; $\kappa = 0.568 \pm 0.08$) than for "trivial" cases where rules could be applied (130 ears; $\kappa = 0.72 \pm 0.06$).

4.6.3 Agreement with Standard Rules

Part of our work involved assessing the extent to which the annotations provided by the audiologists in this study agree with accepted classification rules, more specifically the rules defined in Table 4.1. The agreement between the annotations provided by the audiologists and the aforementioned rules is shown in Table 4.5. These measurements are calculated over all ears which meet the "all-thresholds-within-10-dB" criterion for classification by the rules.

Table 4.5: Agreement between the rules and the audiologist's classifications of audiogram configuration

For all ears where the rules in Table 4.1 could be applied, the resulting configuration annotation was compared with that provided by the audiologist.

Annotator	Fleiss' κ
A	0.55 ± 0.06
B	0.64 ± 0.06
C	0.52 ± 0.06

Using a t-test with equal variance, it's been determined that annotator B agrees with the rules significantly more than both annotators B and C ($p < 0.0001$), although the agreement remains in the "moderate" range for all annotators.

4.6.4 Annotation Analysis

In order to gain more insight into the types of audiograms that constitute the training set, we looked into the breakdown of the annotations for different descriptors.

Looking at the distribution of configurations (Figure 4.7), one can immediately notice that most ears fit within the "sloping" configuration where the hearing loss is more appreciable at higher frequencies. Rarer configurations such as "cookie bite", "reverse cookie bite", and "reverse sloping" were only rarely assigned. Altogether, these observations are consistent with our expectations.

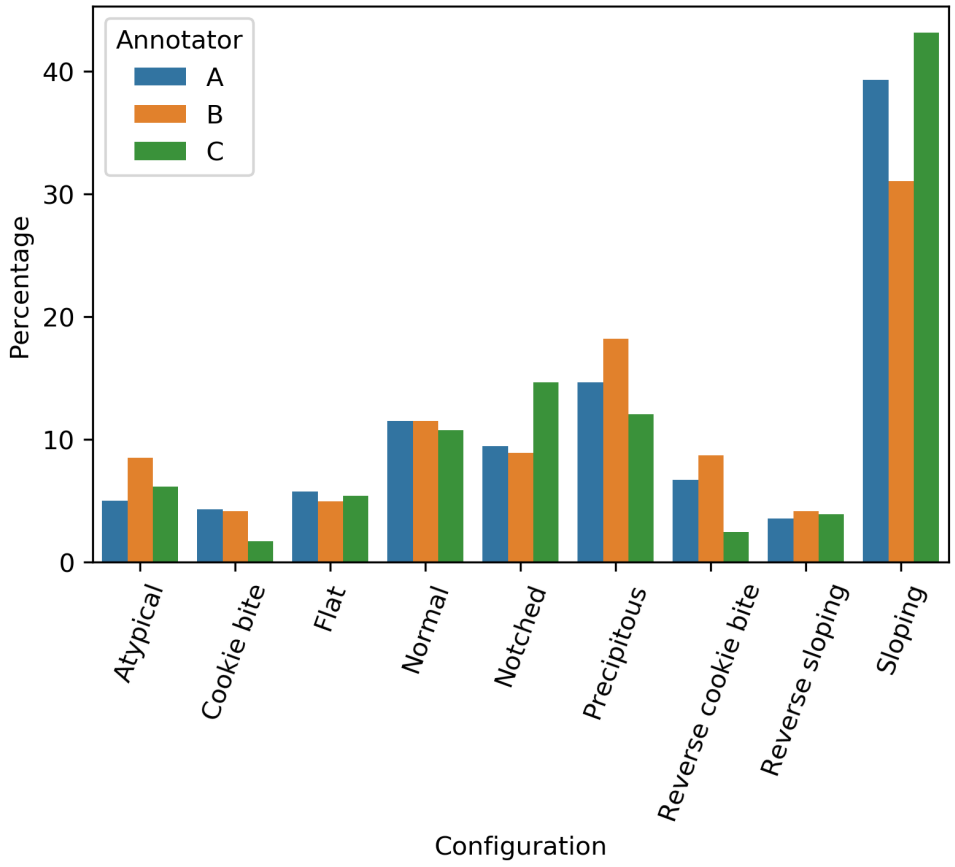


Figure 4.7: Distribution of configurations assigned by all three audiologists
This plot displays, for all possible configuration descriptors, the corresponding percentage of ears to which the descriptor was selected.

With regards to symmetry, we found that the audiograms presented to the audiologists were overwhelmingly asymmetrical (78%, averaged over all three audiologists). We did not request any comments or description of the extent of the asymmetry.

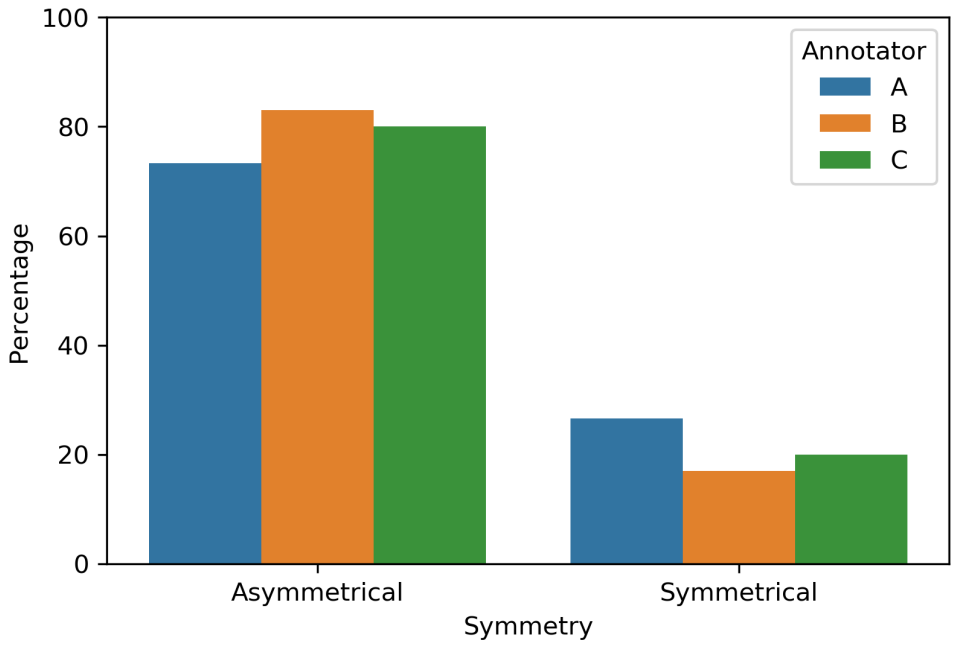


Figure 4.8: Distribution of symmetry descriptors assigned by all three audiologists
The proportion of asymmetrical and symmetrical descriptors assigned to the audiograms is shown here.

Because the severity of the hearing loss is described by a varying number of descriptors for the different frequency ranges, an breakdown similar to what was done for configuration and symmetry appears inappropriate.

4.6.5 Annotation Sessions

We investigated the timing of the annotation process, both in terms of the time required to provide a complete audiogram annotation, but also with respect to the duration of a sitting. This was of interest, as we did not control these variables.

Table 4.6: Temporal information related to the annotation process

We recorded the time at which an annotation is started by means of a "start" button and the time at which the corresponding annotation was submitted to the server. The following information was derived from this data. We considered consecutive annotations occuring within 2 hours to have been completed in the same sitting. Annotations requiring more than 5 minutes to complete were not considered in the calculations of the time required per annotation. Standard deviation are shown in parentheses.

	Annotator A	Annotator B	Annotator C
Number of sittings	8	7	8
Annotations provided per sitting	31 (21)	20 (12)	39 (26)
Duration of a sitting (h)	1.1 (1.0)	0.8 (0.6)	3.8 (7.8)
Time required per annotation (s)	69.4 (41.9)	99.9 (62.7)	53.1 (32.3)

We found that the annotators completed the annotations in 7 or 8 sittings. There was a very large difference between the annotators with respect to the duration or length of a sitting. Suprisingly,

Annotator C provided the most annotations per sitting and had the longest sittings. This not surprising because this annotator almost consistently had higher intra-rater reliability. Being presented an audiogram twice in the same sitting is presumably more likely to lead to the same annotation than if it is presented for a second time days after the first display. This is also indicative that the task may not have been associated with significant fatigue, i.e. decreased intra-rater reliability due to the sessions being too long, at least for Annotator C. Annotator C also spent the least time to per audiogram with 53.1 s on average; this could indicate less hesitation and more confidence in the annotations.

4.7 Discussion

We found that, in general, intra-rater reliability was high, particularly for classification of symmetry, configuration and severity. In the worst case (potentially unreliable threshold identification), the Landis and Koch interpretation indicated moderate agreement. The

intra-rater reliability was slightly lower for the identification of audiometric notches, but still substantial. The poorest intra-rater reliability measurements were observed on the task involving the identification of potentially unreliable thresholds. In another study [71], 5 clinicians were found to have substantial intra-rater agreement in assessing the severity of the hearing loss, whereas we found the agreement to be almost perfect here. The difference is likely attributable to the fact that we insisted that the annotators refer to the same severity assessment scale. We are unaware of any previous intra-rater reliability measurements for all 4 other tasks in the current study.

On the other hand, inter-rater agreement was generally poorer. There was no agreement for the tasks pertaining to the identification of audiometric notches and suspicious thresholds. The participants indicated, in the retrospective meeting, that identifying suspicious thresholds using only air conduction thresholds was particularly challenging. Usually, assessing the quality of the audiogram often requires additional data such as bone conduction thresholds or other tests. Unsurprisingly, the inter-rater agreement was worst for configuration (moderate), when considering the remaining tasks. Agreement about symmetry and severity was almost perfect. We believe that the ill-defined boundary separating the ‘‘flat’’, ‘‘sloping’’ and ‘‘precipitous ’’ configurations generates much of the variability. Margolis and Saly [27] observed slightly better agreement (substantial) for classification of configuration, but worse agreement, although still almost perfect, for the classification of severity. The variation across studies may result from slightly different experimental designs. For example, the configuration and severity schemes used in the two studies, although similar, were not exactly identical. Furthermore, we provided more detailed instructions with respect to severity classification.

We found that the agreement with the standard rules (Table 4.1), where they could be applied, varied between 0.52 (moderate) and 0.64 (substantial), depending on the annotator. This shows that, in general, the annotators’ annotation do not necessarily reflect what is considered to be a standard set of rules. We suspect that this may be

due to the fuzzy boundaries that separate a "flat" hearing loss from a "sloping" loss, and a "sloping" from a "precipitous loss" in human annotators. Audiologists must make a judgement do not necessarily fitting a "curve of best fit" through the thresholds exactly.

It is important to raise two factors that may have introduced some bias in the annotations. First, the audiologists were presented with complete audiograms including both ears. The participants did point out, at the end of the experiment, that seeing the contra-lateral ears may affect the annotation of the ear considered. For this reason, Margolis and Saly chose to only show audiograms of a single ear [27]. We argue here that, in a clinical setting, audiologists refer to the complete audiogram when completing an annotation. They are thus influenced by this source of bias. Furthermore, removing this source of bias completely would have prevented us from collecting annotations regarding audiogram symmetry. Second, the interpretation of an audiogram is likely to be influenced by the visual representation of the audiogram. For example, what may appear as a slightly sloping audiogram on one particular graph, may appear to be sharply sloping in another where the vertical axis is longer or stretched.

In Sections 4.2 and 4.3, we presented our approach for sampling an heterogeneous set of audiograms such that redundancy is minimized. We note, however, that we operated in the "unsupervised" machine learning framework, and consequently, it was not possible to objectively evaluate our approach. Given that the data were unlabeled, we could not use labels to optimize for the parameter α dictating the relative importance of exploring the audiogram space over the sampling of audiograms from large clusters. Furthermore, it was not possible to directly evaluate the quality of our audiogram representation (extracted features). It is conceivable that a different feature set could have led to a totally different number of clusters, and hence, to a different sample of audiograms.

4.8 Conclusions

In this chapter we presented the approach we developed to sample a set of unique audiograms for subsequent annotation by expert audiologists, before proceeding to their analysis.

We assembled a set of 320 audiograms, with 270 distinct audiograms. The audiograms in this dataset were homogeneous in terms of configuration, severity and symmetry, and included duplicates for estimation of intra-rater reliability. The dataset was annotated by 3 professional audiologists by means of the RAAE developed specifically for this purpose.

The overall findings reveal that intra- and inter-rater reliability for tasks pertaining to symmetry, configuration and severity classification were generally good, indicating that it may be possible to automate the task. Unfortunately, there was essentially no inter-rater agreement in what constitutes an audiometric notch or an unreliable threshold. This implies that the dataset of annotations collected is unsuitable for training a learning algorithm for the identification of such audiometric data. Agreement with ''reference'' rules was found to be moderate, prompting a reflection on the applicability of these rules.

5

Supervised Learning for Audiogram Classification

All models are wrong, but some are useful.

-- George Box

In the previous chapter, we presented the methodology applied to assemble a dataset of audiogram annotations for 270 unique audiograms. We also analyzed this dataset, putting significant emphasis on the measurement of the agreement of the audiologists who generated that dataset. That dataset, in and of itself, is of little value -- unless it is leveraged in a way that allows the automatic annotation of audiograms, that is.

This chapter details how insight is extracted from the dataset assembled in the previous chapter, and integrated into an automated system capable of classifying audiograms by configuration, symmetry and by the severity of the hearing loss. This approach departs from previous approaches in that it is purely data-driven; i.e. the classification rules are derived from data, as opposed to manually. Furthermore, by incorporating additional data, the accuracy of the classification system can be improved. In fact, as the training set size increases, the accuracy will tend towards the maximum

achievable accuracy, a value limited by inter-rater reliability.

First, we outline the development of the different modules composing this system which we shall henceforth refer to as the DDAE. Next, we explain how they come together to form a whole before proceeding to an evaluation of its performance and a comparison with the current state of the art.

5.1 Classification Requirements

In the previous chapter, we showed that the intra and inter-rater agreement of audiologists ranged from moderate to almost perfect for classification of audiogram symmetry, configuration and severity. The agreement for the identification of thresholds belonging to a notch or potentially unreliable thresholds was essentially nonexistent. As such, in this chapter, we leave these two descriptors aside to focus our efforts on the description of hearing loss in terms of the former three descriptors.

In contrast with the current state of the art, the classification system developed here was to possess the following properties:

1. **Dimension-independence:** The system can accept a variable number of thresholds. This is enabled by the use of dimension-independent features. This property is important as many audiologists consider certain inter-octave frequencies, such as 3,000 Hz and 6,000 Hz, to be optional, leading to incomplete audiograms in certain cases. As such, the number of measurements varies from one audiogram to the next.

2. **Confidence estimates:** Provides an estimate of the confidence of the labels assigned to the audiograms

3. **Retrainability:** Can easily be retrained as new data become available

4. **Multi-label classification (configuration):** Relieves the assumption that a single configuration descriptor can describe the audiogram

5. **Data-driven:** The classification rules are obtained through the optimization of objective classification accuracy criteria, instead of manually, to prevent the injection of bias

6. **Interpretability:** The classification process should be easily understood by a human

5.2 Problem Formulation

One way -- perhaps the most obvious -- to formulate the classification problem would be to treat this as a classification problem where the possible classes encompass all possible combinations of symmetry, configuration and severity. Assuming that, in most cases, at least two severity labels are necessarily to accurately portray the extent of the hearing loss along the frequency spectrum, this formulation would form a $2\ (symmetry) \times 8\ (configuration) \times 6^2\ (severity) = 576$-class problem. The training assembled in Chapter 4 consists of annotations for 270 unique audiograms. Even assuming a perfectly balanced dataset, this corresponds to less than 1 instance per class. However, we saw previously that the distribution of configuration annotations was largely uneven. This problem formulation is therefore intractable.

Instead, we elected to decompose the problem into three separate, and independent sub-problems, under the assumption that configuration, symmetry, and severity are independent:

1. **Symmetry labelling:** Simple binary classification problem consisting in determining whether the audiogram is symmetrical or asymmetrical. The classes are mutually exclusive in this sub-problem.

2. **Configuration labelling:** Classification of ears by configuration using an ensemble of decision trees, which we term *decision forest* -- not to be confused with the widely known *random forest* classifier. Every tree in the forest is is a binary classifier corresponding to one configuration, whose goal is to determine whether the corresponding configuration accurately describes the audiogram, partially or in its entirety. In this formulation, configurations are not mutually exclusive.

3. **Severity labelling:** Assignment of 1, 2 or 3 severity labels to an ear. The number of labels depend on the configuration.

5.3 DDAE Development

5.3.1 Feature Engineering

In the supervised learning framework that we employ here, a representation, or feature set, that mathematically describes the instances to classify is developed. The learning algorithm then finds patterns and associations within the representation and the target variable. In this section, we present the features that were derived from audiograms for the purpose of classifying audiogram configuration, symmetry and severity.

The reader may wonder why feature engineering is required, because features had already been defined in the previous chapter to define the audiogram space for clustering and sampling. We can motivate a second iteration of feature engineering with three reasons. First, the features developed previously represented a full audiogram (both ears), whereas here, we need to develop a representation for classification of configuration and hearing loss severity which are considered on a ear-per-ear basis. Second, many features derived for clustering and sampling do not necessarily have a clear, interpretable meaning. For example, the normalized threshold at 1,000 Hz means very little to a human as opposed to the slope of the line of best fit, or the curvature of

the audiogram. Third, the features derived previously assume that 7 specific thresholds are provided, which may not be the case. In other words, the features should be somewhat independent of the number of thresholds measured, although we do make the reasonable assumption that at least 4 major thresholds will be provided (500 Hz, 1,000 Hz, 2,000 Hz and 4,000 Hz).

Configuration Features

We derived a number of features that can be derived to classify the configuration of an ear. The features are presented in Table 5.1.

Table 5.1: Features defined for the configuration classification models

A set of features describing audiograms on an ear-by-ear basis was defined. The features are all independent of the number of thresholds measured, as long as there are at least 4 thresholds per ear.

Feature	Description
1	Slope of the line of best fit
2	Proportion of positive slopes joining consecutive thresholds
3	Proportion of negative slopes joining consecutive thresholds
4	Maximum threshold (worst threshold)
5	Minimum threshold (best threshold)
6	Average threshold
7	Standard deviation of the thresholds
8	Average of thresholds in the low frequency range (below 1,000 Hz)
9	Average of thresholds in the mid frequency range (between 1,000 Hz and 3,000 Hz)
10	Average of thresholds in the high frequency range (4,000 Hz and above)
11	Proportion of slopes that change signs with respect to the previous slope
12	Mean absolute residual from the line of best fit
13	Audiogram curvature; highest-order coefficient of the quadratic of best fit
14	Audiogram range; difference between the maximum and minimum thresholds
15	Notch index [70]

We derived a total of 15 features derivable from the thresholds of an ear. To our knowledge, these features were not used by other groups. Given the relatively small size of the training dataset, it was clear that many of these features may not be used at all in the algorithm. In general, the size of the dataset should scale with the number of

features to prevent overfitting [72]. Furthermore, most of these features were derived to be easily interpretable, but it was unclear whether they were discriminative or not before training the model, and consequently whether they were useful for the model. Given that decision trees effectively perform a form of entropy-based "feature selection" during training, we chose to derive a large set of features, and to let the decision trees identify the ones that are most discriminative.

Symmetry Features

To guide our development of a classification classification rule set, we referred to an existing symmetry classification system [27]. We selected a set of 6 features described in Table 5.2.

Table 5.2: Features defined for the symmetry classification model

A set of 6 dimension-independant features were derived to develop a symmetry classification model.

Feature	Description
1	Maximum inter-aural threshold difference
2	Minimum inter-aural threshold difference
3	Average inter-aural threshold difference
4	Average inter-aural threshold difference
5	Difference in the slopes of the lines of best fit
6	Difference between the average threshold across ears

Severity Features

We defined a series of features for severity to be used for severity classification. These features are presented in Table 5.3.

Table 5.3: Features defined for severity classification

A total of 13 features were used to classify the severity of the hearing loss across the frequency spectrum. We define the low frequency range as all frequencies below 1,000 Hz, the midrange as frequencies between 1,000 Hz and 3,000 Hz, inclusively, and the high range as frequencies greater or equal to 4,000 Hz. These range definitions are consistent with was can be found in the literature [28, 73].

Feature	Description
1	Average threshold
2	Maximum (worst) threshold
3	Minimum (best) threshold
4	Average of thresholds in the low range
5	Maximum (worst) threshold in the low range
6	Minimum (best) threshold in the low range
7	Average of thresholds in the mid range
8	Maximum (worst) threshold in the mid range
9	Minimum (best) threshold in the mid range
10	Average of thresholds in the high range
11	Maximum (worst) threshold in the high range
12	Minimum (best) threshold in the low range
13	Maximum (worst) threshold in notch-susceptible frequencies (between 3,000 and 6,000 Hz, inclusively)

5.3.2 Model Training

Configuration

To account for the possibility that multiple configurations may accurately represent all, or a portion of an audiogram, we used a committee of decision trees (forest). Each tree in the forest corresponds to one of the 8 possible configurations, and is a binary classifier whose task is to determine whether the corresponding configuration adequately describes the audiogram or a portion of the audiogram. A decision can then be made based on the output of the trees in the forest. The design of the configuration classification pipeline is illustrated in Figure 5.1.

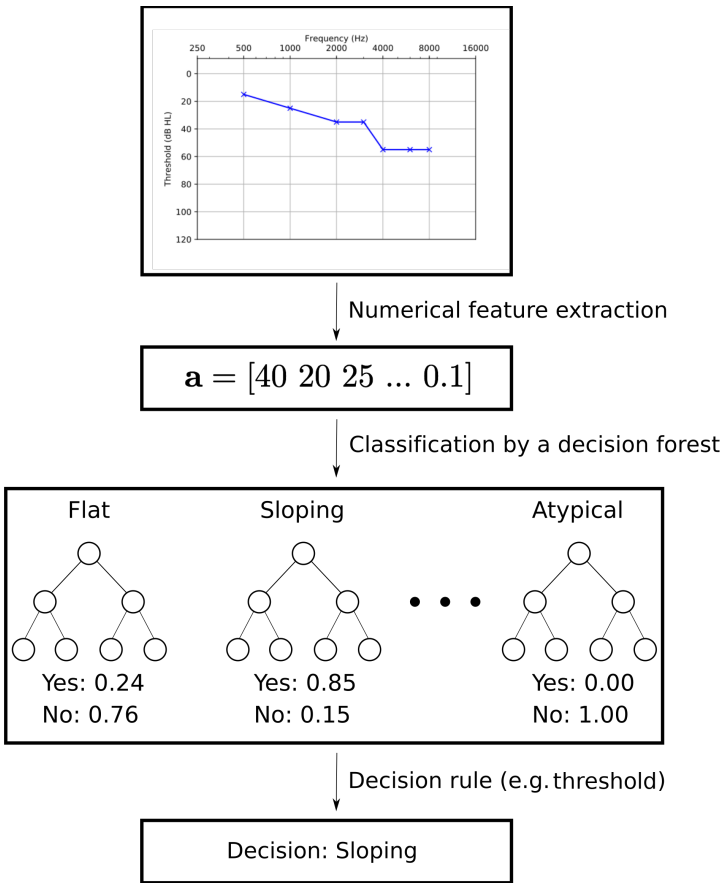


Figure 5.1: Configuration classification pipeline
The configuration classification pipeline is illustrated here. Numerical features are extracted from the audiogram to form the feature vector input to the decision forest composed of 8 decision trees. Each tree outputs the probability that the corresponding label accurately describes the audiogram.

First, a 270×15 design matrix, A, was built, by extracting features for all 270 annotated audiograms:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \ldots & a_{1,15} \\ a_{1,1} & a_{2,2} & a_{2,3} & \ldots & a_{2,15} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{270,1} & a_{270,2} & a_{270,3} & \ldots & a_{270,15} \end{bmatrix} \tag{5.1}$$

where $a_{i,j}$ is the the value of feature j (see Table 5.1) of the audiogram i.

The training process is almost identical for all 8 decision trees. The only distinction is the label vector $\mathbf{t}$ used during the training procedure. For every model, a target vector was obtained:

$$\mathbf{t_c} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_{270} \end{bmatrix} \tag{5.2}$$

where the entry t_i takes the value 1 if the configuration $c \in \{flat, sloping, precipitous, ...\}$ was selected by at least 1 annotator for the ith audiogram, and 0, otherwise. Clearly, this definition relies upon the assumption that all configurations labels provided by the audiologists are correct and equally valid.

Generally, it is desirable to perform some form of feature selection prior to training a learning algorithm. This can be achieved through a variety of means. Given that our approach relies on the application of decision trees, this step is not necessary, because it is implicitly performed during the training of the decision trees. Decision trees iteratively select the features that provide the highest information gain, or decrease the entropy the most, with respect to the class. For these reasons, no feature selection was performed.

Using the scikit-learn [74] implementation of the CART decision tree learning algorithm and information gain as the quality criterion, we trained the decision trees using 3-fold class-stratified cross-validation. The same design matrix was used in all cases, but the target vector was changed appropriately for each configuration. The use of 5-fold or 10-fold cross-validation is often seen in machine learning literature, but given the low prevalence of certain configurations, that may result in folds with too few instances of the positive class. We tuned the hyperparameters controlling the depth of the tree and the number of features using grid search (exhaustive search of hyperparameter combinations), and selected, for all trees, the model that maximized the average F_1

score over the three folds. Both parameters were varied between 1 and 5, to limit the odds of overfitting and for facilitating the interpretation of the resulting rules.

Finally, we trained the model on the entire dataset using the optimal set of hyper-parameters.

Note that, in the scikit-learn implementation of the decision tree algorithm, a decision tree can output a score for a prediction. This score is simply the fraction of the training samples of majority class in the leaf node corresponding to the test instance [74].

Symmetry

Given that annotators were not asked to provide details on the extent of the asymmetry in this study, classification of audiogram symmetry collapses to a binary classification problem. This problem calls for a single decision tree, instead of a forest.

Much like for configuration, we defined a 270×5 design matrix A whose rows represent the audiograms in the training set, and whose columns represent the features presented in Table 5.2. To build the target vector, a simple majority rule was implemented. Given that there were three audiologists and only two possible labels, we simply took as ground truth the label that was assigned by at least 2 audiologists. One audiologist failed to annotate 20 audiograms, and as such, a subset of audiograms were only annotated by 2 audiologists. In all but one case, the 2 other audiologists had assigned the same label. That instance was removed from the design matrix and target vector.

The decision tree for symmetry classification was trained in the same way as the decision trees for configuration classification, with one exception being that 5-fold cross-validation was used instead of 3-fold cross-validation.

Severity

Normally, to obtain a severity descriptor, one would first compute the PTA, and look in a reference table (Table 2.3) for the correct descriptor (e.g. mild) corresponding the decibel value. In reality, a single severity descriptor is only sufficient in "flat" hearing losses. For "sloping", "precipitous" and "reverse sloping" audiograms, two descriptors are required to describe the severity in the best and worst ranges (lows and highs or vice-versa) across the frequency spectrum. For "cookie bite", "reverse cookie bite", "notched" and "atypical" audiograms, it is very difficult to accurately convey the extent of the loss across the spectrum with less than three descriptors.

As opposed to the approach used to classify configuration configuration and symmetry, decision trees are not used to label the severity of the hearing loss. Here, we use a simple "feature selection" approach where we simply seek to maximize agreement with the audiologists.

For the "flat" configuration, a single severity descriptor is sufficient. We wish to determine what audiogram feature, when looked up in a reference table, will most often reproduce the label provided by the audiologists. For all audiograms, we simply perform table lookups for all features in Table 5.3. If table lookups for *Feature 1* most often reproduces the label provided by the audiologists for "flat" audiograms, then *Feature 1* is taken to be the best feature. Subsequently, when the system is presented a "flat" audiogram, it will extract *Feature 1*, and perform a table lookup with that value to identify the severity label.

For the "sloping", "precipitous" and "reverse sloping" configurations, we require two descriptors. We therefore identify, for every configuration, the best feature to use in lookups for the descriptor corresponding to the "lows" and the best feature for the severity descriptor corresponding to the "highs".

For the remaining configurations, the procedure is identical, except that we identify the best feature for the "mids" in addition to the "lows" and the "highs".

In short, the "training" or "feature selection" procedure for the severity classification system simply consists in identifying the feature that most often produces the label provided by the audiologists with a table lookup for all configurations, and for all relevant frequency ranges associated with the configurations. Please refer to Tables D.2 to D.9 in Appendix D for the percentage agreements obtained with the different features.

The classification process for a sloping audiogram, following the identification of the best features for the "lows" and the "highs" is illustrated in Figure 5.2.

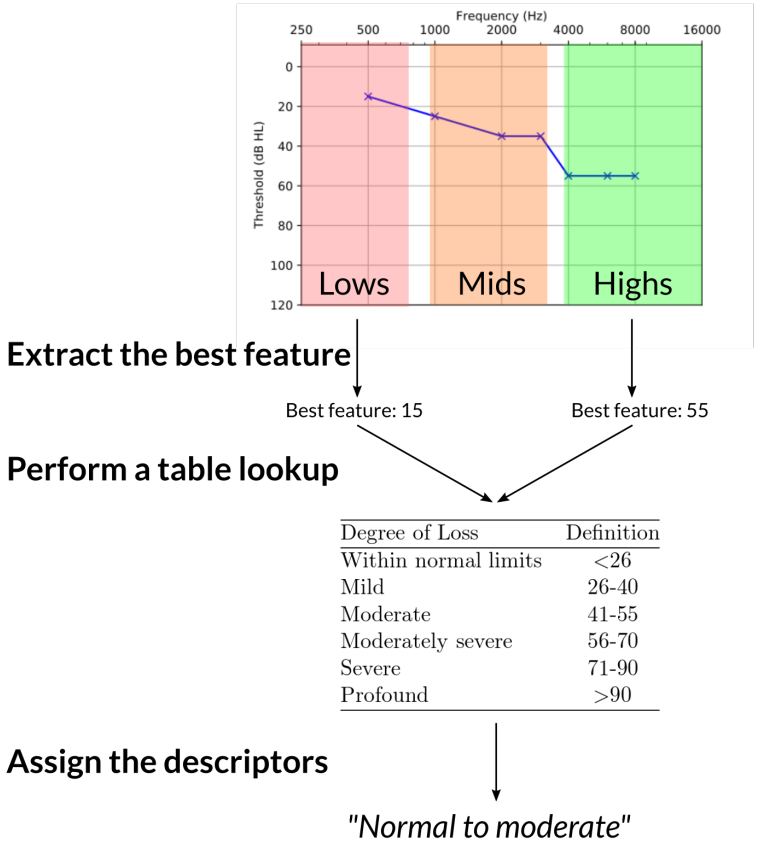


Figure 5.2: Severity classification process
Classification is done by extracting the feature whose lookup in the reference scale achieved the highest agreement with the annotations in the training set. In the example in this figure, the sloping audiogram is adequately described with two descriptors, one for the low range and one for the high range.

5.4 System Integration

In the previous section, the classification problem formulations for classification of configuration, symmetry and severity were presented. The methodology applied to develop classification models was also discussed. Now, one may wonder how these different components come together to form a unified classification system. The overall architecture of the classification system is presented in Figure 5.3, and consists of three distinct modules corresponding to the the three descriptors.

The first module, is concerned with the classification of configuration and operates on an ear-by-ear basis; it is thus run twice for each audiogram. For both ear, the audiogram attempts to classify the audiogram using the rules presented in Table 4.1. If unsuccessful, features related to audiogram configuration (Table 5.1) are extracted and input to the decision forest, which then assigns a probability for every possible configuration. If some decision thresholds established in advance are met by one configuration, this configuration is output. If multiple configurations are output, the ''simplest'' one is selected, as done in Carhart's audiogram classification system [42]. Configurations that are most similar to ''flat'' get precedence (e.g. sloping or reverse sloping over cookie bite). In the event that none of the configuration probabilities exceed the threshold, the audiogram is classified as ''atypical''.

The second module, concerned with the assignment of severity labels, relies on the output of the first module. For every ear, the module accepts the configuration(s) output by the first module to determine whether the hearing loss should be labeled with one, two or three severity descriptors. For every configuration input, the best features are extracted and table lookups performed to assign the severity labels. The module thus returns the a set of severity labels for every configuration that is passed as an input.

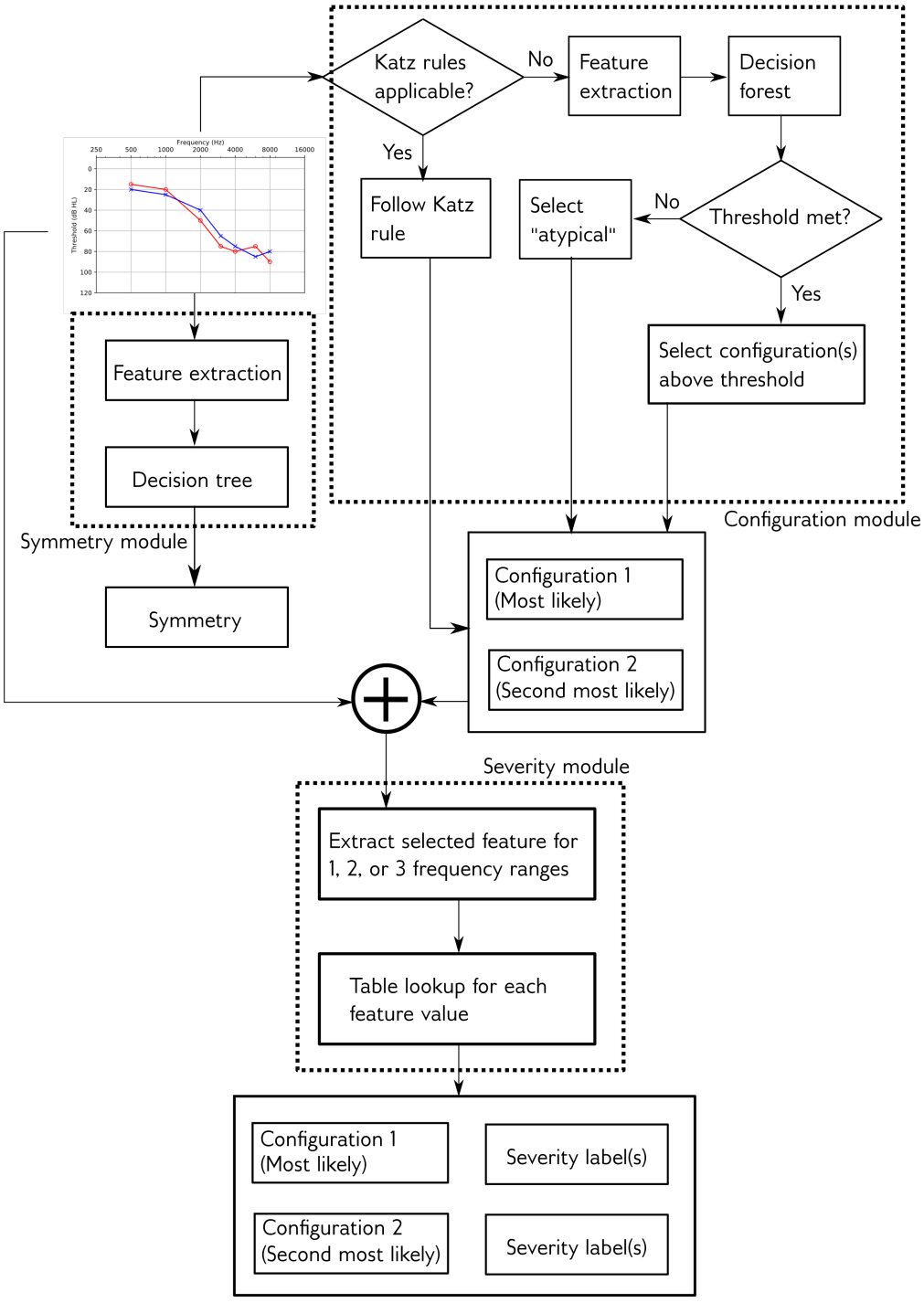


Figure 5.3: Flow chart of the DDAE architecture
This flow chart illustrates the inner workings of the different classification modules as well as their inputs and outputs. The configuration and severity modules operate on an ear-by-ear basis whereas symmetry requires a complete audiogram with thresholds for two ears.

The third module, which classifies audiograms as symmetrical or asymmetrical, operates in a binaural basis, because symmetry is undefined if measurements are only performed for one ear. This module operates independently from the other two modules.

5.5 Comparison with the State of the Art

We compared our algorithm with AMCLASS™ system which is currently the state of the art. The authors of the algorithm developed a web interface through which one can submit audiograms for classification (Figure 5.4). The interface is illustrated below:

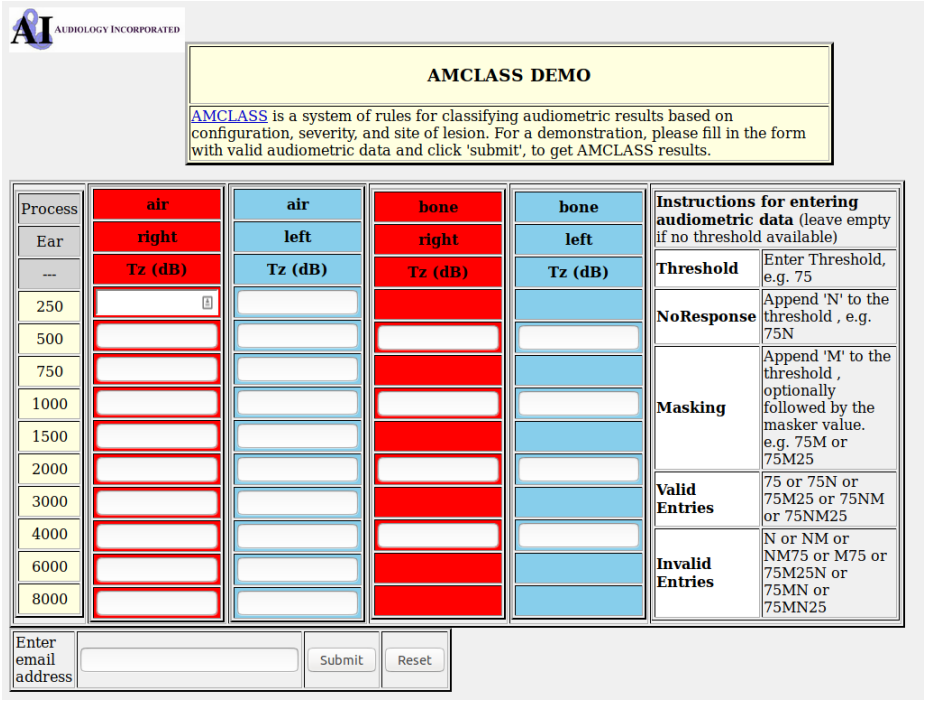


Figure 5.4: The AMCLASS™ web interface
The AMCLASS™ web interface allows users to submit an audiogram for classification. The user may provide the pure tone threshold values by means of a simple web form.

Given that the classification results for 270 unique audiograms had to be retrieved from the interface, we judged that the automating the task was justified. We used the

Selenium library [75] for Python script to develop a script that fills out the form with the thresholds of the audiograms in our dataset and downloads the algorithm's output.

The algorithm outputs a graphical representation of the audiogram and its classification as an image the JPEG format (Figure 5.5). As such, it was not possible to retrieve the classification directly.

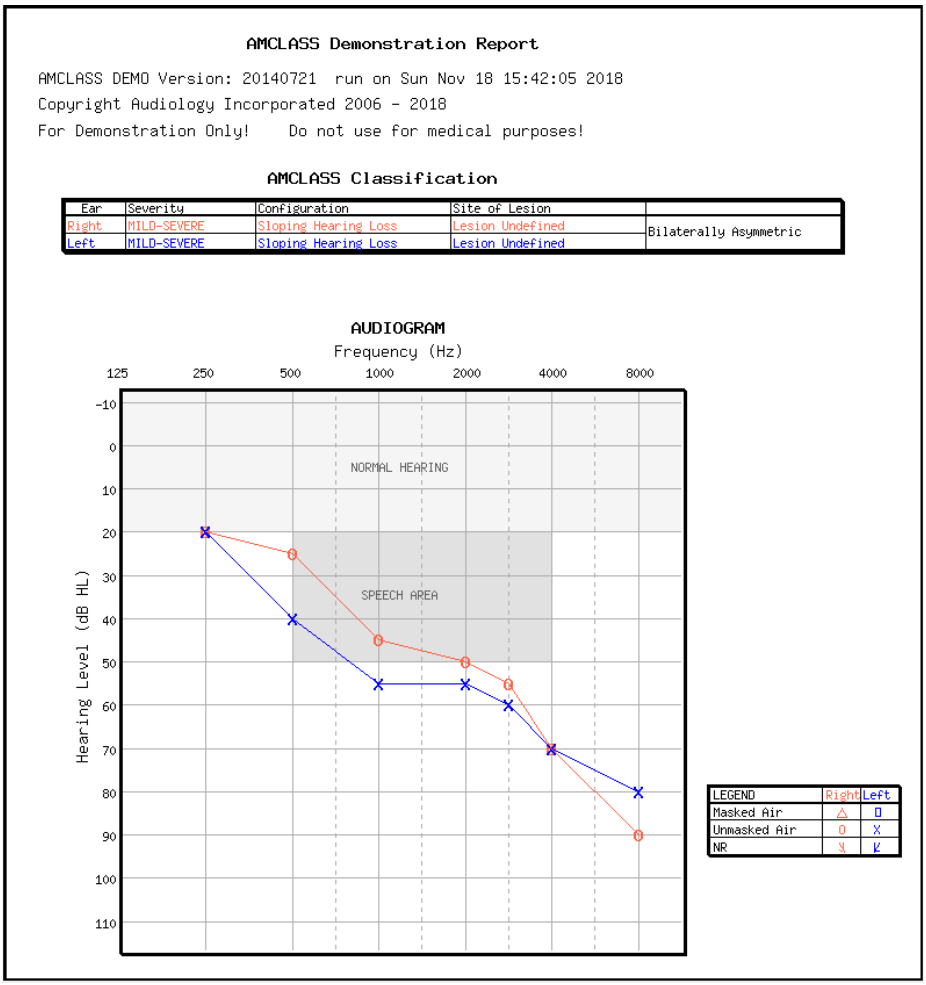


Figure 5.5: The output of the AMCLASS™ web interface
The AMCLASS™ web interface outputs an image containing a visual representation of the audiogram, some metadata (top) and a table with the classification results for both ears.

We used template matching using the OpenCV library [76] to extract the classes from the image for all of the 270 audiograms in our training set. This was simplified by

the fact that the positioning of the class labels in the image was fixed.

5.6 Results

5.6.1 Configuration Classification Performance

We evaluated the performance of each individual decision tree in the decision forest for configuration classification Table 5.4. The performance in terms of accuracy, recall (sensitivity), precision and F_1 score are presented.

Table 5.4: Performance of the configuration classification module

The 3-fold cross-validation accuracy of the trees in the decision forest is presented. A positive prediction is considered correct if at least one audiologist assigned the label to the audiogram. The performance is shown for a decision threshold set at 50%, meaning that a decision tree reports a positive if it is more than 50% confident.

Configuration	Imbalance	Accuracy	Recall	Precision	F_1
Flat	1:10.0	0.94 ± 0.02	0.84 ± 0.10	0.80 ± 0.05	0.81 ± 0.07
Sloping	1:1.0	0.81 ± 0.04	0.81 ± 0.04	0.81 ± 0.04	0.81 ± 0.04
Precipitous	1:3.3	0.88 ± 0.01	0.82 ± 0.03	0.85 ± 0.02	0.83 ± 0.01
Reverse sloping	1:20.6	0.96 ± 0.01	0.83 ± 0.06	0.81 ± 0.05	0.81 ± 0.01
Cookie bite	1:19.0	0.95 ± 0.01	0.62 ± 0.02	0.74 ± 0.06	0.65 ± 0.01
Reverse cookie bite	1:9.0	0.93 ± 0.01	0.80 ± 0.06	0.83 ± 0.05	0.81 ± 0.04
Notched	1:5.5	0.84 ± 0.04	0.71 ± 0.04	0.70 ± 0.06	0.70 ± 0.05
Atypical	1:6.3	0.84 ± 0.04	0.61 ± 0.03	0.70 ± 0.13	0.62 ± 0.04

The classification performance is rather uneven across configurations. While we would expect the performance to be lower on classes with high imbalance, this association is not clear in this case. For example, the binary classifier for the "reverse sloping" configuration performs very well in spite of the large class imbalance (1:21) for this configuration. A similar argument can be made for the atypical configuration, which has poor performance despite a relatively smaller class imbalance (1:6). However, the low class imbalance of atypical audiograms is overpowered by the "difficulty" of the concept. In fact, the "atypical" descriptor represent a very large set of configurations

that is much more heterogeneous than all other descriptors.

We also computed the performance of the AMCLASS™ system, the current state of the art, on the configuration annotation task. Given that the configuration schemes used in this study and that of Margolis and Saly [27] differ slightly, we established the correspondance pairs shown in Table D.1 in Appendix D. To make a direct comparison possible, a prediction by AMCLASS™ was considered to be correct if at least one of the audiologists provided an equivalent annotation, as defined by the equivalence pairs.

Table 5.5: Performance of the AMCLASS™ classification system on the training set

We measured the configuration classification performance of AMCLASS™. As was done for the DDAE performance evaluation, a prediction by the rules was considered correct if at least one audiologist assigned an equivalent configuration label. Given that the AMCLASS™ system does not require training, its performance was estimated from 1000 subsamples of size 150 with the bootstrap method.

Configuration	Imbalance	Accuracy	Recall	Precision	F_1
Flat	1:10	0.90 ± 0.02	0.66 ± 0.14	0.47 ± 0.12	0.52 ± 0.11
Sloping	1:0.6	0.81 ± 0.03	0.80 ± 0.04	0.88 ± 0.03	0.84 ± 0.03
Reverse sloping	1:21	0.81 ± 0.03	0.80 ± 0.04	0.88 ± 0.03	0.84 ± 0.03
Cookie bite	1:19	0.96 ± 0.02	0.37 ± 0.20	0.65 ± 0.29	0.44 ± 0.19
Reverse cookie bite	1:9	0.93 ± 0.02	0.66 ± 0.13	0.64 ± 0.15	0.65 ± 0.12
Other	1:2.8	0.75 ± 0.04	0.09 ± 0.04	0.62 ± 0.23	0.15 ± 0.07

We performed hypothesis testing to determine whether there is a statistically significant difference between the performance of the DDAE and AMCLASS™, using the F_1 score as a basis for comparison. More specifically, we compared the performance of the two systems using Welch's t-test on configurations where there is a perfect correspondence between the configurations of the two systems. The annotations made by DDAE were associated with a statistically significant (with a significance level of 0.05) increase in F_1 score for the "flat" configuration ($p < 0.01$), and decrease for the "reverse sloping" configuration ($p = 0.04$). Other differences for the "cookie bite" ($p = 0.12$) and "reverse cookie bite" ($p = 0.06$) configuration were not statistically significant. A direct comparison for the "sloping" configuration was not possible, given that AMCLASS™ does

not make a distinction between sloping and preciptious, while the DDAE does.

5.6.2 Symmetry Classification Performance

The decision tree trained to classify audiograms by symmetry performs slightly better than the AMCLASS™ symmetry classification rule (Table 5.6), although statistical significance could not be achieved ($p = 0.83$).

Table 5.6: Performance comparison between the DDAE and AMCLASS™ for symmetry classification

The performance of the DDAE was computed alongside AMCLASS™. The performance evaluation was done through 5-fold stratified cross-validation for DDAE, and through bootstrap sampling for AMCLASS™ with 1000 bootstrap samples of size 150. Here, the class "asymmetrical" was considered to be a "positive", because this type of loss tends to be more concerning.

Configuration	Accuracy	Recall	Precision	F_1
DDAE	0.98 $\pm$ 0.02	0.94 $\pm$ 0.05	0.99 $\pm$ 0.01	0.96 $\pm$ 0.04
AMCLASS™	0.89 $\pm$ 0.02	1.00 $\pm$ 0.00	0.88 $\pm$ 0.03	0.93 $\pm$ 0.01

5.6.3 Severity Classification Performance

By selecting the features that maximize the agreement between the label provided by a table lookup and the audiologists, we were able to obtain an agreement of 90% or more for most configurations and frequency ranges. In other words, by extracting the selected features and doing a table lookup in a reference table, we could predict the annotators' judgement 90% of the time. The percentage agreement differs from accuracy in that there is no ground truth label. Here, we merely measure how frequently a table lookup for the best feature leads to a prediction that agrees with the labels assigned by the audiologists.

Table 5.7: Percentage agreement between the DDAE's severity predictions and the annotations

For every configuration, the optimal feature for every relevant frequency range was determined. The proportion of table lookups whose result agree with the severity labels provided by the audiologists are indicated.

Configuration	Lows	Mids	Highs
Flat	0.897	N/A	N/A
Sloping	0.926	N/A	0.965
Precipitous	0.947	N/A	0.976
Reverse sloping	0.948	N/A	0.948
Cookie bite	0.978	1.0	0.956
Reverse cookie bite	0.920	0.987	0.920
Notched	0.988	0.723*	0.594
Atypical	0.841	0.593	0.714

* This descriptor represents frequencies susceptible to host audiometric notches.

Given that the ways through which the DDAE and AMCLASS™ assign severity descriptors differ significantly, a direct comparison of performance was deemed to be inappropriate. For example, the DDAE assigns three severity descriptors per audiogram if the audiogram is ''cookie bite'', ''reverse cookie bite'', ''notched'' or ''atypical'' while AMCLASS™ only assigns one.

Surprisingly, the agreement between the prediction and the actual severity label for the ''flat'' configuration is rather low when compared to more ''complex'' configurations. The agreement between the predictions and the labels assigned by the audiologists is worst for ''atypical'' audiograms, as expected. This observation is not unexpected, as ''atypical'' audiograms encompass a large variety of different shapes and the specific features considered in the assignment of severity labels vary from one audiogram to the next.

5.6.4 Model Analysis

We sought to uncover what features are considered to be most important by the decision trees. This information can reveal the extent to which the concepts learned by purely

data-driven learning algorithms align with the intuition of a human interpreter; expert or not.

To quantify the "importance" of the features from the perspective of each tree in the forest, we computed the Gini importance [77] of all features. While trees were trained using the information gain as a criterion to split nodes, Gini importance is a closely related quantity that also measures the discriminability of a feature, but in terms of a reduction in sample impurity. The Gini importance values of all feature from the perspective of each tree are presented in Table 5.8.

In general, it appears that the features determined to be most important by the trees are intuitive, but not always. For example, one might expect that the decision tree responsible for classifying an audiogram as "flat" or "not flat" would learn that the slope of the line of best fit of a flat audiogram should be near zero. The slope, however, appears to bear little weight in the decision. The decision appears instead to be mostly driven by the range of threshold values. In other words, it is as though it has learned that audiograms with a small threshold range are most likely to be classified as flat by our audiologists. While this may not be immediately intuitive, it provides a different perspective into what is most important based on the data.

One can also observe the difference in concepts learned for the "sloping" and "precipitous" audiogram configurations. The decision to classify an audiogram as "sloping" or not depends mostly on the slope of the line of best fit and the proportion of slopes between consecutive thresholds that are positive. In the "precipitous" case, we might expect similar results because a "precipitous" audiogram differs from a "sloping" audiogram by the rate of change in threshold as a function of frequency. Surprisingly, these features are considered irrelevant by the classifying, which instead, considers the thresholds range to be much more important. It has learned that audiograms with a large thresholds range are very likely to be classified as "precipitous".

Unsurprisingly, however, audiogram curvature is the most important feature in

determining whether an audiogram is likely to be classified as "(reverse) cookie bite" or not. Also, the notch index, an index developed previously to measure the depth of a notch appears to be important in determining whether the audiogram is likely to be classified as "notched". The notch index is higher in notched audiograms, and lower in flatter audiograms.

Taken together, we see that the features considered to be most important are largely intuitive to a human, although their combination into complex rules may not necessarily be so (see Figures D.1-D.8 in Appendix D).

5.7 Discussion

Here, we presented a classification system capable of classifying audiograms by configuration, symmetry and severity. We also evaluated the individual modules in the system.

Configuration is determined by a decision forest, or an ensemble of trees, all of which look at the probability that a particular configuration accurately describes the audiogram, completely or partially. We evaluated the performance of each individual tree in the forest using the F_1 score as the main performance metric. We found that the F_1 score of the trees varied between 0.62 (atypical) to 0.83 (precipitous). Interestingly, we found that class imbalance did not explain variability in the F_1 score to the extent that we anticipated. For example, even if the "reverse sloping" configuration is notably uncommon, its corresponding decision tree performs particularly well. This is likely because this concept is relatively straightforward as opposed to "flat". The "flat" configuration may allow for some deviation from the mean value, for example, but by how much exactly? In fact, a large part of the decision to classify an audiogram as "reverse sloping" relies on the slope of the line of best fit. Intuitively, we would expect the slope to be negative (rising on the audiogram), and this is what the tree appears

Table 5.8: Importance of the 15 features used in configuration classification

The Gini importance values, commensurate with the information gain, for every feature used to train the trees in the decision forest for configuration classification are presented. The top 3 features for every configuration are in bold font.

	Flat	Sloping	Precipitous	Reverse sloping	Cookie bite	Reverse cookie bite	Notched	Atypical
Slope of LOBF[1]	0.07	**0.44**	0.00	**0.58**	0.00	**0.30**	0.00	**0.42**
Proportion of positive slopes	0.00	**0.50**	0.00	**0.37**	0.00	0.00	0.00	0.13
Proportion of negative slopes	0.00	0.00	0.00	0.00	**0.09**	0.00	0.00	0.00
Maximum threshold	**0.22**	0.00	0.00	**0.05**	**0.08**	0.00	0.00	0.00
Minimum threshold	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Threshold average	**0.16**	0.02	0.00	0.00	0.00	**0.09**	**0.24**	0.00
Threshold standard deviation	0.00	**0.05**	**0.09**	0.00	0.00	0.00	0.00	0.00
Mean threshold in the lows	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mean threshold in the mids	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mean threshold in the highs	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Proportion of direction changes	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mean absolute residuals of LOBF	0.00	0.00	0.00	0.00	0.00	0.00	**0.34**	0.13
Curvature of the QOBF[2]	0.00	0.00	0.00	0.00	**0.76**	**0.55**	0.07	**0.15**
Threshold range	**0.51**	0.00	**0.91**	0.00	0.00	0.02	0.00	**0.17**
Notch index	0.00	0.00	0.00	0.00	0.07	0.05	**0.35**	0.00

[1] Line of best fit

[2] Quadratic of best fit

to have learned. The tree responsible for recognizing ''atypical'' audiograms performs the worst. This is likely due to the fact that there is no single definition of what an atypical audiogram is. It is thus reasonable to expect that this is a difficult concept for an algorithm to learn, particularly when the complexity of the decision tree is limited and the dataset is small. One possibility to deal with this issue would be to remove the ''atypical'' tree from the decision forest and to classify all audiograms for whom no probability meets the minimum decision threshold as ''atypical''.

In the system presented here, In deployment, we envision that the system could return multiple classifications per ear, ordered probabilistically.

We compared the F_1 score obtained by DDAE in cross-validation with that of AMCLASS™ over bootstrap samples. We found that the DDAE has a statistically better performance in classifying an audiogram as ''flat'' or not, while AMCLASS™ does better for ''reverse sloping'' audiograms. Other score differences, where configuration correspondance existed between the two systems, i.e. for cookie bite and reverse cookie bite, were not statistically significant.

The DDAE performed slightly better (0.96) in terms of symmetry classification than AMCLASS™ (0.93) in terms of F_1 score, although statistical significance could not be achieved. Interestingly, the annotators mentioned that they were trained to use the rule applied in AMCLASS™ to classify symmetry in school. The fact that AMCLASS™ does not achieve a perfect accuracy suggests that the annotator do not apply this rule systematically. Overall, the high accuracy of the both approaches to symmetry classification seem to indicate that classification of audiogram symmetry is almost trivial.

Our severity label assignment algorithm relying on table lookups for pre-selected features provided labels that agreed with the annotations between 59% and 99% of the time. Severity assessment through our method was least successful for the ''notched'' and ''atypical'' configurations. Surprisingly, severity annotations for ''flat'' audiogram is rather challenging. We might expect audiologists to assign the severity corresponding to

the mean or the worst threshold, but the severity label obtained through a lookup for the optimal feature (mean threshold) only agrees with the annotation 71% of the time.

Although we did not explore this here, one advantage of our classification system is that it allows users to optimize the decision threshold, or in other words, to set a minimum confidence level that must be met for a configuration to be assigned. By default, this threshold is set to 50%, but this could be optimized, and we shall leave this as an future research direction.

For the sake of transparency, we wish to raise the fact that the comparison between the DDAE and AMCLASS™ may be biased. One could argue that the DDAE was trained using the annotations provided by the audiologists which annotated the dataset whereas AMCLASS™ was developed from annotations provided by other audiologists. In this way, the the DDAE may overfit to the test data, highlighting the importance of using cross-validation as done here.

One of the requirements of this study was to create a system that is highly interpretable. For this reason, we limited the maximum depth of the decision trees to 2 for classification of configuration. Given more data, and assuming that this was not such a priority, one could grow deeper decision trees or train different classification algorithms such as support vector machines or neural networks.

We wish to emphasize that the components of the system presented here were trained on a small data set. Significant effort was placed on selecting a dataset that was broadly representative of the entire space of expected audiograms. Additional resources will enable the collection of larger datasets covering more of the audiogram space (i.e. drawing samples from a greater number of distinct clusters) and from obtaining more annotations per audiogram, resulting in greater confidence in the acquired labels.

5.8 Conclusions

In this chapter, we presented the DDAE, a classification system capable of classifying audiograms in terms of configuration, symmetry and severity. The system assigns definite symmetry and severity labels to an audiogram, but provides a probabilistic breakdown for all configuration labels. We have shown that the system produces results that are on par with the current state of the art for configuration and symmetry. These results are achieved, however, with significantly less effort due to the data-driven nature of our approach as opposed to manual derivation and refinement of rules.

6

Other Applications of Machine Learning in Audiology

Our greatest weakness lies in giving up. The most certain way to succeed is always to try just one more time.

-- Thomas Edison

Our efforts, up to this point, were geared towards the development of a data-driven audiogram classification system: the DDAE. In the last three chapters, we described how we designed and applied to achieve this goal; from training set assembly to model development and validation.

The potential applications of machine learning in the field of audiology extend far beyond what we have concerned ourselves with so far. In this chapter, we explore two additional applications of machine learning in audiological research. First, we present a Gaussian Mixture Model-based method whose aim is to detect potentially unreliable audiograms. Second, we demonstrate that machine learning can be used in the context of threshold imputation to address incomplete audiograms.

6.1 Unsupervised Learning for Quality Assurance

It goes without saying that data quality is a considerable concern for anyone doing any serious data analysis work. In computer science language, this idea is often referred to as the *garbage in, garbage out* principle.

Audiology researchers must also deal with unreliable data. A variety of scenarios may lead to unreliable audiometric data, including:

1. **Excessive ambient noise:** The ambient noise masks the sound stimulus presented, resulting in threshold overestimation
2. **Poor earphones selection/placement:** The wrong type of earphones are used or are not placed correctly
3. **Poor instructions:** The instructions provided by the tester are insufficiently clear, and the subject does not respond consistently to a stimulus
4. **Uncooperative subject:** The subject claims to have heard the stimulus, while they actually have not or vice-versa

In studies on the prevalence of hearing loss, researchers analyze large collections of audiograms to understand the patterns of hearing loss in a population. To our knowledge, most of these studies do not consider the reliability of the audiograms, although they do discard incomplete ones [78], which may adversely affect the quality of the results.

Qualind™, developed by Margolis *et al.* [34], is a system capable of producing a quantitative assessement of the "quality" of an audiogram obtained through automated audiometry. This method relies on a set of variables including the patient age, gender, the length of the audiometry procedure, the false alarm rate, test-retest difference, and uses a linear model trained using these variables to predict the "quality" of the audiogram. In this study, the "quality" of the audiogram is defined as the expected difference between the thresholds obtained through automated audiometry and conventional audiometry.

A significant shortcoming with this method is that the variables its requires may not be readily available, particularly when using audiograms in retrospective studies.

Here, we are interested in producing a reasonable assessment of audiogram reliability using the least possible amount of data, i.e. audiograms alone.

6.1.1 Problem Formulation

Consistent with the assumption that we have made in Chapter 4, we again assume that there are different hearing health states that generate audiograms. These different hearing health states can be thought of as generating processes.

Assuming that this finite number of hearing disorders give rise to different "prototypical" audiograms around which some variation can occur, we want to determine how likely is it that a specific audiogram will be encountered, given these audiogram-generating processes. Audiograms which are unlikely to be generated by the possible hearing states may be at increased risk of being unreliable, because these audiograms a infrequently observed.

This problem can be viewed as a density estimation problem. Given our assumption that generating processes give rise to audiograms, the Gaussian mixture model (GMM) is an appropriate to estimate the density of the landscape. This idea is illustrated in Figure 6.1.

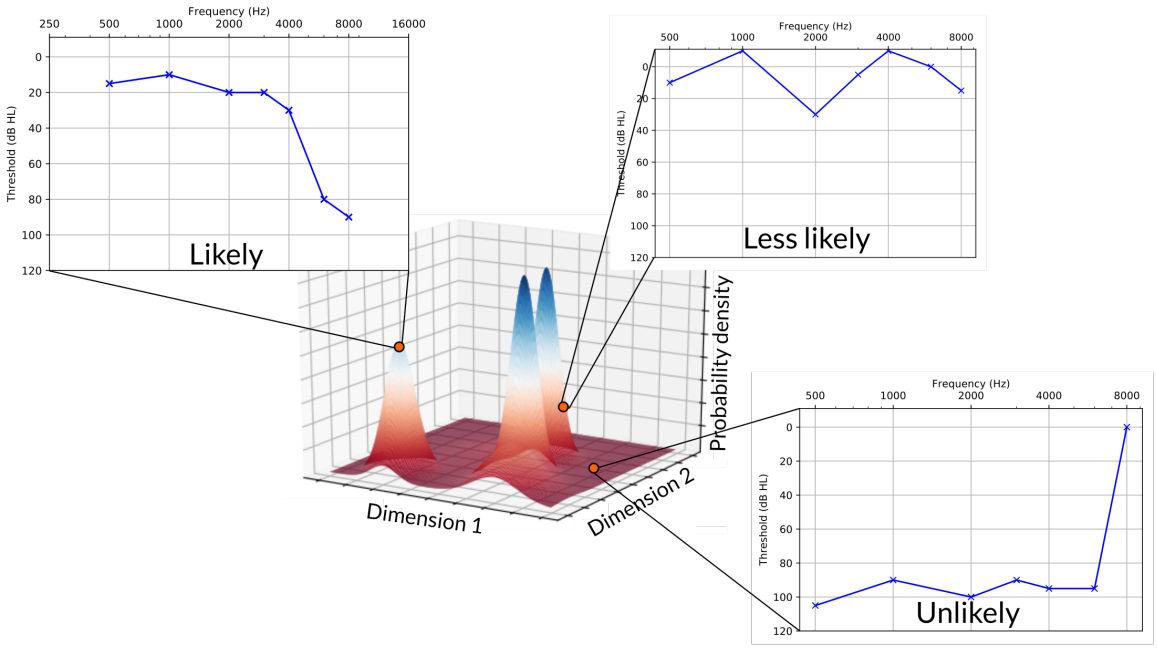


Figure 6.1: Conceptual representation of the density estimation problem
Assuming that audiograms spanned a two-dimensional space, the probability density landscape may look similar to what is shown here. Zones of high probability density corresponding to the generating processes (Gaussian Mixture components) could be seen. These regions are modeled as Gaussian distributions with a mean and covariance matrix which allow for some variation. Audiograms that fall too far away from these regions are unlikely to be observed under the model. Of course, representing an audiogram requires many more dimensions, and we use 2 dimensions here to convey the intuition behind the problem formulation which is to estimate that landscape.

6.1.2 Methodology

To model the landscape, we used the same dataset that was used in Chapter 4 to assemble a training set, namely the NHANES dataset. We used a standard GMM of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\mu_k, \mathbf{S}_k) \tag{6.1}$$

where k are the components (generating processes) defined by a normal distribution $\mathcal{N}$ with 7-dimensional mean vector μ_k, and 7×7 covariance matrix $\mathbf{S}_k$, and w_k is the mixing weight of the component k. We considered each ear separately and used the 7

raw thresholds for each ear as features.

To determine the number of processes in the GMM, we relied on the Bayesian information criterion (BIC) [59]. This index assesses the likelihood of a model while penalizing overly complex models, in this case, models with too many components. Alternatively, we could have varied the number of components until all the desired ''prototypical'' audiograms (''flat'', ''sloping'', etc.) are present, similar to the approach in [51], but this is rather subjective and may require too many components before all the prototypes are observed. We minimized the BIC by iteratively fitting GMM models with varying numbers of components to the NHANES data with the expectation-maximization algorithm, using the Scikit-learn library [74] for Python. A final model was developed using the optimal number of components. We ranked the audiograms in decreasing order of log-probability under the GMM model and visually inspected the lowest ranking audiograms.

6.1.3 Results

We found that the GMM had minimal BIC when the mixture consisted of 10 components (Figure 6.2-A). The mean of the components (Gaussians) are plotted in Figure 6.2-B. The most common audiogram configurations are represented by the mean components.

The bottom ranking audiograms are shown in Figure 6.3. Most audiograms have features that make them particularly atypical.

Multiple signs may indicate that the quality of these audiograms questionnable. For example, certain severe notches occur at unusual frequencies (2,000 Hz) in audiograms 3, 4 and 7. In another example, it is particularly unlikely that an individual is profoundly deaf, except for one frequency where hearing is excellent (audiogram 1). In other audiograms, the ''zig-zagging'' appears to make audiograms, which would otherwise not be especially remarkable, unlikely (audiograms 6 and 8).

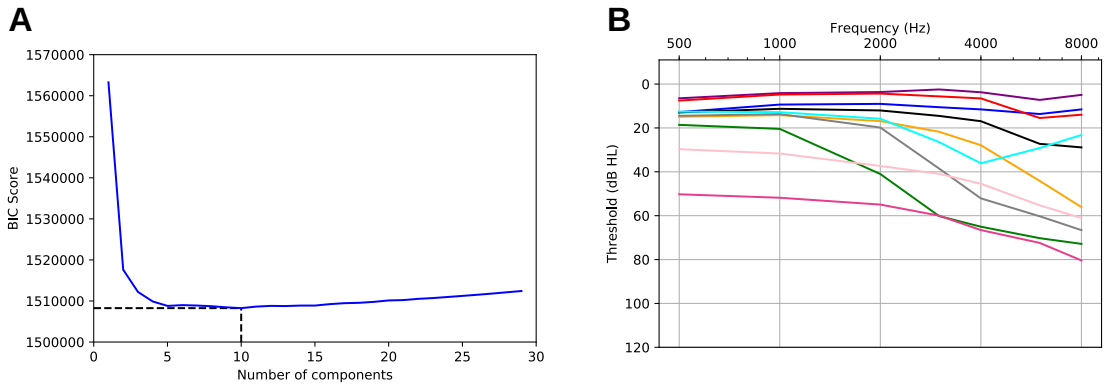


Figure 6.2: Our Gaussian Mixture Model
(A) The optimal value of the BIC is obtained at 10 components. (B) The components of the GMM model tend to correspond to prototypical audiogram types such as flat (blue), notched (cyan), and sloping/precipitous (green).

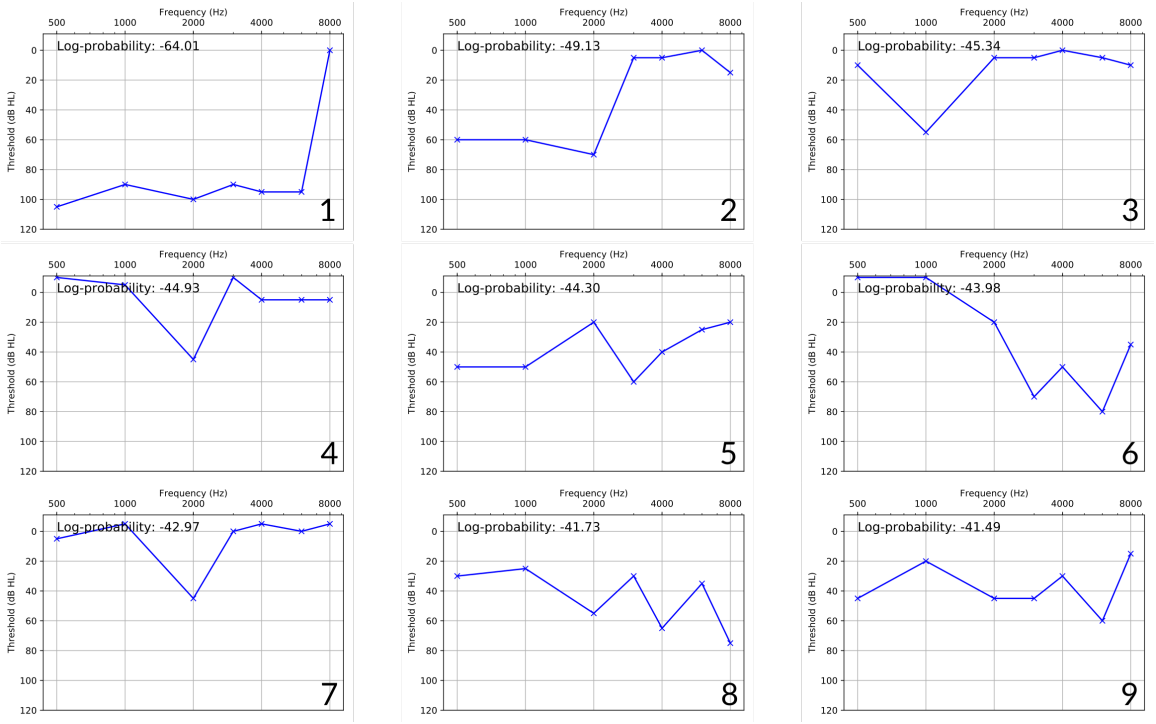


Figure 6.3: Least probable audiograms in the NHANES dataset
The 9 audiograms from the NHANES dataset determined to be least probable by our GMM model are presented. The patterns present in this set of audiograms are unambiguously atypical.

6.1.4 Discussion

Here, we have shown that GMMs can be used to model audiogram density. This landscape can, in turn, be used to assess whether an audiogram is likely to be observed or not. A visual inspection of bottom ranking audiograms in terms of log-probability revealed that these audiograms are highly atypical, as expected. The advantage of this approach over the existing approach, Qualind™, is that no additional information beyond the audiogram itself is required. Of course, a large collection of audiograms is required, making this approach widely applicable in retrospective studies and for real-time applications.

Here, we have used the NHANES dataset to determine the probability landscape. Given that the NHANES survey is conducted on the general population, the audiograms in the dataset are heterogenous. Using audiograms representative of the general population is essential for deployement of the approach proposed here. A dataset collected from a specific clinical population, veterans or workers exposed to noises, may be appropriate if the approach is to be deployed in these specific populations.

Beyond applications in audiology research, the method presented has implications for mobile audiometry, particulary as it relates to quality assurance. For example, it may be useful for non-expert users to get real-time feedback on the reliability of an audiogram. For example, on can envision that this method could be integrated in a mobile audiometer such that the user gets warned if the audiogram recorded is highly atypical. The user could then be suggested to confirm the results by repeating the procedure, in full or partially.

6.2 Supervised Learning for Threshold Imputation

The pure tone audiometry procedure involves delivery of multiple sound stimuli at multiple frequencies. Given that the threshold search can become time-consuming and lengthy, some thresholds are only queried if warranted, i.e. in certain populations or if

certain conditions are suspected. In particular, the 3,000 Hz and 6,000 Hz inter-octave frequencies are not often recorded, as opposed to standard octave frequencies. These thresholds, however, are of interest to researchers who study the prevalence of hearing loss. The 6,000 Hz threshold may be important in determining the prevalence of NIHL, for example, because the notch observed at 4,000 Hz and 6,000 Hz are often indicative of NIHL.

Inter-octave thresholds are missing from many audiograms in large databases. While it will now be relatively easy and cheap to query these thresholds with a mobile audiometer in future studies, retrieving participants in previous studies to measure these missing thresholds is inconvenient and likely to lead to highly inconsistent measurements. Imputation of missing thresholds hence makes the most sense in the context of retrospective studies, such as existing studies on the prevalence of hearing loss in specific populations. The average 6,000 Hz threshold, for example, may hold information that is useful to researchers, healthcare providers, and insurance companies who must make decisions regarding hearing aids reimbursement policies.

Currently, many researchers deal with these missing measurements in one of two ways. The first option is to simply discard audiograms where these inter-octave thresholds are missing [78]. This may result in discarding audiograms which are absolutely valid otherwise. Other choose to impute a value by averaging the surrounding octave thresholds, an approach which has been evaluated before [78, 79]. This averaging method produces measurements within ± 5 dB for 72% of audiograms at 3,000 Hz [78], but the error is not reported for 6,000 Hz.

Here we propose an alternative method that relies, again, on the assumption that the set of possible audiograms is finite and that generating processes yield audiograms which occur within clusters defined by the anatomy and physiology of the ear and hearing conditions. If that assumption holds, then imputation may benefit from comparing an incomplete audiogram with surrounding audiograms in the audiogram space to impute a

value based on the incomplete audiogram's neighbourhood.

6.2.1 Methods

To better inform imputation, we leveraged the NHANES dataset used in other sections of this thesis. We applied the same filtering procedure as done in 4.2, with the exception that we did not remove audiograms which are "easily" classified or that show no hearing loss. We then removed audiograms with a log-probability below -38. The audiograms with a lower log-probability were very unusual, and showed many signs of unreliability. The log-probability of the audiograms was computed using the GMM model trained in the previous section.

Next, we split the dataset into 3 distinct sub-datasets:

1. **All ears (D1):** includes all ears (left and right) in the dataset
2. **Adult ears (D2):** includes all ears (left and right) from patients aged 18 years or older
3. **Pediatric ears (D3):** includes all ears (left and right) from patients aged less than 18 years

We split the dataset in this way to determine whether including audiograms from the other group age affects imputation accuracy.

Given the assumptions made, the *k*-Nearest Neighbors (*k*-NN) algorithm is an intuitive choice for imputation. The algorithm, in this case, proceeds in two steps. First, the k audiograms with the lowest Euclidean distance to the incomplete audiogram are determined, using all thresholds, except the missing one. Finally, the value imputed at the missing frequency is taken to be the mean of the value at that frequency of the k nearest audiograms.

To evaluate our method, we set aside 10% of audiograms and created two test sets: one containing adult audiograms and the other containing pediatric audiograms.

All other audiograms were used to train two *k*-NN models, one designed to impute the 3,000 Hz threshold, and the other to impute the 6,000 Hz threshold. We optimized the hyperparameter k which specifies the number of neighbors to consider for imputation by selecting the value that minimized the mean absolute imputation error in 10-fold cross-validation.

6.2.2 Results

The error associated with imputation through averaging and *k*-NN-based imputation on a test set consisting of 2331 (adults) and 665 (children) ears are shown in Table 6.1 and Table 6.2, respectively. We found that in all of the cases, *k*-NN imputation tended to give less error than threshold averaging. No statistically significant improvement was seen however ($p > 0.05$; Wilcoxon's rank-sum test).

Table 6.1: Mean absolute error of threshold imputation on the adult test set

The imputation error obtained through threshold averaging and *k*-NN-based imputation are shown here.

Training Set	3,000 Hz Threshold		6,000 Hz Threshold	
	Averaging	k-NN	Averaging	k-NN
D1	5.49 $\pm$ 4.95	5.38 $\pm$ 4.75	7.36 $\pm$ 6.19	6.92 $\pm$ 5.74
D2		5.39 $\pm$ 4.76		6.96 $\pm$ 5.74

Table 6.2: Mean absolute error of threshold imputation on the pediatric test set

The imputation error obtained through threshold averaging and *k*-NN-based imputation are shown here. The results are show $\pm$ s.e.

Training Set	3,000 Hz Threshold		6,000 Hz Threshold	
	Averaging	k-NN	Averaging	k-NN
D1	4.68 $\pm$ 3.84	4.39 $\pm$ 3.54	7.66 $\pm$ 6.34	6.53 $\pm$ 5.19
D3		4.40 $\pm$ 3.50		6.50 $\pm$ 5.31

We observed that, in most cases, the age-specific (children or adults) models did

not results in improved accuracy, except when eliminating adult ears to predict the 6,000 Hz threshold in children.

We also found that, for both imputation methods, the imputation accuracy is lower at 6,000 Hz than at 3,000 Hz.

6.2.3 Discussion

We obtained poorer accuracy while attempting to impute the 6,000 Hz threshold in children using a training set consisting of both adult and pediatric ears. One possible explanation is that including adult ears to train the predictor causes the model to overpredict notches at 6,000 Hz. Such notches are more common in adults than in children, as they often result from long-term noise exposure. In that case, eliminating adult ears lessens the likelihood of predicting a notch at 6,000 Hz in children. In the other cases though, having a higher number of training data was the dominant factor in achieving improved accuracy.

A reason explaining the difference in imputation accuracies between the 3,000 Hz and 6,000 Hz thresholds is the difference in prevalence of audiometric notches at 6,000 Hz as opposed to 3,000 Hz in our dataset. Notches, by definition, correspond to narrow drops in hearing and, as such, they are independent from neighboring thresholds, making them difficult to predict without additional context (patient's occupation, noise exposure, etc.)

Taken together, our results show that a data-driven method can result in slightly -- albeit, not significantly -- more accurate imputation when addressing incomplete audiograms. The clinical significance of the improvement in imputation error is less clear, however. One bottleneck limiting the robustness our approach, and likely all other imputation approaches, is the occurrence of audiometric notches in these inter-octave frequencies which are difficult to predict.

We note that while we invested some effort in identifying and discarding potentially

unreliable audiograms using a GMM model prior to imputation, it is unlikely that these audiograms would have drastically affected imputation with our *k*-NN-based approach, as it is unlikely they would have been selected as "neighbours" by the algorithm due to their uniqueness. Nonetheless, the technique we developed remains useful for the detection of unusual audiometric patterns.

6.3 Conclusions

In this chapter, we have shown two additional applications of machine learning in audiology: one using unsupervised learning, and the other, supervised learning.

The first application is concerned with the indirect estimation of audiogram reliability. We have shown qualitatively that this method can be useful to detect highly atypical audiogram patterns. This method assumes that most audiograms are variations around "generating processes" (prototypical audiograms), and that audiograms that are unlikely to have been the result of one or a mixture of these processes are more likely to have reliability issues.

Finally, in the second half of this chapter, we investigated the use of a data-driven approach for imputation of missing data in audiograms. We have shown that the improvement in accuracy of a data-driven method over the current standard approach is modest, but consistent.

These two methods developed here are of interest to audiology researchers who perform retrospective studies using large datasets of audiograms with information beyond the audiogram itself. The first method, beyond research applications, may find a role in mobile audiometry and be used as a quality assurance check that alerts the user if the audiogram recorded is unlikely to be observed.

7

Concluding Remarks

If I have seen further than others, it is by standing upon the shoulders of giants.
-- Sir Isaac Newton

Let us conclude this thesis with a summary of the work done up to this point, and engage in a reflection regarding the limitations of this work and possible upcoming research avenues.

7.1 Summary of Contributions

The contributions presented in this thesis are as follows:

1. **Development of a Rapid Audiogram Annotation Environment (Chapter 3)**

   To our knowledge, there is no readily available software for easily annotating sets of audiograms over the internet. This was addressed through the development of a web-based software allowing for convenient, consistent, and systematic annotation of audiograms. This software was developed and developed using cloud-based technologies, so that the current study can be scaled up in the future, as needed. The web application allows authenticated users to review a personalized set of audiograms and to annotate them easily and quickly from any computer or a

tablet with an internet connection.

2. **Development of an audiogram sampling and annotation pipeline (Chapter 4)**

   A pipeline to maximize the value extracted from a small set of audiograms, due to limited annotation resources was developed. This multistep process involves data cleaning and preprocessing, feature extraction, clustering, and an iterative greedy search sampling algorithm. Altogether, the pipeline developed here aims to maximize the audiogram space coverage and to balance that objective with that of annotating common audiograms.

3. **Estimation of intra- and inter-rater reliability over five audiogram annotation tasks (Chapter 4)**

   Inter- and intra-rater reliability indicators are estimated using a more comprehensive set of annotation tasks and on a larger dataset than that used in [80]. The intra- and inter-rater reliability were assessed for the following five tasks: classification of audiogram configuration, classification of hearing loss severity, classification of audiogram symmetry, identification of audiometric notches and identification of potential reliability issues.

4. **Development and validation of a data-driven audiogram classification system (Chapter 5)**

   A data-driven audiogram classification system based on the application of a decision tree ensemble is developed, cross-validated, and compared with the state of the art. This system relieves the assumption that audiograms belong to only one configurations and provides a measure of confidence in the classification, unlike prior systems. The classification of configuration and symmetry is on par with the current state of the art, although it was not possible to compare performance for severity assessment.

5. **Development of two additional tools designed to indirectly evaluate audiogram reliability and to impute missing data in audiograms (Chapter 6)**

   Published methods for assessing audiogram reliability such as [34] cannot be used after the audiogram has been performed, as they rely on repeating certain measurements or may require values which are not routinely recording during an audiological exam. These methods are therefore not applicable on retrospective studies on hearing loss where only thresholds values are available. A simple method for detecting potential reliability relying on Gaussian Mixture Model-based density estimation was developed and tested qualitatively here. Furthermore, we presented a data-driven approach relying on the k-NN algorithm to impute values for missing inter-octave thresholds in audiograms. The method performs better than the current averaging method in most cases.

7.2 Limitations of this Work

One important limitation of the work done here is the size of the dataset that was annotated. We collected annotations for a set of 270 unique audiograms due to the high cost of audiological expertise, but we are well aware that this only constitutes a very small portion of the audiogram space.

We also note that only three expert audiologists contributed audiogram annotations in this study, and all three annotators were trained in Canada. Our assessment of inter-rater reliability may not apply for experts trained in countries where audiological training and clinical practices differ significantly. It is possible that agreement would have been adversely affected by disparities in training and practice.

Another limitation of the work done here relates to the age of the subject, which we disregarded altogether. In reality, however, the age of the patient is a key consideration in audiogram interpretation.

7.3 Future Work

7.3.1 Classification System Improvements

The most important requirement, and sometimes limitation, is data. Ideally, the data used to train the machine learning models would be high volume, high quality and highly heterogeneity. Machine learning models, while extremely flexible and convenient, can only learn so much from small datasets. As such, we believe that it would be highly beneficial to collect additional audiogram annotations. This additional data would allow to, on one hand, improve the performance of the classification system, and, in addition, to allow for more complex models to be built.

One interesting avenue that could easily be leveraged to generate more data is data augmentation. Small alterations could be made to annotated audiograms, and the resulting audiograms could retain the original annotations. It is unclear, however, whether this technique could result in the generation of synthetic audiograms that are clinically implausible, or whether the change would justify a change in the annotation.

Another consideration, in light of the insight acquired in this work as it relates to inter-rater reliability, concerns the implementation of some more extensive annotator training. A longer training session where a larger set of audiograms had been annotated collectively could have potentially improved inter-rater reliability. Another possibility involves the development of a basic rubric that guides the annotators without being overly specific as to bias their judgement.

7.3.2 Audiogram Reliability Evaluation

The approach we presented for estimating the reliability of an audiogram operates in a "global" fashion instead of "locally", meaning that it cannot pinpoint exactly what threshold(s) make the audiogram potentially unreliable. This information would be particularly invaluable, as it would allow users of an mobile audiometry system to only

repeat the measurement for the problematic threshold instead of the entire audiogram.

Future research could investigate methods to identify issues at the threshold-specific level. For example, one potential experiment would involve training GMM models with one threshold dismissed and measuring the different in log-probability with a GMM model considering all thresholds. If the audiogram is improbable overall, but quite probable when a particular threshold is dismissed, then, perhaps one could deduce that this particular threshold is problematic.

7.3.3 Identification of Site of Lesion

In this thesis, we did not attempt to classify hearing loss by site of lesion (type). Typically, this decision is based on both air conduction and bone conduction thresholds. Given that the NHANES dataset used in this study included air conduction thresholds only, there was insufficient data to identify the site of lesion for audiograms in our dataset. One could train a separate model capable of classifying the hearing loss in an audiogram as ''sensorineural'', ''conductive'' or ''mixed'', assuming that complete audiograms containing air and bone conduction thresholds are available.

7.3.4 Automated Model Refinement

It may be interesting to investigate how the classification evolves as it is retrained by adding incoming audiogram annotations to the training set. For example, one could investigate the stability of the classification system with increasing dataset size. One may be interested in answering the following questions:

1. Are there major structural changes in the decision trees when using datasets that differ in size, provenance and annotators?

2. Are there important drifts in the relative importance of certain features in making decisions when the training set changes?

These questions are interesting, because if proven to be robust, the system could be set up such that it is retrained automatically as annotations arrive in to refine the decision boundaries and improve the performance of the classification system.

7.3.5 Inclusion of Additional Data Sources

As mentioned previously, audiologists rely on the patient's history and many additional tests beyond the pure tone audiogram to paint an accurate portait of a patient's hearing. In this work, we concerned ourselves with the pure tone audiogram only. An obvious way forward involves including additional sources of data in order to better inform machine learning models. Typanogram data, questionnaires, otoscopic images could all provide a machine learning system with invaluable information that could help not only to summarize or classify the hearing loss, as done in this work, but also to identify the most probable etiology or to make tailored recommendations.

7.3.6 Recommendation Engine

Another research opportunity would be concerned with the implementation of an automated recommendation engine capable of determining whether the hearing loss constitutes a medical emergency. Furthermore, the system could be extended such that it can leverage the output of the system presented in this work, in addition to other data sources, to determine whether a referral is in order, and if so, to whom the patient should be referred: an audiologist, an ENT surgeon, or a hearing instrumentation specialist.

References

[1] François Charih, Matthew Bromwich, Renée Lefrançois, Amy E. Mark, and James R. Green. Mining Audiograms to Improve the Interpretability of Automated Audiometry Measurements. In *Proceedings of the 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, Rome, Italy, 2018.

[2] François Charih, Ashlynn Steeves, Matthew Bromwich, Amy E. Mark, Renée Lefrançois, and James R. Green. Applications of Machine Learning Methods in Retrospective Studies on Hearing. In *Proceedings of the IEEE Life Sciences Conference 2018*, Montréal, Canada, 2018.

[3] De Wet Swanepoel, Jackie L. Clark, Dirk Koekemoer, James W. Hall III, Mark Krumm, Deborah V. Ferrari, Bradley McPherson, Bolajoko O. Olusanya, Maurice Mars, Iêda Russo, and Jose J. Barajas. Telehealth in audiology: The need and potential to reach underserved communities. *International Journal of Audiology*, 49(3):195--202, January 2010.

[4] Bolajoko O Olusanya, Katrin J Neumann, and James E Saunders. The global burden of disabling hearing impairment: A call to action. *Bulletin of the World Health Organization*, 92(5):367--373, May 2014.

[5] World Health Organization. *Global Costs of Unaddressed Hearing Loss and Cost-Effectiveness of Interventions.* 2017. OCLC: 975492198.

[6] World Health Organization. Deafness and hearing loss. http://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, 2018.

[7] M Wake. Hearing impairment: A population study of age at diagnosis, severity, and language outcomes at 7-8 years. *Archives of Disease in Childhood*, 90(3):238--244, March 2005.

[8] Christine Yoshinaga-Itano, Allison L. Sedey, Diane K. Coulter, and Albert L. Mehl. Language of Early- and Later-identified Children With Hearing Loss. *Pediatrics*, 102(5):1161--1171, November 1998.

[9] Johanna Grant Nicholas and Ann E. Geers. Will They Catch Up? The Role of Age at Cochlear Implantation in the Spoken Language Development of Children With Severe to Profound Hearing Loss. *Journal of Speech Language and Hearing Research*, 50(4):1048, August 2007.

[10] Stig Arlinger. Negative consequences of uncorrected hearing loss—a review. *International Journal of Audiology*, 42(sup2):17--20, January 2003.

[11] Susan D. Emmett and Howard W. Francis. The Socioeconomic Impact of Hearing Loss in U.S. Adults:. *Otology & Neurotology*, 36(3):545--550, March 2015.

[12] Dafydd Stephens, Liz France, and Kelly Lormore. Effects of Hearing Impairment on the Patient's Family and Friends. *Acta Oto-Laryngologica*, 115(2):165--167, January 1995.

[13] Michel Picard, Serge André Girard, Marc Simard, Richard Larocque, Tony Leroux, and Fernand Turcotte. Association of work-related accidents with noise exposure in the workplace and noise-induced hearing loss based on the experience of some 240,000 person-years of observation. *Accident Analysis & Prevention*, 40(5):1644--1652, September 2008.

[14] The World Bank. World Bank national accounts data, and OECD National Accounts data files. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD, 2018.

[15] Statistics Canada Government of Canada. Hearing loss of Canadians, 2012 to 2015. https://www150.statcan.gc.ca/n1/pub/82-625-x/2016001/article/14658-eng.htm, May 2015.

[16] A. L. Oyler. Untreated Hearing Loss in Adults—A Growing National Epidemic. https://www.asha.org/Articles/Untreated-Hearing-Loss-in-Adults/, January 2012.

[17] H. Goulios and R.B. Patuzzi. Audiology education and practice from an international perspective. *International Journal of Audiology*, 47(10):647--664, January 2008.

[18] Samantha J. Kleindienst, Sumitrajit Dhar, Donald W. Nielsen, James W. Griffith, Larry B. Lundy, Colin Driscoll, Brian Neff, Charles Beatty, David Barrs, and David A. Zapala. Identifying and Prioritizing Diseases Important for Detection in Adult Hearing Health Care. *American Journal of Audiology*, 25(3):224, September 2016.

[19] Bruno M.C. Silva, Joel J.P.C. Rodrigues, Isabel de la Torre Díez, Miguel López-Coronado, and Kashif Saleem. Mobile-health: A review of current state in 2015. *Journal of Biomedical Informatics*, 56:265--272, August 2015.

[20] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22):2402, December 2016.

[21] Michael P. Menden, Francesco Iorio, Mathew Garnett, Ultan McDermott, Cyril H. Benes, Pedro J. Ballester, and Julio Saez-Rodriguez. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS ONE*, 8(4):e61318, April 2013.

[22] Yanwei Xing, Jie Wang, Zhihong Zhao, and andYonghong Gao. Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease. pages 868--872. IEEE, November 2007.

[23] Xinyu D. Song, Brittany M. Wallace, Jacob R. Gardner, Noah M. Ledbetter, Kilian Q. Weinberger, and Dennis L. Barbour. Fast, Continuous Audiogram Estimation Using Machine Learning:. *Ear and Hearing*, 36(6):e326--e335, 2015.

[24] Marco Cox and Bert de Vries. A Gaussian process mixture prior for hearing loss modeling. In *Benelearn 2017: Proceedings of the Twenty-Sixth Benelux Conference on Machine Learning, Technische Universiteit Eindhoven*, page 4, Eindhoven, 2017.

[25] Thomas Janssen Christie. *Improving the Fitting of Hearing Aids by Using Prior Information from Community Audiograms*. PhD thesis, Radboud University Nijmegen, Eindhoven, Netherlands, August 2008.

[26] Muhammad N Anwar and Michael P Oakes. Data mining of audiology patient records: Factors influencing the choice of hearing aid type. *BMC Medical Informatics and Decision Making*, 12(Suppl 1):S6, 2012.

[27] Robert H. Margolis and George L. Saly. Toward a standard description of hearing loss. *International Journal of Audiology*, 46(12):746--758, January 2007.

[28] Robert S. Schlauch and Peggy Nelson. Pure-tone Evaluation. In *Handbook of Clinical Audiology*, pages 29--47. Wolters Kluwer Health, seventh edition, 2015.

[29] James Jerger. Why the audiogram is upside-down. *International Journal of Audiology*, 52(3):146--150, March 2013.

[30] American Speech-Language-Hearing Association. Guideline: Audiometric Symbols. https://www.asha.org/policy/gl1990-00006.htm, 1990.

[31] Robert H. Margolis and Donald E. Morgan. Automated Pure-Tone Audiometry: An Analysis of Capacity, Need, and Benefit. *American Journal of Audiology*, 17(2):109, December 2008.

[32] Ian M. Windmill and Barry A. Freeman. Demand for Audiology Services: 30-Yr Projections and Impact on Academic Programs. *Journal of the American Academy of Audiology*, 24(5):407--416, May 2013.

[33] De Wet Swanepoel, Felicity Maclennan-Smith, and James W. Hall. Diagnostic Pure-Tone Audiometry in Schools: Mobile Testing without a Sound-Treated Environment. *Journal of the American Academy of Audiology*, 24(10):992--1000, November 2013.

[34] Robert H. Margolis, George L. Saly, Chap Le, and Jessica Laurence. Qualind™: A Method for Assessing the Accuracy of Automated Tests. *Journal of the American Academy of Audiology*, 18(1):78--89, January 2007.

[35] Gregory P. Thompson, Douglas P. Sladen, Becky J. Hughes Borst, and Owen L. Still. Accuracy of a Tablet Audiometer for Measuring Behavioral Hearing Thresholds in a Clinical Population. *Otolaryngology-Head and Neck Surgery*, 153(5):838--842, November 2015.

[36] Joe Saliba, Mahmoud Al-Reefi, Junie S. Carriere, Neil Verma, Christiane Provencal, and Jamie M. Rappaport. Accuracy of Mobile-Based Audiometry in the Evaluation of Hearing Loss in Quiet and Noisy Environments. *Otolaryngology-Head and Neck Surgery*, 156(4):706--711, April 2017.

[37] Jeff Davies. Delivering an audiology outreach clinic in Gujarat: Clinical observations and challenges. *Indian Journal of Otology*, 24(1):28, 2018.

[38] Ryan Rourke, David Chan Chun Kong, and Matthew Bromwich. Tablet Audiometry in Canada's North: A Portable and Efficient Method for Hearing Screening. *Otolaryngology-Head and Neck Surgery*, 155(3):473--478, September 2016.

[39] Doreen Nakku, Victoria Nyaiteera, Evelyn Llowet, Dennis Nanseera, Gladys Nakalema, Brian Westerberg, and Francis Bajunirwe. HIV status and hearing loss among children between 6 and 12 years of age at a large urban health facility in south western Uganda. *International Journal of Pediatric Otorhinolaryngology*, 101:172--177, October 2017.

[40] Scott Kohlert and Matthew Bromwich. Mobile tablet audiometry in fluctuating autoimmune ear disease. *Journal of Otolaryngology - Head & Neck Surgery*, 46(1), December 2017.

[41] Mitchell J. Isaac, Deanna H. McBroom, Shaun A. Nguyen, and Lucinda A. Halstead. Prevalence of Hearing Loss in Teachers of Singing and Voice Students. *Journal of Voice*, 31(3):379.e21--379.e32, May 2017.

[42] Raymond Carhart. An Improved Method for Classifying Audiograms. *Laryngoscope*, 55(11):640--662, 1945.

[43] C. A. Mangham. Hearing threshold difference between ears and risk of acoustic tumor. *Otolaryngology–Head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery*, 105(6):814--817, December 1991.

[44] J Northern and M Downs. *Hearing in Children*. William Wilkins Co., Baltimore, MD, 2nd edition edition, 1978.

[45] A. Goodman. Reference Zero Levels for Pure-Tone Audiometers. *American Speech-LanguageHearing Association*, 7:262--263, 1965.

[46] J Jerger and S Jerger. Measurement of hearing in adults. In M.M. Paparella and D.A. Shumrick, editors, *Otolaryngology*. W.B. Saunders, Philadelphia, PA, 2nd edition edition, 1980.

[47] G. A. Gates, N. N. Couropmitree, and R. H. Myers. Genetic associations in age-related hearing thresholds. *Archives of Otolaryngology–Head & Neck Surgery*, 125(6):654--659, June 1999.

[48] Samuli Hannula, Risto Bloigu, Kari Majamaa, Martti Sorri, and Elina Mäki-Torkko. Audiogram configurations among older adults: Prevalence and relation to self-reported hearing problems. *International Journal of Audiology*, 50(11):793--801, November 2011.

[49] A. L. Pittman and P. G. Stelmachowicz. Hearing Loss in Children and Adults: Audiometric Configuration, Asymmetry, and Progression:. *Ear and Hearing*, 24(3):198--205, June 2003.

[50] K. Demeester, A. van Wieringen, J. J. Hendrickx, V. Topsakal, E. Fransen, L. Van Laer, D. De Ridder, G. Van Camp, and P. Van de Heyning. Prevalence of tinnitus and audiometric shape. *B-ENT*, 3 Suppl 7:37--49, 2007.

[51] Nikolai Bisgaard, Marcel S. M. G. Vlaming, and Martin Dahlquist. Standard audiograms for the IEC 60118-15 measurement procedure. *Trends in Amplification*, 14(2):113--120, June 2010.

[52] Cheng-Yung Lee, Juen-Haur Hwang, Szu-Jen Hou, and Tien-Chen Liu. Using cluster analysis to classify audiogram shapes. *International Journal of Audiology*, 49(9):628--633, September 2010.

[53] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009.

[54] Wei-Yin Loh. Classification and regression trees: Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14--23, January 2011.

[55] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: An overview, II: Algorithms for hierarchical clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(6):e1219, November 2017.

[56] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53--65, November 1987.

[57] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1--38, 1977.

[58] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716--723, December 1974.

[59] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461--464, March 1978.

[60] Kilem L. Gwet. Intrarater Reliability. In *Wiley StatsRef: Statistics Reference Online*. American Cancer Society, September 2014.

[61] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37--46, April 1960.

[62] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378--382, 1971.

[63] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159--174, March 1977.

[64] Facebook Inc. React – A JavaScript library for building user interfaces. https://reactjs.org/index.html, 2018.

[65] Jordan Novet. Microsoft gains cloud market share in Q1, but AWS still dominates. https://www.cnbc.com/2018/04/27/microsoft-gains-cloud-market-share-in-q1-but-aws-still-dominates.html, 2018-04-27T16:42:02-0400.

[66] John Bradley, Nat Sakimura, and Michael Jones. JSON Web Token (JWT). https://tools.ietf.org/html/rfc7519, 2015.

[67] J. G. Clark. Uses and abuses of hearing loss classification. *ASHA*, 23(7):493--500, July 1981.

[68] K. C. P. Yuen and B. McPherson. Audiometric configurations of hearing impaired children in Hong Kong: Implications for amplification. *Disability and Rehabilitation*, 24(17):904--913, January 2002.

[69] Ting Cai, Bradley McPherson, Caiwei Li, and Feng Yang. Pure tone hearing profiles in children with otitis media with effusion. *Disability and Rehabilitation*, 40(10):1166--1175, May 2018.

[70] Peter M. Rabinowitz, Deron Galusha, Martin D. Slade, Christine Dixon-Ernst, Kanta D. Sircar, and Robert A. Dobie. Audiogram Notches in Noise-Exposed Workers:. *Ear and Hearing*, 27(6):742--750, December 2006.

[71] Christopher G Brennan-Jones, Robert H Eikelboom, Rebecca J Bennett, Karina FM Tao, and De Wet Swanepoel. Asynchronous interpretation of manual and automated audiometry: Agreement and reliability. *Journal of Telemedicine and Telecare*, 24(1):37--43, January 2018.

[72] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin, Heidelberg, 2006.

[73] Rolf Quam, Ignacio Martínez, Carlos Lorenzo, Bonmatí A, Manuel Rosa-Zurera, Jarabo P, and Arsuaga JL. Studying audition in fossil hominins: A new approach to the evolution of language? In *Psychology of Language*, pages 1--37. Nova Science Publishers, Inc., New York, NY, October 2012.

[74] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825--2830, 2011.

[75] SeleniumHQ. Selenium - Web Browser Automation. https://www.seleniumhq.org/, 2018.

[76] Gary Bradski. The OpenCV Library. http://www.drdobbs.com/open-source/the-opencv-library/184404319, 2000.

[77] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, page 21, 2007.

[78] Richard K. Gurgel, Gerald R. Popelka, John S. Oghalai, Nikolas H. Blevins, Kay W. Chang, and Robert K. Jackler. Is it valid to calculate the 3-kilohertz threshold by averaging 2 and 4 kilohertz? *Otolaryngology–Head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery*, 147(1):102--104, July 2012.

[79] Richard H. Wilson and Rachel McArdle. A treatise on the thresholds of interoctave frequencies: 1500, 3000, and 6000 Hz. *Journal of the American Academy of Audiology*, 25(2):171--186, February 2014.

[80] Christopher G Brennan-Jones, Robert H Eikelboom, and De Wet Swanepoel. Diagnosis of hearing loss using automated audiometry in an asynchronous telehealth model: A pilot accuracy study. *Journal of Telemedicine and Telecare*, 23(2):256--262, February 2017.

[81] M. A. Hersh and M. A. Johnson. *Assistive Technology for the Hearing-Impaired, Deaf and Deafblind*. Springer-Verlag, London, 2003.

[82] Encyclopaedia Britannica, Inc. Basilar membrane. https://www.britannica.com/science/basilar-membrane, 2018.

[83] Kiyofumi Gyo. Measurement of the Ossicular Vibration Ratio in Human Temporal Bones by Use of a Video Measuring System. *Acta Oto-Laryngologica*, 103:87--95, 1987.

[84] Keerthana Devarajan, Hinrich Staecker, and Michael S Detamore. A Review of Gene Delivery and Stem Cell Based Therapies for Regenerating Inner Ear Hair Cells. *Journal of functional biomaterials*, 2:249--270, December 2011.

[85] J Lim. Functional structure of the organ of Corti: A review. *Hearing Research*, 22:117--146, 1986.

[86] Encyclopaedia Britannica, Inc. Structure of the Human Ear. https://www.britannica.com/science/ear/images-videos, 2018.

[87] Raymond Carhart and James Jerger. Preferred method for clinical determination of pure-tone thresholds. *Journal of Speech & Hearing Disorders*, 24:330--345, 1959.

[88] J. Tonndorf. Sensorineural and pseudosensorineural hearing losses. *ORL; Journal for Oto-Rhino-Laryngology and Its Related Specialties*, 50(2):79--83, 1988.

Appendices

Appendix A

Anatomy and Physiology of the Ear

The ear is a truly fascinating organ that enables us to interact with our environment in ways that our other senses cannot. If the sound of the summer breeze or an old love song can rewind time and awaken memories hidden in the deepest corners of our conscience, it is partly thanks to this intricate organ. Its task is to funnel, modulate, convert and transmit sound waves in an exquisitely well-organized series of steps, which is worth visiting -- at a high level, at least.

The discussion in this section summarizes the description of the inner ear made in [81]. The reader is invited to consult this source for further details. The ear subdivides into three main components (Figure A.1): the outer ear, the middle ear and the inner ear.

The outer ear consists of the auricle (or pinna), the visible part of the ear and the ear canal. In engineering terms, it is said to act both as a resonator and high-pass filter for sound.

The middle ear, is located in a cavity along the mastoid process, a small portion of the skull behind the ear. It is comprised of the tympanic membrane (eardrum) as well as three ossicles, the *malleus* (or hammer), the *incus* (or anvil) and the *stapes* (or footplate). The ossicles, whose common names accurately describe their shape, are coupled such that they form a chain that moves cooperatively. The *malleus* is directly attached to the

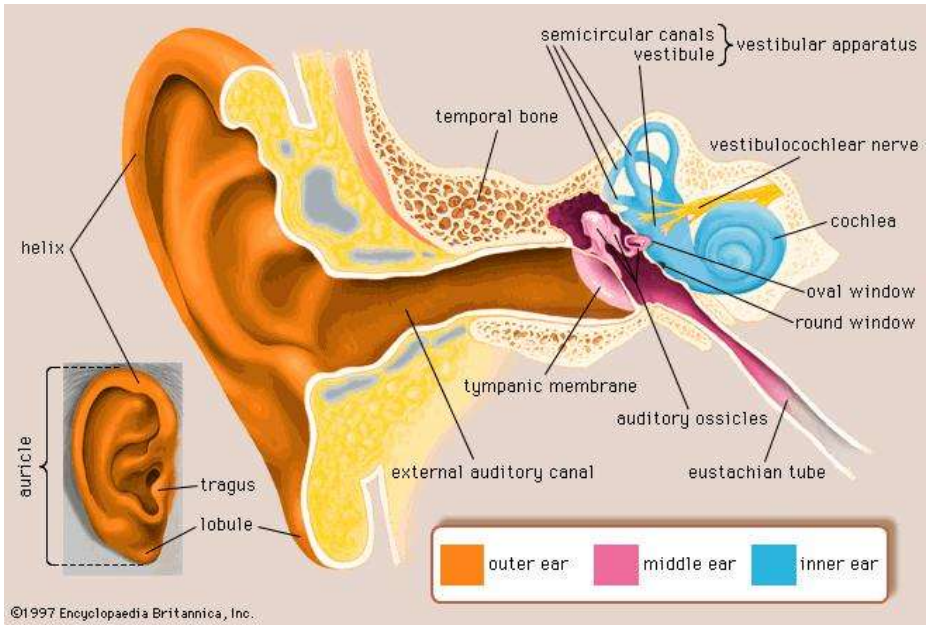


Figure A.1: High-level organization of the ear
In this illustration (reproduced from [82]), we can clearly distinguish the structures belonging to the outer ear (orange), the middle ear (pink) and inner ear (light blue).

eardrum that it connects to the *incus*. The *incus*, in turn, transmits the motion to the stapes. The tympanic membrane responds to sound waves by vibrating back-and-forth with the compression and rarefaction of air molecules. This ossicular chain is responsible for transduction and impedance matching. It converts energy in the form of vibration of the eardrum, into pressure waves of the fluid filling the cochlea through the oval window, a small opening accessible to the flat end of the footplate. Given the energy loss in converting between these energy forms, the ear must compensate somehow. It does so by leveraging the elongated shape of the ossicular chain and the lever principle. It is well known that the application of a small force onto the farthest extremity from the pivot point generates a large force at the opposite extremity, and this is precisely how small movements of the tympanic membrane translate to vigourous motion of the inner ear fluid. It is believed that the amplification factor is frequency-dependant, and may reach up to 6x near 2 kHz [83].

The last component, the inner ear, consists principally of the aforementioned

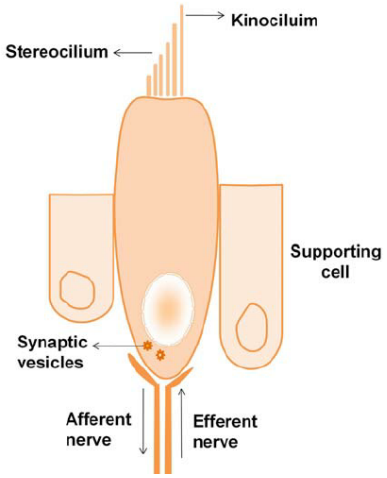


Figure A.2: A hair cell anchored in a supporting cell
Hair cells are anchored in supporting cells and expose stereocilia to the perilymphatic fluid in the cochlea's lumen (illustration reproduced from [84]).

cochlea, a snail-shaped fluid-filled organ attached to the auditory nerve (cranial nerve VIII). This intricate organ is directly responsible for converting mechanical energy into an electric impulse. The motion of perilymph in the Organ of Corti, a compartment in the cochlea, stimulates small hair cells (Figure A.2) from which cilia protrude into the lumen. As these cilia shear, the cell membrane at the base of hair cells allows the release of neurotransmitters to the neurons of the auditory nerve at proximity which then transmit the signal to the central nervous system for processing. It is worth pointing out that hair cells are organized *tonotopically* [85]. If we were to unwrap the spiral-shaped cochlea, adjacent groups of hair cells along it would respond to low frequencies at one extremity and gradually respond to higher frequencies as we work our way down towards the opposite extremity (Figure A.3).

The human ear can perceive sounds ranging from 20 Hz to 20,000 Hz, allowing us to appreciate and enjoy a vast range of timbres and textures, from the low-pitched complaint of a double bass to the high-pitched cries of a violin.

It goes without saying that the organization of the ear is appreciably more complex than what was described here. However, even taking only this basic description into

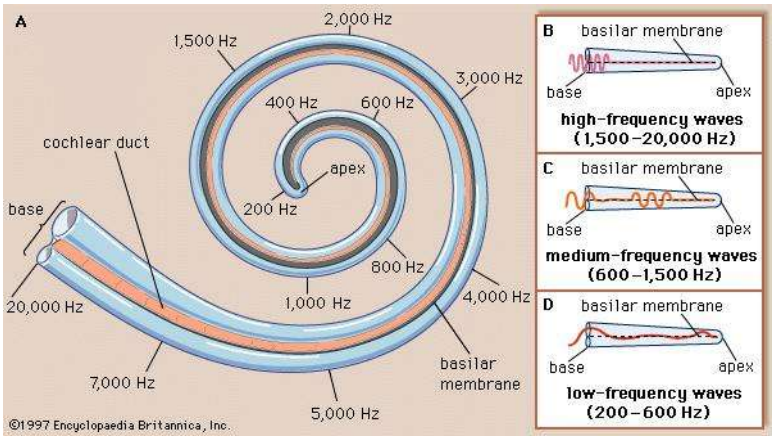


Figure A.3: Tonotopic map of the cochlea
This illustration (reproduced from [86]) displays a map of the frequency reponse of the different locations in the cochlea. The organization of the cochlea is reminiscent of the tonal organization of a piano's keyboard.

account, it should be apparent to the reader that there are multiple points of failure from which diseases of the ear and hearing loss could potentially arise.

Appendix B

Pure Tone Audiometry

Audiologists are trained to deliver a large range of tests to assess the hearing of a patient. Most people will be familiar with otoscopy, wherein an otoscope is inserted in the ear canal to look for obstruction by ear wax and inspect the tympanic membrane, as it is often conducted as part of a routine medical examination. In tympanometry, the pressure in the ear canal is varied and the transmission (or *admittance*) of a tone of fixed frequency to the middle ear is measured over the pressure range. This test is useful for determining the middle ear status of a patient. Establishing a comprehensive list of tests in audiology is beyond the scope of this thesis. Nevertheless, no discussion about audiology would be complete without a proper treatment of audiometry, and more particularly *pure tone audiometry*, the most common hearing test. Incidentally, the output of this hearing test underpins the work done as part of this thesis. Hence, this section will aim to present the most relevant aspects of this procedure, as they relate to this thesis.

B.1 Etymology

The etymology of the word *audiometry* is relatively straightforward, with its prefix, *audio-*, coming from the latin word *audire* for "listen" and its suffix, *-metry*, coming from Ancient

Greek word *métron* for "measure". To state what should have become obvious by now, the purpose of audiometry is to measure how well one can listen or hear.

In pure tone audiometry, the tones presented to the patient are pure, that is to say that they are sinusoidal waves of fixed frequency and amplitude. This detail is important, as audiologists also conduct other types of audiometry procedures, among which warble tone audiometry and speech audiometry. In warble tone audiometry, the tones have modulated frequency, whereas in speech audiometry, the speech that is heard is composed of a complex frequency spectrum.

B.2 Objectives and Procedure

The reader probably wonders what exactly is measured in pure tone audiometry and how.

The ultimate goal of pure tone audiometry is to quantitatively measure the threshold of hearing of a patient on the decibel (dB) scale. This threshold is defined as the sound level at which there is a 50% chance that the sound will be heard, or in other words, the sound level where the sound is heard half of the time. Naturally, this threshold may vary along the frequency spectrum. Recall the tonotopic organization of the cochlea. If a lesion of the hair cells occurs in the outer portion of the cochlea, the capacity to hear high frequencies will be adversely affected, but not necessarily that to hear low frequency sounds.

Notwithstanding recent advances in mobile audiometry, traditional pure tone audiometry is usually performed in a soundproof booth compliant with maximum ambiant noise standards specifieds in ANSI S3.1. The procedure is performed with an audiometer capable of generating pure tones of specific frequencies and amplitudes, earphones and, occasionally, bone transducers to test sound conduction through bone. Multiple types of earphones are used, *supra-aural* (over the ear), *insert* (inside the ear) and *circumaural*

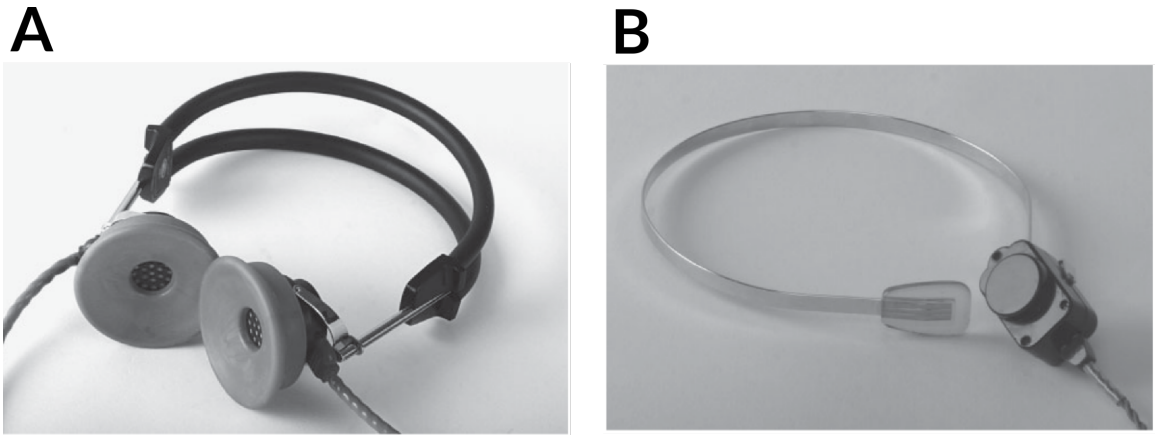


Figure B.1: Equipment used in conventional pure tone audiometry
Multiple types of headphones are used to administer pure tone audiometry, among which: (A) Supra-aural headphones and (B) Bone transducers. (Adaptation from [28])

(over and around the ear). The type of headphones selected depends on a variety of factors including the ease of calibration, the presence of ear canal collapse, the age of the patient, the required amplitude to be delivered, among others. Bone transducers are often positioned on the mastoid bone behind the ear.

Prior to the start of the test, the examiner will place the headphones on the patient's head and deliver an unambiguous set of instructions. Typically, the patient will be told to react as soon as a tone is heard, even if the tone is perceived to be very faint to prevent bias. The patient is normally asked to raise a hand or to tap a button to indicate that the tone was heard.

During the procedure, auditory stimuli are delivered through the earphones one ear at a time. Normally, if there is an expected asymmetry in hearing, the better ear is tested first, followed by the *contralateral* (opposite) ear. Otherwise, the convention is to test the right ear and to then assess the left ear. Hearing thresholds are measured systematically frequency-by-frequency.

The frequencies typically measured range from 250 Hz to 8,000 Hz -- or 16,000 Hz, if the audiometer allows for extended high frequencies (EHF) testing, and if the audiologist judges it to be appropriate. The conventional set of tested frequencies comprises:

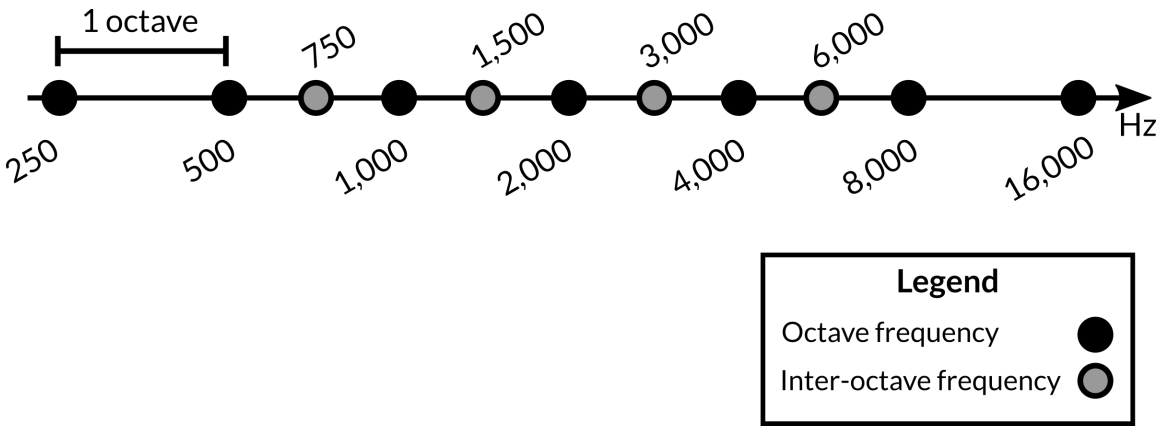


Figure B.2: Common test frequencies measured in pure tone audiometry
The common test frequencies are plotted along the frequency spectrum. Notice that octave frequencies are equidistant on a $\log_2$ scale. Inter-octave frequencies lie approximately, but not exactly, in the center of consecutive octave frequencies.

250 Hz, 500 Hz, 1,000 Hz, 2,000 Hz, 4,000 Hz and 8,000 Hz. These frequencies are separated by a distance called an *octave*. An octave corresponds to the distance between a frequency and its doubled value. For example, 500 Hz is one octave lower than 1,000 Hz and 3 octaves lower than 4,000 Hz. The test may also comprise frequencies that fall between two octaves, such as 1,500 Hz, 3,000 Hz or 6,000 Hz. These frequencies are called *inter-octave* frequencies. These concepts are represented schematically in Figure B.2.

There exists multiple threshold search methods, but most audiologists perform some variation of the modified Hughson-Westlake threshold search algorithm to measure the thresholds of hearing.

B.3 Modified Hughson-Westlake Threshold Search Algorithm

The modified Hughson-Westlake threshold search algorithm [87] was introduced over 50 years ago and its variants are still taught in many audiology schools. The purpose of

the procedure is to determine, for every test frequency, a reliable threshold estimate. The algorithm uses a "10-down, 5-up" approach wherein the amplitude of the pure tone is decreased by 10 dBs whenever the patient hears the stimulus, and increased by 5 dBs otherwise. The threshold corresponds to the minimum amplitude at which the patient responds in two consecutive ascending runs or in 50% of ascending runs, depending on the version of the ASHA guidelines followed by the tester. The procedure for a single test frequency is illustrated in Figure B.3.

In the traditional variant of the approach, the patient is presented with a pure tone at 1,000 Hz at 30 dB. This initial presentation allows for two things: familiarize the patient with the test by delivering a tone at a comfortable level, and for test-retest evaluation purposes, as the 1,000 Hz threshold is sometimes retested at the end of the procedure to estimate the reliability of the test.

B.4 Bone Conduction Evaluation

A complete audiogram comprises measurements for bone conduction hearing thresholds. The purpose of these measurements is to measure the function of the inner ear. Contrarily to air conduction measurements, which are measured by presenting pure tones through earphones, bone conduction thresholds are measured using a bone transducer. The transducer, most often placed on the mastoid process behind the ear, although sometimes on the forehead, transmits sound waves through the skull and directly to the inner ear.

Pure tones presented through the bone bypass a large portion of the sound transmission chain, namely the tympanic membrane and ossicular chain. By obtaining bone conduction thresholds in addition to air conduction threshold, the practitioner can make a judgement with respect to the *site of lesion* (sometimes referred to as the hearing loss *type*), or the ear component responsible for the impairement. In other

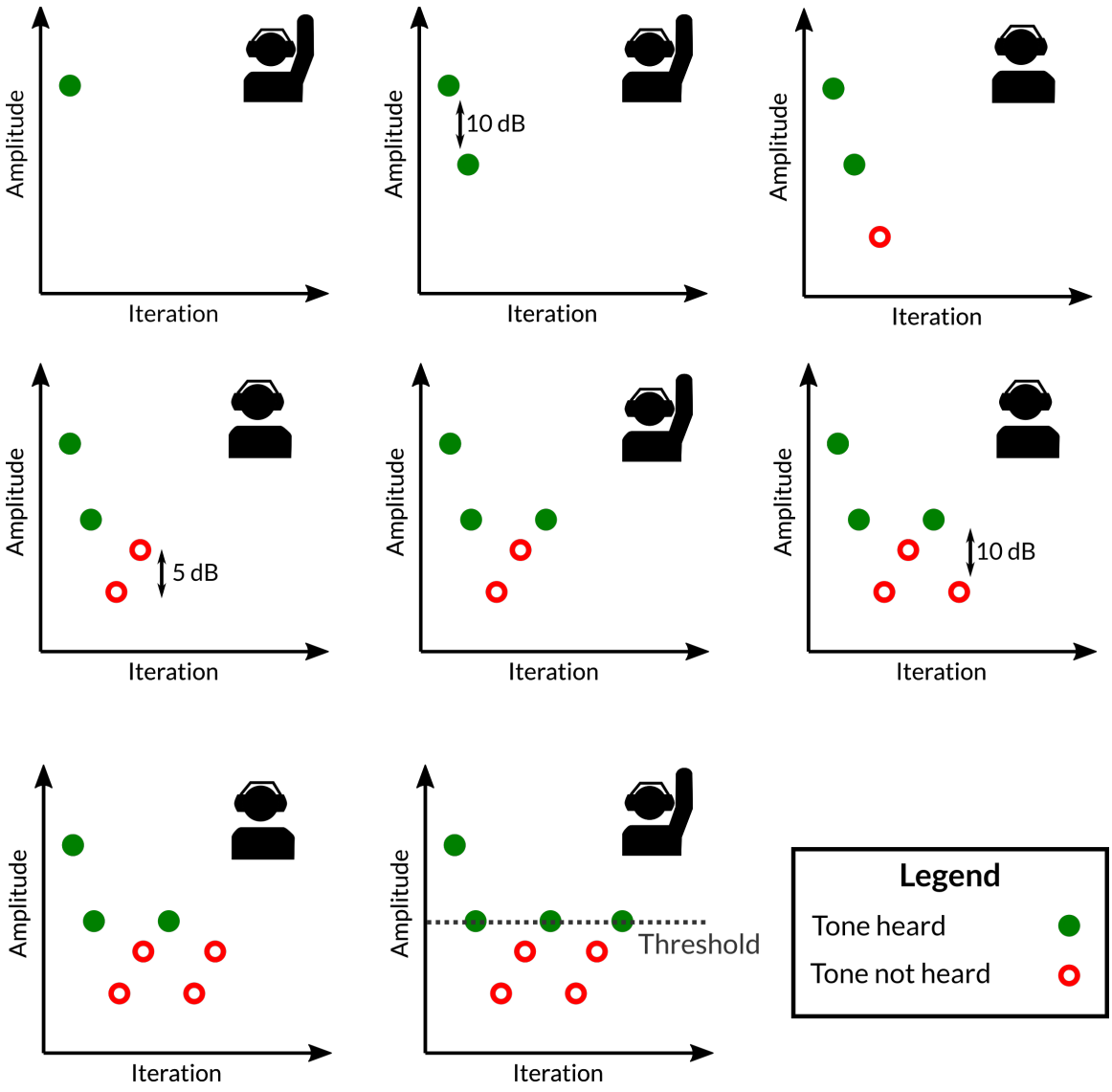


Figure B.3: The modified Hughson-Westlake procedure
This diagram, meant to be read from left to right, top to bottom, illustrates how the threshold of hearing for a specific frequency is determined. This procedure depends on the reaction of the patient to sound stimuli.

words, bone conduction evaluation allows the tester to determine the bottleneck in sound transmission from the outer ear to the brain centres.

If bone conduction thresholds are normal, but air conduction thresholds are elevated, the practionner will likely suspect a *conductive* hearing loss, because thresholds are normal if the sound is directly transmitted to the inner ear. This type of hearing loss results from a dysfunction within the sound transmission chain in the outer or middle ear, for example a wax occlusion or a tympanic membrane perforation. If both air thresholds and bone thresholds are increased roughly equally, a *sensorineural* hearing loss will typically be suspected as the bottleneck is not in the sound conduction chain. This type of hearing loss results from lesions in or beyond the inner ear. A third type of hearing loss is the *mixed* hearing loss, wherein there are conductive and sensorineural components to the impairement. There are other types of hearing loss (e.g. pseudo-sensorineural [88]), but they are less common.

Identifying the site of lesion is pivotal in establishing a prognosis and selecting the most adequate treatment. Some types of hearing losses, particularly conductive hearing losses, can be corrected surgically, whereas others cannot. Sensorineural hearing losses are most often permanent and may require amplification through hearing aids or the implantation of cochlear implants.

B.5 Masking

For a variety of reasons, a patient's nontest ear may respond to the stimulus. This unintended situation is referred to as *cross hearing*. To lower the risks of cross hearing, the tester may apply *masking*. To mask the sound which may travel through the skull from the test ear, the tester may apply a carefully selected masking noise such as a white noise to the nontest ear.

Appendix C

Ethics

You will find, in the following pages, the documentation that was provided to the participants in this study. The documentation included a formal invitation letter, and an electronic form to be read and agreed to by the participants (through a "Submit" button in the Rapid Audiogram Annotation Environment).

Carleton UNIVERSITY
Canada's Capital University

Invitation email

Subject: Invitation to participate to a study on the classification of SHOEBOX audiograms

Dear Sir or Madam,

My name is Francois Charih. I am a Master's student in Electrical and Computer Engineering working under the supervision of Prof. James R. Green at Carleton University. We are leading a study named *Classification of SHOEBOX audiograms* in partnership with Clearwater Clinical Limited, an Ottawa-based company specializing in mobile audiometry.

The principal objective of this study funded by NSERC/OCE and Clearwater Clinical Limited is to develop a decision support system capable of classifying an audiogram and generating a summary describing its configuration, its symmetry, the severity of the hearing loss, as well as a list of etiologies the audiogram may be consistent with. This system will increase the interpretability of audiograms by non-specialists users (family doctors, technicians, etc.) of the SHOEBOX Audiometry mobile audiometer developed and marketed by Clearwater Clinical Limited, while simultaneously serving as an education tool.

In order to develop this system, a large collection of audiogram annotations needs to be assembled for a heterogeneous set of audiograms. We are recruiting participants who hold a graduate degree in audiology (or equivalent), have experience in a clinical setting and that are fluent in English. We are contacting you because your training and expertise in the field of audiology puts you in a unique position to contribute to this effort.

In this study, participants will be invited to create an account on a web-based audiogram annotation environment developed specifically for this study. Participants will be notified when the account is created and will be asked to log into the annotation environment to annotate a set of carefully sampled, de-identified audiograms. For every audiogram in the assigned set, participants will be asked to select from lists the descriptors that best describe its reliability, symmetry, configuration, severity, whether the classification was easy or challenging (and to provide a short justification for your annotation in the latter case), and to select the most relevant follow-up questions. This study can be completed from any location where an Internet connection is available, and we expect that participants will complete the task in multiple sittings.

This study does not expose you to more risk than you would experience in your daily life as a professional audiologist, and you will have the right to end your participation in the study any time before October 31st, 2018, for any reason.

As a token of appreciation, you will be compensated at a rate of $3 per fully annotated audiogram. This compensation was fixed under the assumption that an audiogram will take approximately 3.5 minutes to annotate, on average. You will receive compensation from Carleton University after data collection has ended. You will be required to complete some paperwork with Carleton University Human Resources to receive these funds. No other compensation will be provided. Should you choose to withdraw, you will still be compensated for the annotations contributed up until the moment of withdrawal, unless you request the annotations and feedback you provided be destroyed, in which case you will be entitled to $100 for your time.

We will not gather sensitive personal information, but we will collect basic contact information (name, email) and information regarding your professional experience (position, years of experience, affiliation, expertise, etc.). You will be assigned an anonymous annotator identifier upon registration, and all following analyses will be carried out using these anonymous identifiers. Your contact information and professional information will be stored in an

encrypted file on the researcher's password-protected computer at Carleton University in order to allow us to compensate you or reach you to provide additional information or instructions. All other data collected throughout this process will be stored in a password-protected database on Amazon Web Services.

We may compute metrics such as intra- and inter-annotator agreement and the time required to generate the annotations. To help protect all participants from possible economic consequences in the event of a data breach, all personal information will be kept separately from all other data. Furthermore, reasonable security precautions (password protection, encryption, etc.) have been taken such that the risk of such an event is minimal.

At the end of the study, all annotations and feedback regarding the software will be shared with Clearwater Clinical Limited, who may use it for further research and development as they are providing part of the funding for the project. The data shared with Clearwater Clinical Limited will be coded, and will only include the annotations and the feedback concerning the software. Please note, however, that although the data to be transmitted to Clearwater Clinical Limited is in its coded form (de-identified), we cannot guarantee that Clearwater will not be able to link the data back to you using circumstantial information as you were invited to participate to this study based on their recommendation.

This ethics protocol for this project was reviewed by the Carleton University Research Ethics Board, which provided clearance (CUREB-B Clearance # 108278) to carry out the research. The clearance for this experimental protocol expires on January 31st, 2018.

Should you have questions or concerns related to your involvement in this research, please contact Dr. Andy Adler, Chair, Carleton University Research Ethics Board-B (by phone at 613-520-2600 ext. 4085 or via email at ethics@carleton.ca).

If you would like to participate in this research project, or have any questions, please contact Renée Lefrançois (renee@clearwaterclinical.com) or François Charih (francois.charih@carleton.ca).

Cordially,

François Charih, B.A.Sc., B.Sc.(Hons.)
M.A.Sc. Candidate
Carleton University

Carleton UNIVERSITY
Canada's Capital University

Terms of the agreement/consent form

This text will be displayed to the participant before they submit the registration form, as part of the online Rapid Audiogram Annotation Tool software developed in our lab.

Title: Automating classification of SHOEBOX audiograms

Funding Source: Ontario Centers of Excellence (OCE), Natural Sciences and Engineering Research Council of Canada (NSERC), and Clearwater Clinical Ltd.

Date of ethics clearance: January 17th, 2018

Ethics Clearance for the Collection of Data Expires: January 31st, 2019

This is a study on the classification of audiograms. The study aims to develop an automated classification system capable of classifying, i.e. generating a summary describing the audiogram configuration, its symmetry, the severity of the hearing loss, as well as a list of etiologies the audiogram may be consistent with. The researcher for this study is Francois Charih from the Systems and Computer Engineering department at Carleton University. Francois Charih is working under the supervision of Professor James R. Green.

To register and participate to the study, you must hold a graduate degree in audiology (or equivalent), have experience in a clinical setting, be fluent in English and have received an invitation to participate to the study.

Once you provide your consent by submitting this form, the researchers will be in possession of the contact information and professional information you provided. You will be asked to annotate a set of de-identified audiograms to the best of your knowledge. You will also be periodically invited to provide feedback on the audiogram annotation software used in this study for future improvement. During the course of this research project, we will be computing various metrics such as the number of audiograms annotated per hour, inter-annotator variability, etc.

You will be completing the tasks using the Rapid Audiogram Annotation Environment developed specifically for this study. You will be free to complete the task from the location of your choosing.

Given that basic contact and professional information are collected and that reliability metrics will be computed, there are some economic risks to you in case of a data breach. The risks are minimal, and we will take all the necessary precautions to limit these risks and to protect your identity. Upon registration, you will be assigned an anonymous identifier that will allow you to log into the annotation environment. This identifier will be sent to you by email. Any annotations and feedback provided through the software will be associated to this anonymous identifier. All analyses will be carried out using the anonymous identifiers and all publications will make use of generic pseudonyms to refer to the annotators (e.g. Annotator 1, Expert Audiologist 2, Medical Student 3, etc.). Your personal information will be stored separately on a password-protected computer in Professor James Green's laboratory.

As a token of appreciation, you will be compensated at a rate of $3 per fully annotated audiogram. We expect that each audiogram will take about 3.5 minutes to annotate on average. You will be compensated shortly after the end of the data collection period. You may be required to complete paperwork with Carleton University in order to receive your funds.

You are able to withdraw from the study at any time until data collection has been completed (31 Oct 2018), for any reason. You can withdraw by phoning or emailing the researcher or the research supervisor. Your personal information will be promptly erased, should you decide to withdraw. All annotations and feedback regarding the software provided up until the time of withdrawal will be preserved and shared with Clearwater Clinical Limited for future research and development. You will still be compensated for every complete annotation you have provided unless you request that we destroy your annotations and feedback, in which case you will be entitled to $100 for your time.

Your contact and professional information will be stored separately from all other data in an encrypted file on a password-protected computer in Prof Green's laboratory at Carleton University. Your personal information will only be accessible by the researcher and the research supervisor. Non-identifiable research data (i.e. annotations and feedback regarding the software) will be stored in a password-protected database on a server belonging to Amazon Web Services. Amazon Web Services are

located in the United States and are thus subject to U.S. laws on data privacy. The researchers will not track IP addresses of participants.

We may compute metrics such as intra- and inter-annotator agreement and the time required to generate the annotations. To help protect all participants from possible economic consequences in the event of a data breach, all personal information will be kept separately from all other data. Furthermore, reasonable security precautions (password protection, encryption, etc) have been taken such that the risk of such an event is minimal.

At the end of the project, all annotations and feedback concerning the annotation environment will be downloaded, erased from Amazon Web Services, and archived indefinitely on the researcher's password-protected computer. Annotation and feedback data will also be shared with Clearwater Clinical Limited and may be used for future research by the researchers and/or Clearwater Clinical Limited, as they are providing funding for the project. All personal information (e.g. names, email addresses and affiliations) will be erased permanently and will not be shared with Clearwater Clinical Ltd. However, we cannot guarantee that Clearwater Clinical Limited will not be able to link the data back to you using circumstantial information as you were invited to participate to this study following their recommendation.

The ethics protocol for this project was reviewed by the Carleton University Research Ethics Board, which provided clearance (CUREB-B Clearance # 108278) to carry out the research.

If you have any ethical concerns with the study, please contact Dr. Andy Adler, Chair, Carleton University Research Ethics Board-B (by phone at 613-520-2600 ext. 4085 or via email at ethics@carleton.ca).

Researcher contact information:
Name: Francois Charih
Department: Systems and Computer Engineering
Carleton University
Email: francois.charih@carleton.ca

Supervisor contact information:
Name: James R. Green
Department: Systems and Computer Engineering
Carleton University
Tel: (613) 520-2600 (x1463)
Email: jrgreen@sce.carleton.ca

By clicking "submit", you consent to participate in the research study as described above.

Appendix D

Supplementary Material

Table D.1: Equivalence pairs between the DDAE and AMCLASS™'s configuration schemes

In order to compensate for the differences between the configuration classification system used in this study and that used in the study by Margolis and Saly [27], correspondance pairs were defined. These configurations are considered to be equivalent.

DDAE	AMCLASS™
Flat	Flat
Sloping	Sloping
Precipitous	Sloping
Reverse sloping	Rising
Cookie bite	Trough-shaped
Reverse cookie bite	Peaked
Atypical	Other
Notched	Other

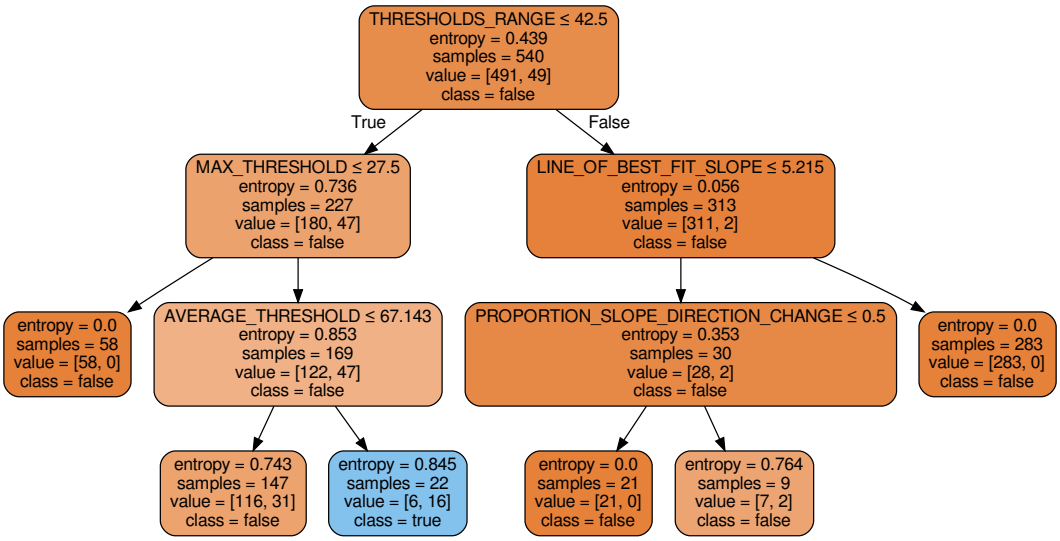


Figure D.1: Decision tree structure for the flat/not-flat classifier
This decision tree for the binary classifier classifying an audiogram as flat or not flat.

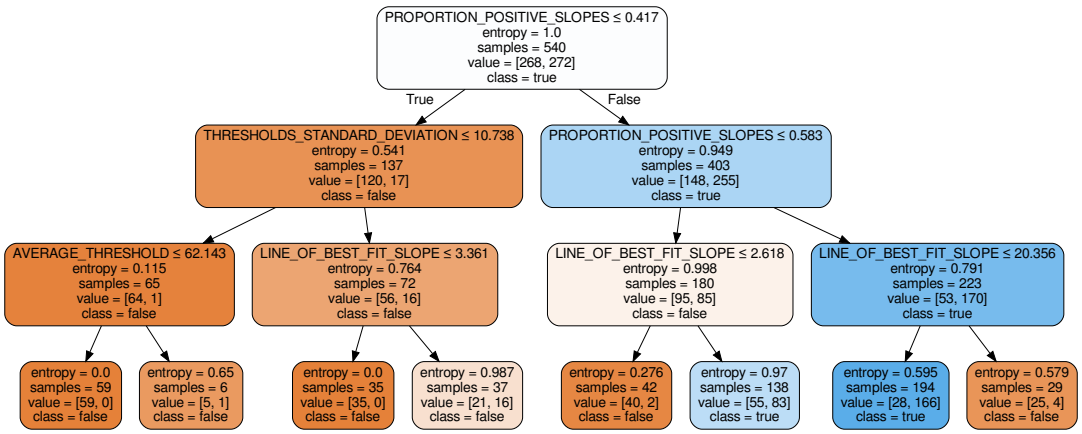


Figure D.2: Decision tree structure for the sloping/not-sloping classifier
This decision tree for the binary classifier classifying an audiogram as sloping or not sloping.

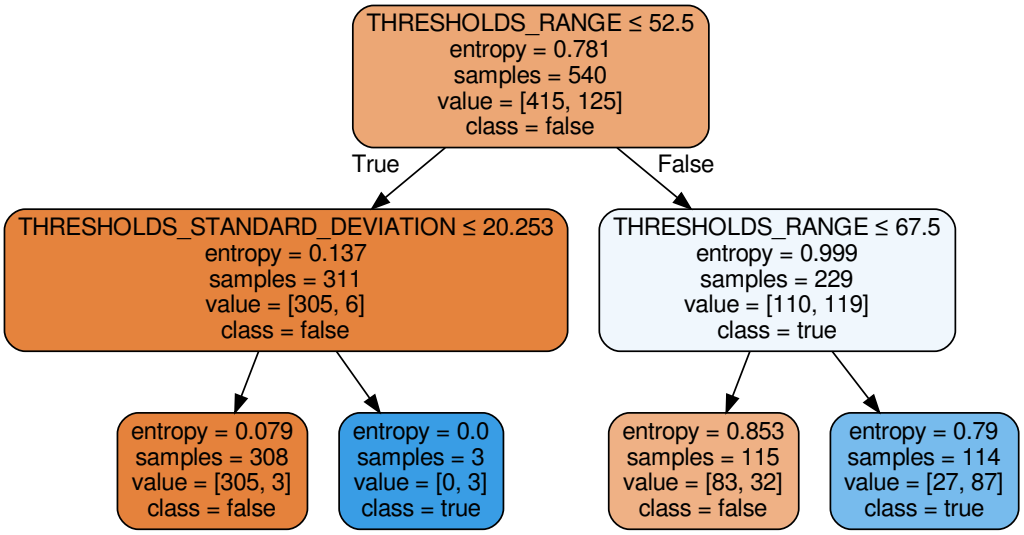


Figure D.3: Decision tree structure for the precipitous/not-precipitous classifier This decision tree for the binary classifier classifying an audiogram as precipitous or not precipitous.

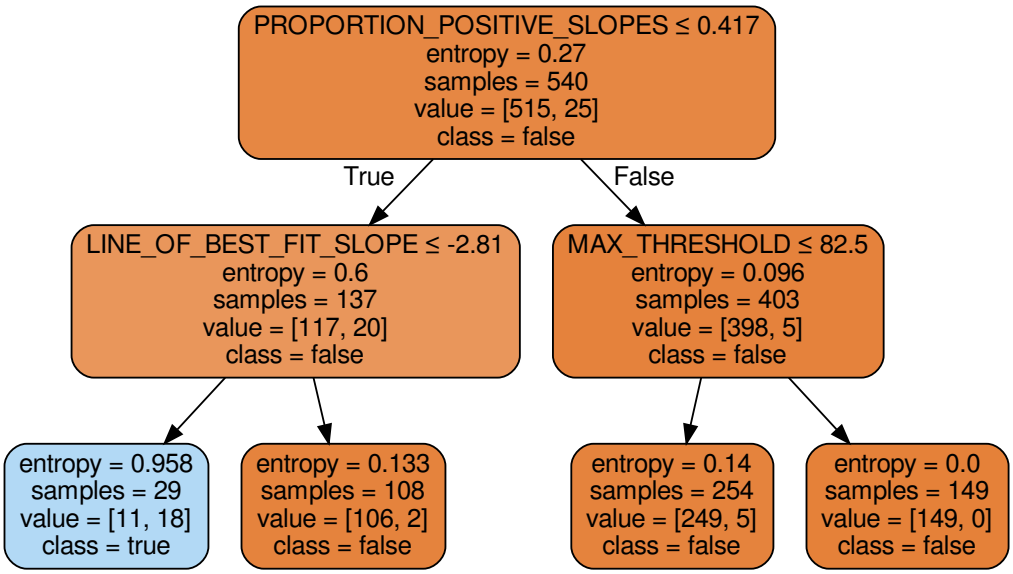


Figure D.4: Decision tree structure for the reverse sloping/not-reverse sloping classifier

This decision tree for the binary classifier classifying an audiogram as reverse sloping or not reverse sloping.

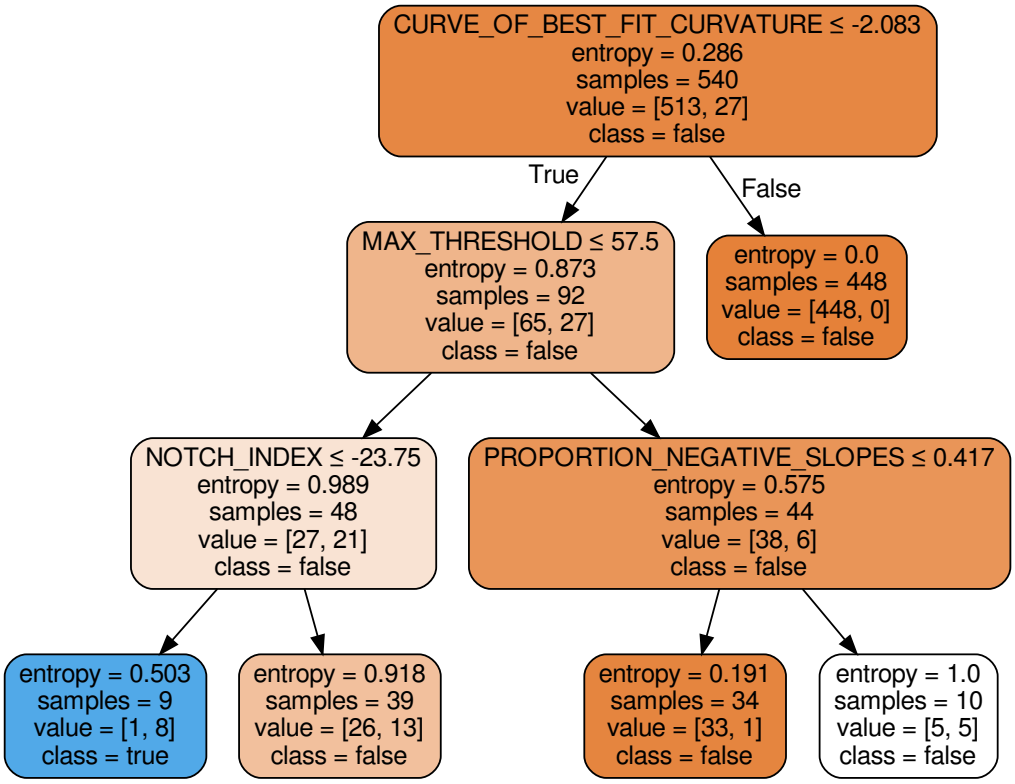


Figure D.5: Decision tree structure for the cookie bite/not cookie bite classifier This decision tree for the binary classifier classifying an audiogram as cookie bite or not cookie bite.

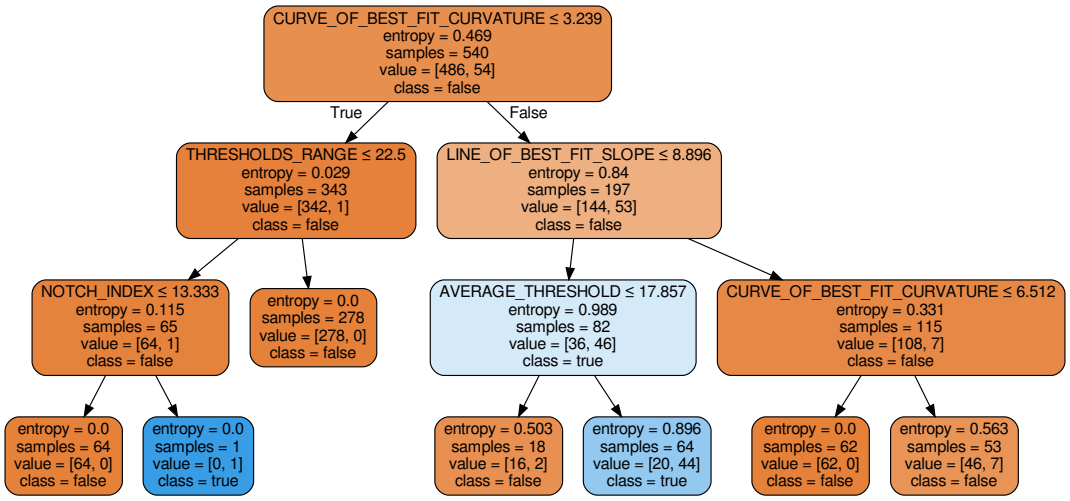


Figure D.6: Decision tree structure for the reverse cookie bite/not-reverse cookie bite classifier

This decision tree for the binary classifier classifying an audiogram as reverse cookie bite or not reverse cookie bite.

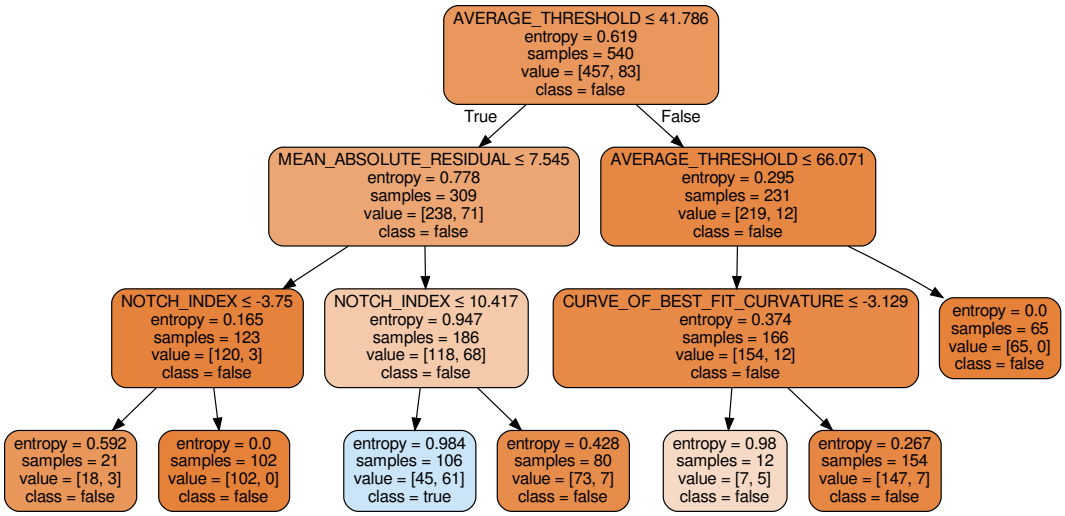


Figure D.7: Decision tree structure for the notched/not-notched classifier
This decision tree for the binary classifier classifying an audiogram as notched or not notched.

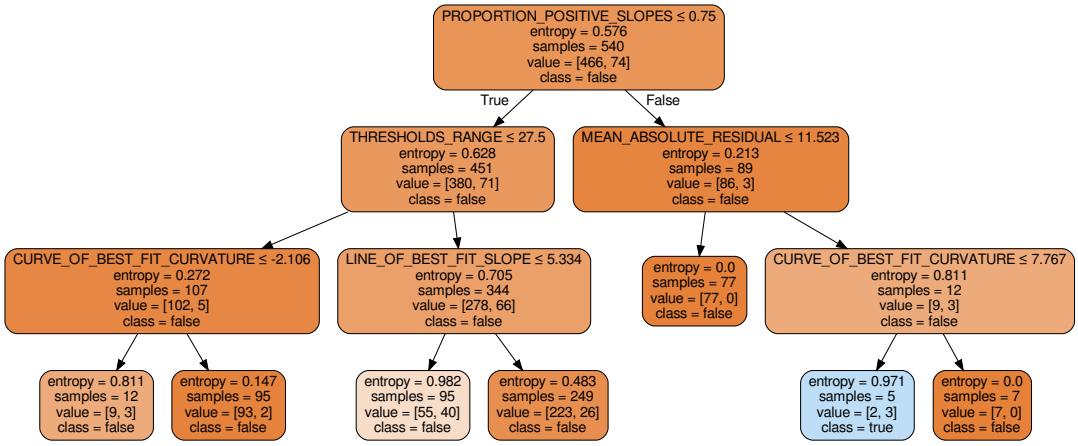


Figure D.8: Decision tree structure for the atypical/not-atypical classifier
This decision tree for the binary classifier classifying an audiogram as atypical or not atypical.

Note that in Tables D.2 to D.9, the feature values for the low range are identical because there is a single measurement at 500 Hz in that range.

Table D.2: Severity feature agreements for the flat configuration

The percentage agreement was computed for the different features and frequency ranges are shown here. The optimal feature(s) for the range of interest is in bold.

Feature	Lows	Mids	Highs
AVERAGE_THRESHOLD	0.735	N/A	N/A
MAX_THRESHOLD	0.382	N/A	N/A
MIN_THRESHOLD	0.662	N/A	N/A
LOW_RANGE_AVERAGE	**0.897***	N/A	N/A
LOW_RANGE_MAX	**0.897**	N/A	N/A
LOW_RANGE_MIN	**0.897**	N/A	N/A
MID_RANGE_AVERAGE	0.75	N/A	N/A
MID_RANGE_MAX	0.647	N/A	N/A
MID_RANGE_MIN	0.676	N/A	N/A
HIGH_RANGE_AVERAGE	0.559	N/A	N/A
HIGH_RANGE_MAX	0.426	N/A	N/A
HIGH_RANGE_MIN	0.647	N/A	N/A
WORST_OF_NOTCH_FREQUENCIES	0.574	N/A	N/A

Table D.3: Severity feature agreements for the sloping configurations

The percentage agreement was computed for the different features and frequency ranges are shown here. The optimal feature(s) for the range of interest is in bold.

Feature	Lows	Mids	Highs
AVERAGE_THRESHOLD	0.227	N/A	0.054
MAX_THRESHOLD	0.0	N/A	0.963
MIN_THRESHOLD	0.908	N/A	0.005
LOW_RANGE_AVERAGE	**0.926**	N/A	0.005
LOW_RANGE_MAX	**0.926**	N/A	0.005
LOW_RANGE_MIN	**0.926**	N/A	0.005
MID_RANGE_AVERAGE	0.443	N/A	0.047
MID_RANGE_MAX	0.261	N/A	0.194
MID_RANGE_MIN	0.712	N/A	0.012
HIGH_RANGE_AVERAGE	0.054	N/A	0.555
HIGH_RANGE_MAX	0.0	N/A	**0.965**
HIGH_RANGE_MIN	0.231	N/A	0.179
WORST_OF_NOTCH_FREQUENCIES	0.048	N/A	0.599

Table D.4: Severity feature agreements for the precipitous configurations

The percentage agreement was computed for the different features and frequency ranges are shown here. The optimal feature(s) for the range of interest is in bold.

Feature	Lows	Mids	Highs
AVERAGE_THRESHOLD	0.043	N/A	0.005
MAX_THRESHOLD	0.0	N/A	**0.976**
MIN_THRESHOLD	**0.947**	N/A	0.0
LOW_RANGE_AVERAGE	0.933	N/A	0.0
LOW_RANGE_MAX	0.933	N/A	0.0
LOW_RANGE_MIN	0.933	N/A	0.0
MID_RANGE_AVERAGE	0.404	N/A	0.0
MID_RANGE_MAX	0.231	N/A	0.144
MID_RANGE_MIN	0.841	N/A	0.0
HIGH_RANGE_AVERAGE	0.0	N/A	0.639
HIGH_RANGE_MAX	0.0	N/A	**0.976**
HIGH_RANGE_MIN	0.053	N/A	0.298
WORST_OF_NOTCH_FREQUENCIES	0.01	N/A	0.764

Table D.5: Severity feature agreements for the reverse sloping configuration

The percentage agreement was computed for the different features and frequency ranges are shown here. The optimal feature(s) for the range of interest is in bold.

Feature	Lows	Mids	Highs
AVERAGE_THRESHOLD	0.19	N/A	0.448
MAX_THRESHOLD	**0.948**	N/A	0.017
MIN_THRESHOLD	0.0	N/A	**0.948**
LOW_RANGE_AVERAGE	0.862	N/A	0.069
LOW_RANGE_MAX	0.862	N/A	0.069
LOW_RANGE_MIN	0.862	N/A	0.069
MID_RANGE_AVERAGE	0.276	N/A	0.414
MID_RANGE_MAX	0.759	N/A	0.069
MID_RANGE_MIN	0.0	N/A	0.776
HIGH_RANGE_AVERAGE	0.0	N/A	0.81
HIGH_RANGE_MAX	0.207	N/A	0.621
HIGH_RANGE_MIN	0.0	N/A	**0.948**
WORST_OF_NOTCH_FREQUENCIES	0.19	N/A	0.586

Table D.6: Severity feature agreements for the cookie bite configuration

The percentage agreement was computed for the different features and frequency ranges are shown here. The optimal feature(s) for the range of interest is in bold.

Feature	Lows	Mids	Highs
AVERAGE_THRESHOLD	0.444	0.045	0.444
MAX_THRESHOLD	0.044	**1.0**	0.022
MIN_THRESHOLD	0.956	0.0	0.956
LOW_RANGE_AVERAGE	**0.978**	0.045	0.933
LOW_RANGE_MAX	**0.978**	0.045	0.933
LOW_RANGE_MIN	**0.978**	0.045	0.933
MID_RANGE_AVERAGE	0.178	0.455	0.178
MID_RANGE_MAX	0.044	**1.0**	0.022
MID_RANGE_MIN	0.756	0.0	0.756
HIGH_RANGE_AVERAGE	0.511	0.114	0.511
HIGH_RANGE_MAX	0.244	0.523	0.222
HIGH_RANGE_MIN	0.956	0.0	**0.956**
WORST_OF_NOTCH_FREQUENCIES	0.178	0.795	0.156

Table D.7: Severity feature agreements for the reverse cookie bite configuration

The percentage agreement was computed for the different features and frequency ranges are shown here. The optimal feature(s) for the range of interest is in bold.

Feature	Lows	Mids	Highs
AVERAGE_THRESHOLD	0.437	0.211	0.207
MAX_THRESHOLD	0.402	0.0	0.828
MIN_THRESHOLD	0.046	**0.987**	0.057
LOW_RANGE_AVERAGE	**0.92**	0.053	0.31
LOW_RANGE_MAX	**0.92**	0.053	0.31
LOW_RANGE_MIN	**0.92**	0.053	0.31
MID_RANGE_AVERAGE	0.092	0.711	0.069
MID_RANGE_MAX	0.425	0.368	0.218
MID_RANGE_MIN	0.046	**0.987**	0.057
HIGH_RANGE_AVERAGE	0.402	0.197	0.31
HIGH_RANGE_MAX	0.299	0.026	**0.92**
HIGH_RANGE_MIN	0.161	0.618	0.115
WORST_OF_NOTCH_FREQUENCIES	0.333	0.25	0.356

Table D.8: Severity feature agreements for the notched configuration

The percentage agreement was computed for the different features and frequency ranges are shown here. The optimal feature(s) for the range of interest is in bold.

Feature	Lows	Mids	Highs
AVERAGE_THRESHOLD	0.667	0.176	0.479
MAX_THRESHOLD	0.0	**0.723**	0.248
MIN_THRESHOLD	0.976	0.25	0.418
LOW_RANGE_AVERAGE	**0.988**	0.243	0.43
LOW_RANGE_MAX	**0.988**	0.243	0.43
LOW_RANGE_MIN	**0.988**	0.243	0.43
MID_RANGE_AVERAGE	0.782	0.209	0.448
MID_RANGE_MAX	0.521	0.318	0.242
MID_RANGE_MIN	0.958	0.25	0.418
HIGH_RANGE_AVERAGE	0.273	0.122	0.388
HIGH_RANGE_MAX	0.109	0.628	0.358
HIGH_RANGE_MIN	0.661	0.149	**0.594**
WORST_OF_NOTCH_FREQUENCIES	0.0	**0.723**	0.248

Table D.9: Severity feature agreements for the atypical configuration

The percentage agreement was computed for the different features and frequency ranges are shown here. The optimal feature(s) for the range of interest is in bold.

Feature	Lows	Mids	Highs
AVERAGE_THRESHOLD	0.46	0.186	0.27
MAX_THRESHOLD	0.206	0.254	0.587
MIN_THRESHOLD	0.302	0.542	0.222
LOW_RANGE_AVERAGE	**0.841**	0.119	0.254
LOW_RANGE_MAX	**0.841**	0.119	0.254
LOW_RANGE_MIN	**0.841**	0.119	0.254
MID_RANGE_AVERAGE	0.413	0.288	0.19
MID_RANGE_MAX	0.381	0.39	0.175
MID_RANGE_MIN	0.317	**0.593**	0.206
HIGH_RANGE_AVERAGE	0.222	0.102	0.508
HIGH_RANGE_MAX	0.111	0.186	**0.714**
HIGH_RANGE_MIN	0.365	0.288	0.349
WORST_OF_NOTCH_FREQUENCIES	0.159	0.288	0.429