Auditory profiles: enabling big data analyses and remote hearing loss characterization

Von der Fakultät für Medizin und Gesundheitswissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

angenommene Dissertation von

Samira Kristina Saak
geboren am 22.05.1996 in Hamburg

Erstgutachter : Prof. Dr. Dr. Birger Kollmeier
Zweitgutachterin: Prof. Dr. Andrea Hildebrandt
Dritter Prüfer : Prof. Dr. Mathias Dietz
Vierte Prüferin : Dr. Mareike Buhl

Tag der Disputation: 16.12.2024

Abstract

Big data analyses in audiology have the potential to uncover patterns of hearing loss that, when linked to audiological findings and treatment recommendations, can significantly provide benefit to audiological patients in the long term. To achieve this, existing datasets must be thoroughly analyzed, and new datasets need to be collected with a sufficient number of participants, which can be achieved efficiently via remote self-testing. Moreover, the development of suitable methods is crucial to integrate and facilitate large-scale datasets. This thesis, therefore, introduces the concept of auditory profiles as a framework for big data analyses and remote hearing loss characterization. Auditory profiles are data-driven representations of patient groups that share similar hearing loss patterns, enabling the identification of novel patterns and relationships within datasets. By leveraging machine learning techniques and remote testing capabilities, this approach can expand the participant pool, facilitate the collection of large-scale datasets and uncover complex relations within the datasets. If these methods are established, patients can be classified into a specific auditory profile even using a minimum set of (remotely performed) hearing tests, which will have a big impact on audiological diagnostics and treatment for the whole population. This thesis consists of three studies that tackle different aspects of big data analytics and remote testing in audiology. The first study proposes a general auditory profile generation pipeline that can generate auditory profiles from a single dataset. The second study extends this pipeline to a federated learning approach, enabling the integration of multiple datasets and the continuous update of the knowledge contained in the auditory profiles. In both studies, classification models are built that can classify patients into the varying auditory profiles. The third study focuses on adapting the matrix sentence test for a smartphone-based implementation to facilitate remote self-testing. It evaluates the feasibility of administering the test via a smartphone and proposes a user-friendly interface for the predominantly elderly target group. In conclusion, this thesis provides a framework for characterizing patient groups across datasets through the profile generation pipeline. Users of remote testing applications can be classified into specific auditory profiles, allowing the additional information contained within these profiles to be used in research and diagnostic processes, and future data collection efforts. In the long term, additional datasets should be integrated to create a comprehensive global auditory profile set that covers the entire audiological patient population. Linking these profiles to audiological findings and treatment recommendations could ultimately enable remote diagnostics for hearing loss.

Zusammenfassung

Die Analyse von großen Datensätzen in der Audiologie hat das Potenzial, Muster von Hörverlusten aufzudecken, die, wenn sie mit audiologischen Befunden und Behandlungsempfehlungen verknüpft werden, audiologischen Patienten langfristig zugutekommen können. Um groß angelegte Datenanalysen durchführen zu können, müssen bestehende Datensätze analysiert und neue Datensätze mit einer ausreichenden Anzahl von Teilnehmern gewonnen werden, was durch mobile Ferntests effizient erreicht werden kann. Anschließend müssen geeignete Methoden entwickelt werden, die sowohl die Integration als auch die Auswertung großer Datensätze ermöglichen. Zu diesem Zweck führt die vorliegende Dissertation das Konzept der Auditorischen Profile als Rahmen für Big-Data-Analysen und die mobile Charakterisierung von Hörverlusten ein. Auditorische Profile sind datengetriebene Darstellungen von Patientengruppen, die ähnliche Hörverlustmuster aufweisen, und ermöglichen die Identifizierung neuer Muster und Beziehungen innerhalb von Datensätzen. Durch den Einsatz von maschinellen Lernverfahren und mobiler Ferntests hat dieser Ansatz das Potenzial, den Teilnehmerpool zu erweitern, die Erfassung großer Datensätze zu erleichtern und Zusammenhänge in den Datensätzen aufzudecken. Wenn diese Methoden etabliert sind, können Patienten auch mit einer minimalen Anzahl von (ferngesteuerten) Hörtests in ein bestimmtes Profil eingeteilt werden, was einen großen Einfluss auf die audiologische Diagnostik und Behandlung der gesamten Bevölkerung haben kann. Diese Dissertation besteht aus drei Studien, die sich mit verschiedenen Aspekten der Analyse von großen Datensätzen und der mobilen Selbsttestung befasst. In der ersten Studie wird eine allgemeine Pipeline zur Erstellung von Auditorischen Profilen entwickelt, die Profile aus einem einzigen Datensatz generieren kann. Die zweite Studie erweitert diese Pipeline um einen föderierten Lernansatz, der die Integration mehrerer Datensätze und die kontinuierliche Aktualisierung des in den Auditorischen Profilen enthaltenen Wissens ermöglicht. In beiden Studien werden Klassifikationsmodelle gebaut, die es ermöglichen, Patienten in eines der Profile zu klassifizieren. Die dritte Studie befasst sich mit der Anpassung des Oldenburger Satztests (Matrix sentence test) für eine Smartphone-basierte Implementierung, um die Selbsttestung aus der Ferne zu erleichtern. Dazu wird die generelle Machbarkeit der Testdurchführung über ein Smartphone evaluiert und eine geeignete Benutzeroberfläche vorgeschlagen, die speziell auf die überwiegend ältere Zielgruppe zugeschnitten ist. Zusammenfassend ermöglicht diese Dissertation, dass Patientengruppen über Datensätze hinweg durch die Profilgenerierungs-Pipeline charakterisiert werden können. Darüber hinaus können Nutzer einer mobilen

Messumgebung spezifischen Auditorische Profile zugeteilt werden, sodass die zusätzlichen Informationen, die in diesen Profilen enthalten sind, für die Forschung, die Diagnostik und für zukünftige Datensammlungsmaßnahmen genutzt werden können. Langfristig sollten weitere Datensätze integriert werden, um ein umfassendes globales Auditorisches Profil zu erstellen, das die gesamte audiologische Patientenpopulation abdeckt. Die Verknüpfung dieser Profile mit audiologischen Befunden und Behandlungsempfehlungen könnte letztendlich Ferndiagnosen von Hörverlusten ermöglichen.

Acronyms

ABG	air-bone gap
ACALOS	adaptive categorical loudness scaling
APs	Auditory Profiles
AUPRC	area under the precision–recall curve
BIC	bayesian information criterion
BILD	binaural intelligibility level difference
CAFPAs	Common Audiological Functional Parameters
CV	cross-validation
DTT	digit-triplet test
FAMD	factorial analysis for mixed data
GOESA	Goettingen sentence test
ILD	intelligibility level difference
LAB	laboratory control session
LOOCV	leave-one-out CV
MICE	multivariate imputations with chained equations
MST	matrix sentence test
oHI	elderly hearing-impaired
OMA	Oldenburg Measurement Application
OVA	one-vs.-all
OVAOVO	OVA and OVO
OVO	one-vs.-one
PTA	pure-tone average
RepCV	10-fold CV repeated 10 times
RMSE	root mean square error
SRT	speech recognition threshold
UCL	uncomfortable loudness level
UI	user interface
WEB	smartphone test session
yNH	younger normal hearing

Contents

1 General introduction

Data is becoming increasingly important in today's society to extract insights and foster knowledge generation. This development is aided by advances in computing powers, machine learning models, and data storage capabilities. To extract valuable insights, analyses need to consider the sampling, the accuracy of the data, and the reasoning behind the analyses. Here, larger datasets can lead to more robust results (Shi, 2022). These developments towards big data analytics have tremendously benefited the medical domain, where maintaining biobanks has facilitated the advancement of personalized medicine, by enabling analyses into prevention, diagnosis, treatment, and monitoring of patients (Kinkorová and Topolčan, 2020; Jagadeeswari et al., 2018; Christensen et al., 2018; Wasmann et al., 2021). The benefits of big data analyses in the medical field are, thus, widely acknowledged and the field of audiology can equally benefit from big data analytics.

Another factor, fostered by the Covid-19 epidemic, is that remote testing has been on the rise (Omboni et al., 2022; Saunders and Roughley, 2021). Remote testing has the potential to provide easy access to hearing health care (Swanepoel et al., 2010) and also facilitate the collection of large-scale datasets. These datasets, in turn, can provide valuable insights into hearing loss patterns and also longitudinal trajectories of hearing loss patterns. To achieve these goals, it is essential that individuals are measured on a comprehensive yet concise test battery of audiological tests, which can also be measured remotely. These tests should not only yield accurate audiological findings, but also encompass enough measures to be able to detect and classify most of the potentially underlying pathologies as well as to ensure robust data for statistical analyses.

1.1 Hearing loss characterization

To characterize hearing loss, different audiological measures exist, and their usage varies according to which institution performs the assessment. An Ear-Nose-Throat (ENT) physician or audiologist, for instance, uses a different set of audiological measures to characterize hearing loss, as compared to a hearing care professional for fitting a hearing aid.

An ENT physician's primary focus in characterizing hearing loss is to diagnose underlying conditions, provide appropriate treatment, or refer patients to spe-

cialized care when needed. Initially, an otoscopy is used to visually inspect the outer ear canal and tympanic membrane, and to assess the presence of active ear diseases, such as ear infections, excessive cerumen, or a perforated tympanic membrane. This is often accompanied by a tympanogram to measure middle ear functioning via the compliance of the tympanic membrane. Next, the type of hearing loss (sensorineural-, mixed-, or conductive hearing loss) is assessed via tuning forks or an audiogram (Hoth and Baljić, 2017; Gelfand, 2016; Akşit and Kösemihal, 2024).

Audiologists are hearing specialists that conduct both objective and subjective measurements to evaluate a patient's hearing status. Subjective measurements include the audiogram, which assesses the general type of hearing loss and the degree of hearing loss across frequencies, as well as speech tests which assess speech understanding in both quiet and noise. Speech tests are highly relevant, as most individuals complaining about hearing loss suffer from reduced speech understanding (especially in noise), which results in a loss of communicative abilities. The Freiburg speech intelligibility test (Hahlbrock, 1953) is one of the most commonly used speech tests in Germany (Gemeinsamer Bundesausschuss, 2021) and employs single words or numbers in the assessment. However, this test has been widely criticized due to test lists that are phonemically imbalanced, vary in their difficulty, and have low sensitivity, among others (Hoth and Baljić, 2017; Baljić et al., 2016; Winkler and Holube, 2016). Better alternatives are tests that use sentences, instead of single words to more adequately represent speech understanding, and at the same time show test list equivalence and high sensitivity. Two candidates for this are the Goettingen sentence test (GOESA, Kollmeier and Wesselkamp (1997)) and the matrix sentence test (MST, Wagener (2004); Kollmeier et al. (2015)). The GOESA uses sentence of every-day-life, and can be used for speech testing, hearing aid indication, and hearing aid assessment. The MST, in contrast, uses sentences that are matrix-based (5x10 words) and follow a fixed semantic structure of name-verb-number-adjective-noun ("Thomas hat fünf grüne Messer", engl. translation: "Thomas has five green knives"). Due to the matrix structure, the sentences are unpredictable and especially suited for repeated use (Kollmeier et al., 2015).

When a patient visits a hearing care professional for a hearing aid fitting, the primary audiological measures used to characterize the hearing loss are the audiogram and a speech test. While the speech test is required for hearing aid indication, its information is rarely used to improve the fitting of the hearing aids. As

a result, hearing aid fittings predominantly rely on audiogram data, overlooking additional factors from audiological tests that could enhance the characterization of the hearing loss and improve the effectiveness of the fitted hearing aid. An example for such an additional factor is the usage of adaptive categorical loudness scaling (ACALOS, Brand and Hohmann (2002)) for loudness compensation in hearing aid fitting.

ENT-physicians, audiologists, and hearing care professionals, thus, use different audiological tools in their daily practice in order to capture different aspects of hearing loss. Regardless, the main audiological test and gold standard is still the audiogram. Research has, however, repeatedly shown that the audiogram alone does not suffice in characterizing individual hearing deficits sufficiently (Musiek et al., 2017; Houtgast and Festen, 2008; Schoof and Rosen, 2014; Humes, 2021; Van Esch and Dreschler, 2015). This highlights the need to incorporate additional measures beyond the audiogram for an accurate data-driven characterization of hearing loss. To cover a broader range of measures, beyond the audiogram, datasets need to be collected that either already contain a variety of measures, or can be combined in a sensible way to use the varying measures, such that a comprehensive yet concise data pool can be obtained for big data analyses in audiology.

1.2 Big data in audiology

Big data is more than just the size of the dataset and is commonly described by the five "Vs": *volume*, *velocity*, *variety*, *veracity* and *value* (Anuradha et al., 2015). *Volume* refers to what one typically conceives as "big" data, namely the size of the data. *Velocity* refers to the speed with which data is added or changed. In audiology for instance, data with high velocity could refer to time-varying signal data from additional sensors in hearing aids. Data *variety* implies that datasets contain varying measures, for instance varying audiological measures next to questionnaire data from cognitive domains or hearing care professionals. *Veracity* refers to the accuracy of the data or potential errors contained in the dataset. Finally, data *value* describes how well the data is suited to extract the desired insights of the dataset (Mellor et al., 2018; Anuradha et al., 2015). If these prerequisites are fulfilled, big data analytics can become a powerful tool for knowledge generation and decision support in the medical field.

In order to analyze and interpret the data, various statistical and machine learning models are available. Traditional research models, such as regression and

classification models, are primarily supervised and used to predict specific outcomes and identify influential factors. Commonly used classification models are random forests and neural nets. Both random forests (Breiman, 2001) and neural nets have their advantages and disadvantages. While neural nets can generally achieve high precision with large datasets, they are often less interpretable and make it more challenging to derive insights about feature importance (Hastie et al., 2009). For smaller sample sizes and improved interpretability, random forests are a more suitable choice, as they are based on decision trees and can effectively handle smaller datasets. This makes them a more attractive option for applications where transparency and understanding of the model's behavior are crucial.

In contrast, unsupervised machine learning models, like clustering, are designed to uncover hidden patterns and relationships within the data, thus offering the potential to reveal previously unknown associations (Hastie et al., 2009). Unsupervised clustering is, therefore, particularly well-suited for identifying patterns of hearing loss in audiological datasets. Clustering enables the stratification of data into distinct clusters, maximizing the distinctions between them. Model-based clustering goes a step further by assuming that the data is generated from a mixture of subgroups based on an underlying model, which it aims to recover (Fraley and Raftery, 2002; Banerjee and Shan, 2010). Hence, this approach allows for the identification of different patterns of hearing loss in a data-driven manner. The detected hearing loss patterns have the potential to provide benefit to audiological patients in the long term, especially when related to subgroup-specific audiological findings and treatments.

Recently, federated learning (McMahan et al., 2017) has emerged as a promising approach to increase the available data pool by overcoming data sharing restrictions due to privacy concerns. With federated learning, data analyses and training of models occur at the data ownership site, allowing for access to data that would otherwise be restricted. This is achieved by sharing only the learned parameters, rather than the underlying data, which preserves data privacy. The combined learned parameters from diverse data sources can then be used to train a global model that benefits from a larger dataset, leading to more robust and accurate analyses. For clustering, however, a sufficiently large initial dataset is still required to produce reliable estimates of the underlying clusters.

1.3 Remote testing

Obtaining large datasets in the traditional way, that means in the lab, is difficult due to the location and time constraints. Participants of such studies need to travel to the site of the lab, which can become infeasible, if a multi-center study is not planned that enables easy access for participants from various regions. In addition, experimenters need to measure each participant individually, which limits the ability to gather large datasets quickly. Remote testing offers a solution for this, allowing research institutes to conduct large-scale and longitudinal studies more efficiently by eliminating the need for travel and experimenter time. This approach can tremendously broaden the participant pool, enabling the collection of large-scale datasets, which in turn facilitates the use of machine learning for knowledge extraction.

In addition to research, clinical practice can also benefit from remote testing, as it has the potential to improve both the access to hearing healthcare and improve healthcare monitoring. Again, the access to healthcare is improved, as tests can be performed remotely, without having to travel to a hearing care professional. This is beneficial for individuals who are underserved with hearing care professionals, as well as individuals who are not mobile enough to frequently travel to a hearing care professional. Healthcare monitoring can be improved in clinical practice by using remote testing for large-scale screening approaches to catch, for instance, disease onset (Mishra et al., 2020). Likewise, it can be used for disease monitoring, when a progression of the disease can be expected (Shaik et al., 2023). In the field of audiology, for example, age-related hearing loss often worsens over time (Fischer et al., 2016). Here, remote testing could first track when a hearing aid would become beneficial. Later it could serve as a quality check for the hearing aid, where the remote tests could indicate when the hearing aid needs to be refitted. For this purpose, the remote-testing application would need to be connected to the hearing aids. Here, Sonova, among other manufacturers, provides an online-tool to check hearing aid candidacy. After hearing aids are purchased, audiograms can be measured via the fitted hearing aids, and remote fitting can be performed (Sonova, n.d.).

To enable the benefits of remote testing, the technology for remote testing needs to be accessible by all. For this, smartphone applications are most suitable, as they are easily accessible, and the majority of the population has access to smartphones (Degenhard, 2024). For remote hearing testing to be applicable in research and clinical practice, relevant audiological measures need to be included

to be able to characterize hearing deficits adequately. For this, information on threshold, speech understanding, and loudness perception is important, as these cover different aspects of hearing loss, and were consistently defined as relevant measures (Sanchez-Lopez et al., 2020; Van Esch and Dreschler, 2015; Buhl et al., 2019). They are, therefore, plausible candidates for inclusion in a remote testing tool, as they can effectively assess users' hearing statuses and be used for the collection of larger datasets.

Traditionally, tests to assess threshold information (audiogram), speech understanding (speech test), and loudness perception (adaptive categorical loudness scaling) are measured in the lab with an experimenter present. The participants can respond to the test material through either verbal reports, button presses, or by selecting answer options on a computer screen. However, to measure these tests on a smartphone, it is crucial to ensure that the interface is accessible and usable by the target group, particularly considering the small screen size of a smartphone.

As hearing loss predominantly affects elderly individuals (World Health Organization, 2021), the main target group for such a smartphone-based implementation will also be elderly. Hence, the small screen size and potential unfamiliarity with smartphones could prove to be problematic for the elderly target group, potentially hindering their ability to accurately complete the tests.

From the three domains (threshold, speech understanding, loudness perception), measures of loudness perception and speech understanding (in noise) can be easily implemented on a smartphone in terms of calibration. Both are suprathreshold measures and therefore less dependent on the absolute level. For measuring loudness perception, adaptive categorical loudness scaling (ACALOS, Brand and Hohmann (2002); Oetting et al. (2014)) can be used. ACALOS measures an individual loudness perception by presenting stimuli with varying sound intensities. Participants can rate the sounds in eleven categories from "not heard" to "too loud" (Brand and Hohmann, 2002). The responses are then mapped to a scale of 50 categorical units (CU) in steps of 5 ("not heard" = 0; "too loud" = 50 CU). The eleven categories have been a compromise between Heller's 50-CU scale and the LGOB-method (loudness growth in 1/2 octave bands) that uses only five to seven categories (Kollmeier, 1997). For a smartphone implementation, these eleven categories would need to displayed.

The audiogram can also be easily implemented on a smartphone in terms of the user interface, as it only requires one to three buttons, depending on the test procedure (Lecluyse and Meddis, 2009; Kaernbach, 1990; Xu et al., 2024). It can, however, only result in valuable results if the smartphone setup (smartphone + headphones) is calibrated. For large-scale studies, where individuals are measured with their own devices, calibration remains difficult, even though there are first attempts to calibrate smartphone setups remotely (Masalski et al., 2016; Scharf et al., 2023). Hence, for the smartphone-based implementation, the main difficulty lies in estimating precise audiograms.

While the Goettingen sentence test (GOESA) is included in the current set of auditory profiles, the matrix sentence test (MST) is more suitable for measuring the speech understanding domain on a smartphone. That is, because the MST is a repeatable sentence test, currently available in more than 20 languages (Hörzentrum Oldenburg gGmbH, n.d.; Kollmeier et al., 2015). Further, the results between GOESA and MST are comparable, if the required training is initially performed (Zinner, 2021). Given this, the MST is the more preferable choice for a smartphone-implementation and could replace the GOESA in the future as the main speech test of the auditory profiles. Sentences of the MST are matrix-based (5x10 words) and follow a fixed semantic structure of name-verb-number-adjective-noun ("Thomas hat fünf grüne Messer", engl. translation: "Thomas has five green knives"). It can either be measured in an open or in a closed version. In the open version, patients verbally report the word they understood to an experimenter; in the closed version participants select the words they understood from the 5x10 matrix on a computer screen. The matrix sentence test, however, is more difficult to implement than the ACALOS or the audiogram. Instead of eleven buttons, as with ACALOS, 50 buttons (5x10 matrix) would need to be displayed on screen for participants to mimic the traditional closed version of the matrix sentence test. This could result in button and font sizes that are too small for the elderly target group, where tactile impairments may be present. It is therefore necessary to investigate appropriate interfaces for the small screen size of smartphones, to integrate the matrix sentence test on a smartphone.

1.4 Auditory Profiles

Comprehensive audiological patient characterization is a well-established goal, and various approaches have been developed to comprehensively describe patients. Traditional methods focused on creating audiological test batteries that

encompass a wide range of measures, enabling individual patients to be characterized across multiple dimensions. For instance, Dreschler et al. (2008) have proposed a comprehensive test battery, which serves as an auditory profile by covering key aspects of hearing ability. By measuring patients across these standardized tests, a comprehensive characterization is possible. Further, the aim was to establish a standardized test battery that would facilitate consistent patient characterization across institutions and countries (Van Esch et al., 2013). Further, Jepsen and Dau (2011) conducted a study involving 13 participants, who underwent an audiogram and additional psychoacoustic masking and discrimination experiments. The goal was to not only quantify individual hearing impairments through test results, but also to apply a model of auditory signal processing and perception. They emphasize the importance of characterizing patients beyond the standard audiogram, as individuals with similar audiograms can exhibit distinct results on other tests.

Subsequently, approaches have focused on deriving patient profiles from test batteries, using either existing, or newly developed test batteries. Lecluyse et al. (2013), for instance, used a test battery that measures absolute threshold, frequency selectivity, and compression, to provide individual auditory profiles that can be visualized. They also defined a normal-hearing reference profile and a hearing-impaired profile for patients with sensorineural hearing loss. Here, they observe that the impaired profile is characterized by significant variability in test scores among patients, implying that it may be composed of multiple distinct sub-profiles.

Building on this foundation, statistical approaches to deriving patient profiles have emerged that use unsupervised learning to extract patient groups from the data, instead of individual profiles. Notably, the standard audiograms by Bisgaard et al. (2010) and the auditory profiles proposed by Sanchez et al. (Sanchez Lopez et al., 2018; Sanchez-Lopez et al., 2020) have demonstrated the potential of this approach. While the standard audiograms rely on a single audiological measure (the audiogram), the auditory profiles by Sanchez Lopez et al. (2018); Sanchez-Lopez et al. (2020) use a comprehensive set of audiological measures. However, the profiles are limited to four profiles, which are assumed to arise from two underlying distortion types. As a result, further sub-profiles may not be identified. Nevertheless, they successfully defined separable patient groups which could benefit from different treatment schemes (Wu et al., 2020), and thus demonstrate the potential for auditory profiling.

In this thesis, we introduce auditory profiles as a framework that can facilitate big data collection, analyses, and remote characterization of hearing loss. Here, auditory profiles describe patient groups within datasets that share similar patterns of hearing loss within a profile but can be distinguished from remaining profiles based on their distributions across audiological measures. As they are derived purely data-driven, they have the potential to uncover patterns of hearing loss that are independent of a priori hypotheses, which could, in the long run, aid in the diagnostic and treatment process of audiological patients. Additionally, they could be used in a remote testing tool for hearing loss characterization as a background classification system, where individuals are tested on certain audiological measures and then classified into one of the auditory profiles to add further profile information to the remote assessment.

However, several aspects need to be addressed to ensure practical usability. First, it should be feasible to derive patient groupings, in the form of auditory profiles, that are audiologically plausible from various datasets. Second, this process should be independent of the underlying feature set, allowing profile derivation even with datasets that vary in included measures. Third, profiles need to be able to integrate multiple datasets in order to cover all distinct patterns of hearing loss and work towards big data analyses in audiology. Fourth, if the profiles are intended to function as a background classification tool, such as in a remote testing application using smartphones, the respective relevant audiological measures need to first be available on a smartphone. Then, it becomes possible to classify users into one of the available profiles. This requires the use of classification models, which can also facilitate feature importance estimation to assess audiological plausibility, and measurement relevance. By employing these classification models, users can be categorized into specific profiles, allowing the knowledge contained within these profiles to be used in the diagnostic and treatment process.

For this purpose, auditory profiles need to cover a broader range of measures beyond the traditional audiogram to capture comprehensive information. The initial auditory profiles, therefore, contain information from the domains of hearing threshold, speech understanding and loudness perception, along with other relevant measures available in the respective datasets used in the analyses. As the number of included measures increases with the incorporation of additional datasets, it is essential that the main measures can be easily measured remotely using a smartphone to facilitate large-scale data collection. However, this would

require addressing challenges related to calibration and developing user-friendly interfaces to ensure accurate and reliable data collection.

1.5 Outline of the thesis

The aim of this thesis is to (1) contribute towards big data analytics in the field of audiology by deriving patient groups from multiple datasets and (2) foster remote testing in audiology. Figure 1.1 illustrates the overarching goals of this thesis, highlighting the connection between big data analysis and remote hearing loss characterization.

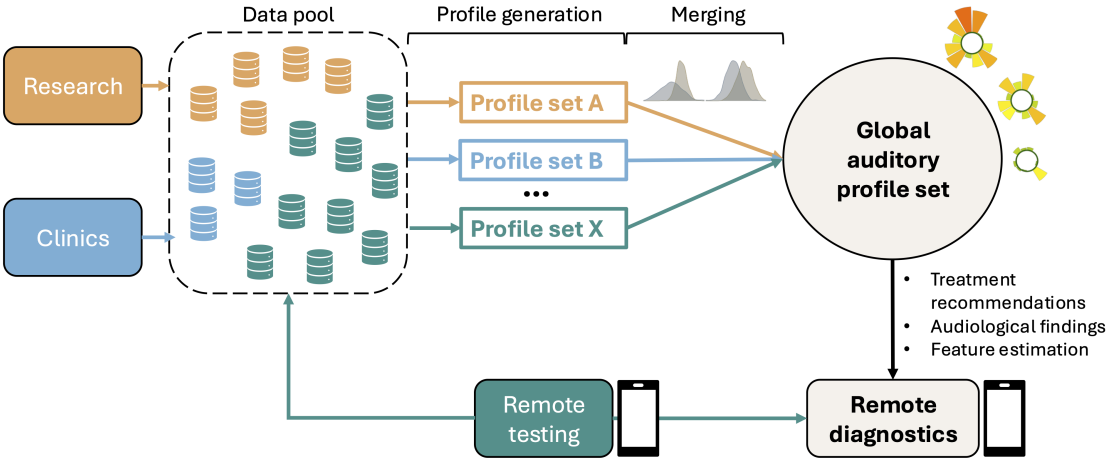

Figure 1.1: Schematic visualization of the auditory profile framework and its connection to remote hearing loss characterization. Auditory profiles can be derived from available datasets in the data pool, including clinical, research, and remote data sources. Remote data sources can significantly increase the data pool, thus facilitating robust data analyses. By merging these profiles, a comprehensive global auditory profile set can be established. This could enable the estimation of treatment recommendations, audiological findings, and missing features, and ultimately facilitate remote diagnostics.

Three different studies were conducted which tackle three different aspects: efficient patient characterization, integration of additional datasets to make use of the large-scale data collection potential, and tailoring a smartphone-based implementation of the matrix sentence test towards the mainly elderly target group to facilitate remote testing.

In the first study (Chapter 2), the general auditory profile generation pipeline is proposed. The introduced profile generation pipeline can generate auditory profiles from a single dataset that contains information on threshold, loudness scaling, and speech understanding. These auditory profiles then describe hearing

loss patterns in the form of auditory profiles and individuals can be classified into one of the auditory profiles using the provided classification models. For the first dataset, 13 auditory profiles were generated ranging from normal hearing to strongly impaired. This study was published as Saak et al. (2022).

In the second study (Chapter 3), the general auditory profile generation pipeline is extended to a federated learning approach such that multiple datasets can be integrated. The aim is to develop a method that enables continuous knowledge integration from available and newly collected datasets into the knowledge container of the auditory profiles. To achieve this, a merging approach was developed that can merge the profiles generated in the first study with a set of newly generated profiles from a second dataset also containing information on threshold, loudness scaling, and speech understanding. This approach should generalize to further datasets that share common information and therefore enables the future integration of remotely collected data from the remote testing application into the auditory profiles. This study was submitted to Trends in Hearing.

In the third study (Chapter 4), a relevant test speech test for remote testing, namely the matrix sentence test, is tailored towards the elderly target population. Here, the general ability to measure the matrix sentence test via a smartphone with household in-ear headphones is assessed. In addition, an appropriate user interface is proposed for the mobile version of the matrix sentence test. This is needed, as the traditional interface does not comply with design guidelines when implemented on the small screen of a smartphone. This study was published as Saak et al. (2024).

Finally, Chapter 5 provides a summary of the main research findings of all three studies and discusses the results in the context of big data analyses and remote testing in audiology.

Bibliography

Akşit, A. M. and Kösemihal, E. (2024). The decision making role of audiological tests in ent practise. *Authorea Preprints*.

Anuradha, J. et al. (2015). A brief introduction on big data 5vs characteristics and hadoop technology. *Procedia computer science*, 48:319–324.

Baljić, I., Winkler, A., Schmidt, T., and Holube, I. (2016). Untersuchungen zur perzeptiven äquivalenz der testlisten im freiburger einsilbertest. *Hno*, 8(64):572–583.

Banerjee, A. and Shan, H. (2010). *Encyclopedia of Machine Learning*, chapter Model-Based Clustering, pages 686–689. Springer US.

Bisgaard, N., Vlaming, M. S., and Dahlquist, M. (2010). Standard audiograms for the iec 60118-15 measurement procedure. *Trends in amplification*, 14(2):113–120.

Brand, T. and Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. *The Journal of the Acoustical Society of America*, 112(4):1597–1604.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Buhl, M., Warzybok, A., Schädler, M. R., Lenarz, T., Majdani, O., and Kollmeier, B. (2019). Common audiological functional parameters (cafpas): statistical and compact representation of rehabilitative audiological classification based on expert knowledge. *International journal of audiology*, 58(4):231–245.

Christensen, J. H., Petersen, M. K., Pontoppidan, N. H., and Cremonini, M. (2018). Big data analytics in healthcare: design and implementation for a hearing aid case study. In *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 296–303. IEEE.

Degenhard, J. (2024). Number of smartphone users worldwide from 2014 to 2029. *Statista*.

Dreschler, W. A., Esch Van, T. E., Larsby, B., Hallgren, M., Lutman, M. E., Lyzenga, J., Vormann, M., and Kollmeier, B. (2008). Charactering the individual ear by the" auditory profile". *Journal of the Acoustical Society of America*, 123(5):3714.

Fischer, N., Weber, B., and Riechelmann, H. (2016). Presbyakusis–altersschwerhörigkeit. *Laryngo-Rhino-Otologie*, 95(07):497–510.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631.

Gelfand, S. (2016). *Essentials of Audiology*. Thieme, 4. edition.

Gemeinsamer Bundesausschuss (2021). Richtlinie des gemeinsamen bundesausschusses über die verordnung von hilfsmitteln in der vertragsärtzlichen versorgung. *Bundesanzeiger*, B3.

Hahlbrock, K.-H. (1953). Über sprachaudiometrie und neue wörterteste. *Archiv für Ohren-, Nasen-und Kehlkopfheilkunde*, 162:394–431.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hoth, S. and Baljić, I. (2017). Current audiological diagnostics. *GMS current topics in otorhinolaryngology, head and neck surgery*, 16.

Houtgast, T. and Festen, J. M. (2008). On the auditory and cognitive functions that may explain an individual's elevation of the speech reception threshold in noise. *International Journal of Audiology*, 47(6):287–295.

Humes, L. E. (2021). Factors underlying individual differences in speech-recognition threshold (srt) in noise among older adults. *Frontiers in Aging Neuroscience*, 13:702739.

Hörzentrum Oldenburg gGmbH (n.d.). Internationale matrixtests. https://www.hz-ol.de/de/diagnostik-matrix.html. Accessed: 2024-01-26.

Jagadeeswari, V., Subramaniyaswamy, V., Logesh, R. t. a., and Vijayakumar, V. (2018). A study on medical internet of things and big data in personalized healthcare system. *Health information science and systems*, 6(1):14.

Jepsen, M. L. and Dau, T. (2011). Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss. *The Journal of the Acoustical Society of America*, 129(1):262–281.

Kaernbach, C. (1990). A single-interval adjustment-matrix (siam) procedure for unbiased adaptive testing. *The Journal of the Acoustical Society of America*, 88(6):2645–2655.

Kinkorová, J. and Topolčan, O. (2020). Biobanks in the era of big data: objectives, challenges, perspectives, and innovations for predictive, preventive, and personalised medicine. *EPMA Journal*, 11(3):333–341.

Kollmeier, B. (1997). *Horflachenskalierung : Grundlagen und Anwendung der kategorialen Lautheitsskalierung fur Hordiagnostik und Horgerate-Versorgung*. Median-Verlag von Killisch-Horn.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., and Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International journal of audiology*, 54(sup2):3–16.

Kollmeier, B. and Wesselkamp, M. (1997). Development and evaluation of a german sentence test for objective and subjective speech intelligibility assessment. *The Journal of the Acoustical Society of America*, 102(4):2412–2421.

Lecluyse, W. and Meddis, R. (2009). A simple single-interval adaptive procedure for estimating thresholds in normal and impaired listeners. *The Journal of the Acoustical Society of America*, 126(5):2570–2579.

Lecluyse, W., Tan, C. M., McFerran, D., and Meddis, R. (2013). Acquisition of auditory profiles for good and impaired hearing. *International journal of audiology*, 52(9):596–605.

Masalski, M., Kipiński, L., Grysiński, T., and Kręcicki, T. (2016). Hearing tests on mobile devices: evaluation of the reference sound level by means of biological calibration. *Journal of medical Internet research*, 18(5):e130.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Mellor, J. C., Stone, M. A., and Keane, J. (2018). Application of data mining to "big data" acquired in audiology: Principles and potential. *Trends in hearing*, 22:2331216518776817.

Mishra, T., Wang, M., Metwally, A. A., Bogu, G. K., Brooks, A. W., Bahmani, A., Alavi, A., Celli, A., Higgs, E., Dagan-Rosenfeld, O., et al. (2020). Pre-symptomatic detection of covid-19 from smartwatch data. *Nature biomedical engineering*, 4(12):1208–1220.

Musiek, F. E., Shinn, J., Chermak, G. D., and Bamiou, D.-E. (2017). Perspectives on the pure-tone audiogram. *Journal of the American Academy of Audiology*, 28(07):655–671.

Oetting, D., Brand, T., and Ewert, S. D. (2014). Optimized loudness-function estimation for categorical loudness scaling data. *Hearing Research*, 316:16–27.

Omboni, S., Padwal, R. S., Alessa, T., Benczúr, B., Green, B. B., Hubbard, I., Kario, K., Khan, N. A., Konradi, A., Logan, A. G., et al. (2022). The worldwide impact of telemedicine during covid-19: current evidence and recommendations for the future. *Connected health*, 1:7.

Saak, S., Huelsmeier, D., Kollmeier, B., and Buhl, M. (2022). A flexible data-driven audiological patient stratification method for deriving auditory profiles. *Frontiers in Neurology*, 13:959582.

Saak, S., Kothe, A., Buhl, M., and Kollmeier, B. (2024). Comparison of user interfaces for measuring the matrix sentence test on a smartphone. *International Journal of Audiology*, pages 1–13.

Sanchez Lopez, R., Bianchi, F., Fereczkowski, M., Santurette, S., and Dau, T. (2018). Data-driven approach for auditory profiling and characterization of individual hearing loss. *Trends in hearing*, 22:2331216518807400.

Sanchez-Lopez, R., Fereczkowski, M., Neher, T., Santurette, S., and Dau, T. (2020). Robust data-driven auditory profiling towards precision audiology. *Trends in hearing*, 24:2331216520973539.

Saunders, G. H. and Roughley, A. (2021). Audiology in the time of covid-19: practices and opinions of audiologists in the uk. *International journal of audiology*, 60(4):255–262.

Scharf, M. K., Schulte, M., Huber, R., and Kollmeier, B. (2023). Microphone calibration estimation for smartphones with resonating beer bottles. DAGA conference 2023, Hamburg.

Schoof, T. and Rosen, S. (2014). The role of auditory and cognitive factors in understanding speech in noise by normal-hearing older listeners. *Frontiers in aging neuroscience*, 6:307.

Shaik, T., Tao, X., Higgins, N., Li, L., Gururajan, R., Zhou, X., and Acharya, U. R. (2023). Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2):e1485.

Shi, Y. (2022). Advances in big data analytics. *Adv Big Data Anal*.

Sonova (n.d.). New technology enables fully-remote hearing aid fittings. https://www.sonova.com/en/story/innovation/new-technology-enables-fully-remote-hearing-aid-fittings. Accessed: 15.09.24.

Swanepoel, D. W., Clark, J. L., Koekemoer, D., Hall Iii, J. W., Krumm, M., Ferrari, D. V., McPherson, B., Olusanya, B. O., Mars, M., Russo, I., et al. (2010). Telehealth in audiology: The need and potential to reach underserved communities. *International Journal of Audiology*, 49(3):195–202.

Van Esch, T. and Dreschler, W. (2015). Relations between the intelligibility of speech in noise and psychophysical measures of hearing measured in four languages using the auditory profile test battery. *Trends in Hearing*, 19:2331216515618902.

Van Esch, T. E., Kollmeier, B., Vormann, M., Lyzenga, J., Houtgast, T., Hällgren, M., Larsby, B., Athalye, S. P., Lutman, M. E., and Dreschler, W. A. (2013). Evaluation of the preliminary auditory profile test battery in an international multi-centre study. *International journal of audiology*, 52(5):305–321.

Wagener, K. (2004). Factors influencing sentence intelligibility in noise. *BIS Verlag*.

Wasmann, J.-W. A., Lanting, C. P., Huinck, W. J., Mylanus, E. A., van der Laak, J. W., Govaerts, P. J., Swanepoel, D. W., Moore, D. R., and Barbour, D. L. (2021). Computational audiology: new approaches to advance hearing health care in the digital age. *Ear and hearing*, 42(6):1499–1507.

Winkler, A. and Holube, I. (2016). Test-retest-reliabilität des freiburger einsilbertests. *Hno*, 8(64):564–571.

World Health Organization (2021). World report on hearing.

Wu, M., Sanchez-Lopez, R., El-Haj-Ali, M., Nielsen, S. G., Fereczkowski, M., Dau, T., Santurette, S., and Neher, T. (2020). Investigating the effects of four auditory profiles on speech recognition, overall quality, and noise annoyance with simulated hearing-aid processing strategies. *Trends in Hearing*, 24:2331216520960861.

Xu, C., Schell-Majoor, L., and Kollmeier, B. (2024). Development and verification of non-supervised smartphone-based methods for assessing pure-tone thresholds and loudness perception. *medRxiv*, pages 2024–06.

Zinner, Christina, W. A. H. I. (2021). Vergleich von fünf sprachtests im sprachsimulierenden störgeräusch. *GMS Zeitschrif für Audiologie (Audiological Acousttics)*, 3.

2 A flexible data-driven audiological patient stratification method for deriving auditory profiles

Bibliographic reference

This chapter is a formatted reprint of the following paper. It is identical in content and has been published in Frontiers in Neurology.

Saak, S., Huelsmeier, D., Kollmeier, B., & Buhl, M. (2022). A flexible data-driven audiological patient stratification method for deriving auditory profiles. Frontiers in Neurology, 13, 959582. https://doi.org/10.3389/fneur.2022.959582

Author contribution

All authors conceptualized and designed the study. S. Saak developed the methodology, programmed the algorithms, performed the analyses, and wrote the first draft of the manuscript. All authors contributed to editing the manuscript.

Prof. Dr. Dr. Birger Kollmeier

Abstract

For characterizing the complexity of hearing deficits, it is important to consider different aspects of auditory functioning in addition to the audiogram. For this purpose, extensive test batteries have been developed aiming to cover all relevant aspects as defined by experts or model assumptions. However, as the assessment time of physicians is limited, such test batteries are often not used in clinical practice. Instead, fewer measures are used, which vary across clinics. This study aimed at proposing a flexible data-driven approach for characterizing distinct patient groups (patient stratification into auditory profiles) based on one prototypical database ($N = 595$) containing audiogram data, loudness scaling, speech tests, and anamnesis questions. To further maintain the applicability of the auditory profiles in clinical routine, we built random forest classification models based on a reduced set of audiological measures which are often available in clinics. Different parameterizations regarding binarization strategy, cross-validation procedure, and evaluation metric were compared to determine the optimum classification model. Our data-driven approach, involving model-based clustering, resulted in a set of 13 patient groups, which serve as auditory profiles. The 13 auditory profiles separate patients within certain ranges across audiological measures and are audiologically plausible. Both a normal hearing profile and profiles with varying extents of hearing impairments are defined. Further, a random forest classification model with a combination of a one-vs.-all and one-vs.-one binarization strategy, 10-fold cross-validation, and the kappa evaluation metric was determined as the optimal model. With the selected model, patients can be classified into 12 of the 13 auditory profiles with adequate precision (*mean across profiles* = 0.9) and sensitivity (*mean across profiles* = 0.84). The proposed approach, consequently, allows generating of audiologically plausible and interpretable, data-driven clinical auditory profiles, providing an efficient way of characterizing hearing deficits, while maintaining clinical applicability. The method should by design be applicable to all audiological data sets from clinics or research, and in addition be flexible to summarize information across databases by means of profiles, as well as to expand the approach toward aided measurements, fitting parameters, and further information from databases.

Keywords: auditory profiles, precision audiology, data mining, machine learning, patient stratification, audiology

2.1 Introduction

It has become increasingly evident that characterizing hearing deficits by the audiogram alone is not enough. In addition to a loss of sensitivity, other factors, such as suprathreshold distortions, determine how well individuals can understand speech in daily life and communicate efficiently (Musiek et al., 2017; Houtgast and Festen, 2008; Schoof and Rosen, 2014; Humes, 2021; Van Esch and Dreschler, 2015). However, it is yet an open issue which measures should be applied to achieve "precision audiology," i.e., to characterize the individual patient as completely and exactly as necessary without losing too much time on comparatively irrelevant measurements. Hence, a number of approaches were described in the literature that differ in their general purpose, their amount of measurements included, and their evaluation method to characterize the most relevant measures.

For instance, Van Esch et al. (2013) proposed a test battery ("auditory profile") for standardized audiological testing comprising eight domains (pure-tone audiometry, loudness perception, spectral and temporal resolution, speech perception in quiet and in noise, spatial hearing, cognitive abilities, listening effort, and self-reported disability and handicap) aiming to describe all major aspects of hearing impairment without introducing redundancy among measures. Similarly, the BEAR test battery was proposed for research purposes to characterize different dimensions of hearing and was evaluated with patients with symmetric sensorineural hearing loss (Sanchez-Lopez et al., 2021). In spite of the benefit of the proposed test batteries, widespread adoption in clinical practice is currently lacking. The complete BEAR test battery, for instance, takes ~ 2.5 h to complete (Sanchez-Lopez et al., 2021), even though a shorter version for clinical purposes was also proposed in (Lopez et al., 2019). Nevertheless, in clinical practice, time is short and the assessment of patients on such a multitude of tests may not be feasible.

To tackle time constraints, Gieseler et al. (2017) aimed at determining clinically relevant predictors for unaided speech recognition from a large test battery, thus, reducing the amount of required tests. They showed that pure-tone audiometry, age, verbal intelligence, self-report measures of hearing loss (e.g., familial hearing loss), loudness scaling at 4 kHz, and an overall physical health score were most important in predicting unaided speech recognition, with the pure-tone audiometry serving as the best predictor. Their model, however, left 38% of the variance in predicting unaided speech recognition unexplained, indicating that further measures may be related to unaided speech recognition. At the same time, their

analyses were tailored toward explaining unaided speech recognition performance as an outcome measure. Predictors for aided speech recognition performance, in contrast, or other outcome measures, may vary. In Lopez-Poveda et al. (2017), for instance, temporal processing deficits as measured by the frequency-modulation detection threshold (FMDT) were shown to be most relevant in predicting aided speech recognition performance. When including only predictors available in clinical situations, however, the unaided speech recognition threshold (SRT) in quiet was determined to be the best predictor. This demonstrates the discrepancy between research and clinical applications and highlights the importance to analyze insights from both clinical and research datasets in combination. It further shows that the relevance of predictors depends on the outcome measures, as different predictors were determined most relevant for unaided and aided speech recognition.

To improve patient characterization in the field of audiology, patient data, therefore, need to be summarized efficiently and flexibly. By summarizing patient data flexibly, the generated knowledge could be used in a variety of settings (e.g., in clinics, for mobile assessments, and decision-support systems in general), and for a variety of outcome measures (e.g., diagnostic outcomes or unaided and aided speech recognition performance). This, however, poses several challenges. First, patients need to be characterized across different dimensions of hearing loss. Second, to gain insights from a diverse patient population, data aggregation across databases is required, which, however, is hindered by the heterogeneity in the applied measures across clinical and research databases in the field of audiology (Buhl et al., 2019). Lastly, for the general applicability of the stored information, it needs to be accessible via measures also applied in clinical settings, such that physicians can be supported.

To tackle these challenges, different approaches toward patient stratification exist that involve identifying subgroups in patient populations based on measurement data from single measures or from interrelations of measures. An example of a data-driven stratification based on single measures is the Bisgaard standard audiograms by Bisgaard et al. (2010). There, a set of 10 standard audiogram patterns occurring in clinical practice were defined. This has subsequently resulted in a variety of studies investigating outcome measures such as aided SRTs in relation to the 10 audiograms [(Gieseler et al., 2017; Dörfler et al., 2020; Folkeard et al., 2020; Kates et al., 2018), to name a few], aiming toward precision audiology, thus, demonstrating the promising nature of finding sub-classes in the field

of audiology. In contrast, an expert-based approach, based on single measures, was proposed by Dubno et al. (2013) that linked four audiometric phenotypes to knowledge about possible etiologies from animal models of presbyacusis via expert decisions. Schematic boundaries for the five phenotypes "older-normal," "pre-metabolic," "metabolic," "sensory," and "metabolic+sensory" are provided which allow for inferences of etiologies, given patient presentations of presbyacusis.

In contrast to patient stratification based on single measures, Sanchez Lopez et al. (2018) introduced a data-driven profiling method based on multiple measures using a combination of unsupervised and supervised machine learning. Based on the hypothesis that two distortion types for the characterization of hearing loss exist, four distinct profiles were generated by means of principal component analysis and archetypal analysis. Thereby, the most important variables for the characterization of each distortion dimension were estimated and employed to identify the most extreme data combinations (archetypes). All patients of two existing research data sets (containing a certain battery of tests) were labeled with the most similar archetype. In a second step, decision trees were built to allow for the classification of new patients into the four auditory profiles. The obtained profiles are interpretable as they were defined based on the hypothesis of two distortion components and the variables used for classification are known. The meaning of the two distortions, however, was different depending on the available measures in the respective data set.

Sanchez-Lopez et al. (2020) improved the profiling method to be more robust (e.g., due to bootstrapping, a more flexible number of allowed variables, and estimating the association of a patient to a profile based on probability) and applied it to the BEAR test battery (Sanchez-Lopez et al., 2021), which was designed for the purpose of including all relevant measures according to the literature and previous work. As a result, a plausible interpretation of the two distortion dimensions was obtained, namely being associated with speech intelligibility and loudness perception, respectively (Sanchez-Lopez et al., 2020). However, by tailoring their analyses toward four extreme distinct profiles and by using archetypal analysis, a priori hypotheses were included in the derivation of the profiles. Consequently, further distinctions between patient groups may be lost.

A further example of summarizing audiological data efficiently is provided by Buhl et al. (2019; 2020). The Common Audiological Functional Parameters (CAF-

PAs) were derived by experts and aim at representing audiological functions in an abstract and measurement-independent way. The CAFPAs further act as an interpretable intermediate layer in a clinical decision-support system. Prediction models allow for a data-driven prediction of CAFPAs (Saak et al., 2020) and a subsequent classification into audiological findings (Buhl, 2022). However, to relate new measures from further data sets to the CAFPAs, experts are currently required for labeling purposes, which consequently does not allow for the automatic integration of new data sets containing additional measures.

The aforementioned methods all contribute toward enhancing patient characterization but are either restricted to single measures or include prior assumptions regarding the distinction of patient groups or audiological functions. Consequently, not all existent differences between patient groups may be detected. In this study, we aim at (1) providing a method for a fully data-driven stratification of patients into subgroups based on audiological measures, namely auditory profiles. This patient stratification approach is not restricted in terms of prior assumptions, the number of patient groups, and contained measures. In that way, all differences between patient groups can be summarized independently of outcome measures. The auditory profiles aim to describe patient groups with similar measurement ranges across audiological measures and are defined based on the contained patient patterns, instead of prior assumptions. In future, profiles could, hence, be combined, added, or removed, depending on the provided insights gained from applying the profiling approach to further data sets, as well as based on the relevance of profile distinctions in clinical routine. The applicability of defined profiles to different settings (e.g., clinical settings) can, however, only be obtained if the knowledge from within the profiles, in the form of plausible ranges for the contained measures, can be linked to patients, given their results on widely used measures (e.g., pure-tone and speech audiometry). We, therefore, further aim at (2) maintaining clinical applicability by building classification models using random forests, based on measures available in clinical routine. This allows for classifying new patients into the auditory profiles. In clinics, it could support physicians to associate a new patient to a profile and in that way exploit statistical knowledge available for the respective profile.

The current study, thus, aims at answering the following two research questions:

RQ1: Does our proposed profiling approach result in a meaningful and distinct grouping (auditory profiles) of patients with respect to important hearing loss

factors contained in the employed data set?

RQ2: Which classification model can provide high precision and sensitivity in classifying patients into the auditory profiles using only a subset of the contained audiological measures?

2.2 Materials and Methods

2.2.1 Data set

To define the first set of auditory profiles, we analyzed an existing data set that was provided by Hörzentrum Oldenburg gGmbH and is described in detail in Gieseler et al. (2017). In contrast to Gieseler et al. (2017), we did not exclude any patients with, e.g., an air-bone gap >10 dB HL but aimed for a diverse patient sample. Our patient sample, consequently, consisted of all patients that completed the full test battery, resulting in 595 patients (*mean age* = 67.6, *SD* = 11.9, *female* = 44%) with normal to impaired hearing. For each patient, information with respect to a broad range of measures, including audiogram data, loudness scaling, speech tests, cognitive measures, and anamnesis questions is contained.

The contained measures either are, or can easily be integrated into clinical routine. The audiogram and the Goettingen sentence test (GOESA, Kollmeier and Wesselkamp (1997)) are commonly used for the assessment of individuals' hearing status. The former assesses an individual's thresholds across frequencies; the latter assesses speech recognition threshold (SRT), here, in noise for the collocated condition (S0N0). Both the audiogram and the GOESA are used in hearing aid fitting, for gain adjustments, and as an outcome measure, respectively. From the contained measures, we used several features to generate the auditory profiles (see Table 2.1 for an overview of the features). For the audiogram, the pure-tone average (PTA, threshold averaged across 0.5, 1, 2, and 4 kHz) for air-, and bone conduction was used for the more severely affected ear. Asymmetric hearing loss was accounted for via the inclusion of an asymmetry score (absolute difference between PTA of left and right ear). Additionally, the air-bone gap (ABG), the PTA of the uncomfortable loudness level (UCL), and the Bisgaard standard audiograms (Bisgaard et al., 2010) were derived from the audiogram. The Bisgaard standard audiograms were included to allow for a separation of different audiogram patterns (e.g., moderately and steeply sloping audiograms), while reducing the dimensionality of the audiogram. A further speech test (digit-triplet test

Table 2.1: Overview of audiological domains and features used for the generation of the profiles.

Domain	Number of features	Features
Audiogram	6	**AC PTA**, BC PTA, **Asymmetry (left/right ear)**, ABG, UCL PTA, **Bisgaard standard audiograms**
Loudness Scaling	6	**ACALOS (L15, L35, L15-L35) for 1.5 & 4 kHz**
Speech tests	3	**GOESA (SRT, slope)**, DTT (SRT)
Cognitive measures	2	DemTect score, WST score
Anamnesis	3	Tinnitus, Socio-economic status, **age**

Features used for the classification into the profiles are shown in **bold**.

(DTT), Smits et al. (2004)) was included to add information to the auditory profiles from a measure mainly used for screening purposes. The adaptive categorical loudness scaling (ACALOS, Brand and Hohmann (2002)) provides relevant information with respect to an individual's loudness perception and recruitment, and has also shown its effectiveness in hearing aid fitting (Oetting et al., 2018). To characterize both the lower and upper part of the loudness curves, both L15, L35, and the difference between L15 and L35 were selected as features. As a relation between cognition and hearing exists (Fulton et al., 2015), the age-normed sum score from a screening test for dementia (Demtect, Kalbe et al. (2004)) and the raw score from a measure of verbal intelligence (Vocabulary test (WST), Schmidt and Metzler (1992)) were also included. Further, information regarding the socio-economic status (sum score of education, income, and occupation) (Winkler and Stolzenberg, 2009), the presence of tinnitus [none (1), unilateral (2), bilateral (3)], and the age of the patients were available.

2.2.2 Generating auditory profiles using model-based clustering

To generate auditory profiles that are capable of separating patients with respect to ranges of audiological tests, we applied clustering, as it has shown promising for purposes of patient stratification. For the current analyses, the clustering pipeline consists of two steps, namely robust learning and profile generation (see Figure 2.1 for visualization).

2.2.2.1 Robust learning

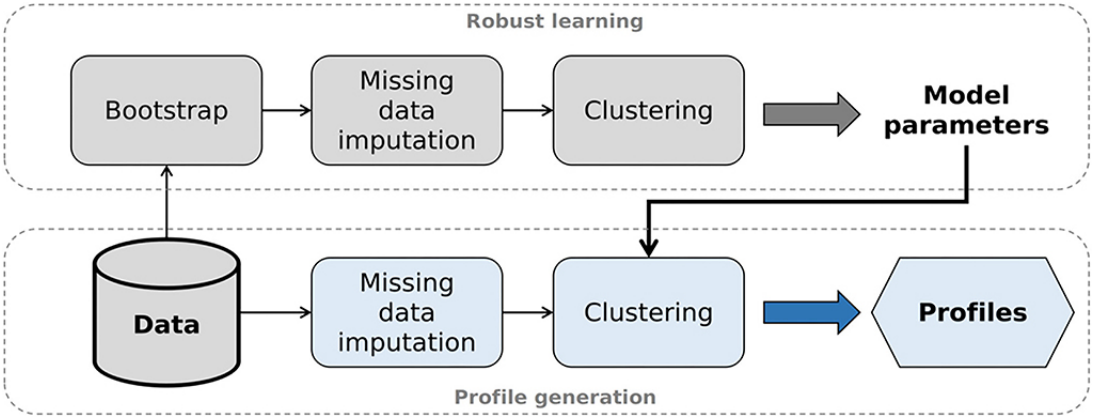

Figure 2.1: Analysis pipeline to generate auditory profiles. After selecting the optimal model parameters (robust learning, upper part), model-based clustering is applied to the original data set (profile generation, lower part).

Bootstrapping and imputation of missing data As bootstrapping techniques have shown to improve the robustness of clustering solutions (Fang and Wang, 2012; Von Luxburg et al., 2010), we first subsampled the data set 1,000 times containing 95% of the original data set. We chose subsampling over resampling with replacement, in order to avoid duplicate samples being seen as a "mini"-cluster, hence, artificially increasing the number of clusters. As missing values existed in the original data set, each of the 1,000 subsamples also contained missing values and needed to be imputed. Missing values pose a common problem in clinical data sets, and a loss of patient information, e.g., complete-case analysis, is often undesirable, thus, requiring an adequate technique to solve it.

Consequently, for audiogram data, prior to extracting pure-tone averages and Bisgaard standard audiograms, missing thresholds were interpolated if the thresholds prior to and after missing values were available. For the remainder of missings (on average 1.5% with a maximum of 2.5%), multivariate imputations with chained equations (MICE) (Azur et al., 2011) was applied. MICE results in multiple completed data sets that account for the uncertainty that stems from imputing missings. With MICE, the analyses of interest are subsequently performed on all completed data sets and the results are combined (Azur et al., 2011). For the present analyses, we generated 20 completed data sets. Accordingly, clustering was performed on each of the 1,000 $\times$ 20 data sets.

Model-based clustering Before clustering, we transformed the features of Bisgaard standard audiograms and tinnitus and treated them as continuous for clustering purposes. Bisgaard standard audiograms were ordered with respect

to increasing PTA; tinnitus with respect to its absence, unilateral, or bilateral presence. All features (see Table 2.1) were then scaled using min–max scaling, resulting in values between 0 and 1. As the number of features ($N = 20$) can be considered small, we refrained from further dimensionality reduction and instead aimed at maintaining a balance of the number of features stemming from the different measures. Depending on the clustering goal, dimension reduction with, e.g., principal component analysis can prove problematic as the reduction of dimensionality could also lead to the removal of information that would have proved to be discriminatory for the clustering goal (Bouveyron and Brunet-Saumard, 2014).

On the scaled feature set, we applied model-based clustering. Model-based clustering was especially suitable for our purposes of uncovering patient groups existent in the data set, as it assumes that the data stem from a mixture of subgroups. The mixture of subgroups is further assumed to be generated by an underlying model which model-based clustering aims to recover (Fraley and Raftery, 2002; Banerjee and Shan, 2010). For this purpose, the number of clusters k and a parameterization of the covariance matrices with respect to their shape, size, and orientation (see Fraley and Raftery (2003) for possible covariance parameterizations) need to be specified beforehand. Subsequently, each cluster's mean vector μ_k and covariance matrix Σ_k is learned and a likelihood estimate for the given clustering solution is computed.

In contrast to simpler clustering techniques such as k-means clustering, model-based clustering is able to detect more complex shapes in the data (Greve et al., 2016). It is, therefore, more suitable for our purposes of detecting all plausible differences in the data. At the same time, the parameterization of the covariance matrices can constrain the complexity of the clustering solution by enforcing stronger restrictions and reducing the number of parameters that need to be estimated (Bouveyron et al., 2019). To select the most suitable model, all candidate parameterizations (k and covariance matrix parameterization) are computed and the model with the highest likelihood of explaining the underlying data structure is selected using the bayesian information criterion (BIC, Schwarz (1978)). More complex clustering structures (i.e., less covariance matrix restrictions) may suffice in explaining the dataset with fewer clusters but require the estimation of a much larger number of parameters and are, thus, not always feasible with smaller datasets. Less complex clustering structures, in contrast, could explain the same underlying data structure by increasing the number of clusters (Bouveyron et al., 2019). This also holds for increasing the number of features used for clustering.

Increasing the number of features increases the number of parameters to be estimated (i.e., the complexity), which, however, can be reduced by restraining the covariance matrices. This may, in turn, increase the number of estimated clusters required to explain the data. To avoid increasing the number of clusters beyond clusters that enhance the explanation of the data structure, however, the BIC penalizes for the complexity of the covariance parameterization and number of clusters k, and thus, results in a trade-off between model complexity and over-parameterization (Fraley and Raftery, 2002).

Here, for each of the 1,000 $\times$ 20 data sets, we computed all potential parameterizations for 2–30 clusters and then derived the optimal model for each data set using the BIC, which resulted in 1,000 $\times$ 20 candidate models. The dimensionality of the candidate models was then reduced across the 20 completed data sets of each of the 1,000 subsamples. The most frequently occurring model parameterization was selected as a candidate model, resulting in a reduced set of 1,000 candidate models. We then defined the overall optimal model via its frequency across the 1,000 candidate models, which resulted in an estimate for the model parameters (i.e., the number of profiles and the model's covariance parameterization).

2.2.2.2 Profile generation In the profile generation step, we generated the auditory profiles using the original data set without prior subsampling. First, we imputed missings using multivariate imputations with chained equations (MICE) in the same manner as described in Section 2.2.2.1. Thus, 20 completed data sets were generated with differing estimates for missings. Second, we applied model-based clustering using the estimated optimal model structure from the robust learning step for each completed data set, which resulted in 20 candidate clustering solutions. From these 20 candidate clustering solutions, we aimed to select the solution showing the highest overlap with the remaining solutions regarding patient allocation into the clusters. The rationale behind this is that, since model parameters are kept constant, differences between clustering solutions stem from differences in the imputed values. The solution showing the highest overlap can then be assumed to be least influenced by imputed values, as patient allocations into the clusters were agreed upon by most solutions.

2.2.3 Building classification models to classify patients into auditory profiles

2.2.3.1 Features and labels To allow for the usage of the auditory profiles for different purposes (e.g., clinical applications), it is necessary to classify patients into the profiles based on a subset of measures widely available. Therefore, we built classification models using the profiles as labels and a reduced set of measures as features. From the aforementioned features used for clustering (see Table 2.1), only the features from ACALOS, GOESA, and the air-conduction audiogram (PTA, Asym PTA, Bisgaard) were used next to the age of the patients (12 features), to simulate the case that these measures were conducted for a to-be-classified patient.

2.2.3.2 Model training For model training, we split the reduced data set, containing the above-mentioned 12 features, into a training (75% of patients) and test data set (25% of patients). The training data set was used for training the model, which included cross-validation (CV), model tuning, and the selection of the best model tuning parameters containing different binarization strategies, CV procedures, and evaluation metrics defining the prediction error, and are described in more detail in the following. The best model is defined as the model minimizing prediction error. We then evaluated the training data set's best model on the test data set to estimate its predictive performance on patient cases not used for model training, which indicates how the classification model would generalize on unseen patient cases.

To build the classification models on the training data set, we used random forests (Breiman, 2001), as it has shown competitive classification performance, while remaining interpretable. It is also less prone to overfitting and handles relatively small sample sizes well (Hastie et al., 2009; Biau and Scornet, 2016). Random forests are an extension of simple decision trees. Multiple decision trees are built, each segmenting the predictor space into several smaller regions, based on derived decision rules. Predictions are consequently derived from the ensemble of trees. For classification purposes, the label predicted most frequently among trees is selected. In other words, it has the highest estimated probability among candidate labels. To avoid building correlated trees, the tuning parameter *mtry* defines the number of features considered at each split. At each split, the specified number of features is then randomly sampled from the feature set, thus, enforcing different tree structures, which in turn reduce the variance of the predictions (Hastie et al.,

2009). For the current analyses, we tuned *mtry* using cross-validation.

To provide optimal prediction models for each of the profiles, we applied different binarization techniques. Binarization strategies to tackle multi-class problems have proved beneficial in enhancing predictive performance. They involve building base learners for binary classification tasks which are subsequently aggregated to provide a prediction (Galar et al., 2011; Adnan and Islam, 2015).

Consequently, we compared multi-class classification to three different binarization strategies. First, we built predictive models for each auditory profile separately (k models), with the one-vs.-all (OVA) technique, allowing the model to learn the specific differences of a profile, as compared to all remaining ones. Thus, for each profile, we built a classification model that decides whether a patient belongs to a given profile, or not. If more than one of the k OVA models predicted that a patient belonged to its profile, the profile with the highest probability among candidate profiles is selected, as defined by the frequency of its prediction in the random forest. Second, we used a one-vs.-one (OVO) technique to build predictive models for all k(k-1)/2 profile combinations. Thus, differences between each pair of profiles were learned. To provide a prediction, voting aggregation was applied, which means that the most frequently predicted profile was selected. Lastly, we used a combination of OVA and OVO (OVAOVO). Here, again, we used OVA to predict profile classes. However, for uncertain cases, if more than one profile was predicted, instead of selecting the profile with the higher probability, we used OVO to decide upon the final profile prediction.

Across profiles, a class imbalance exists, either due to differing profile sizes or due to the applied binarization strategy. Classifiers trained on imbalanced data sets tend to favor the majority class over the minority class in order to reduce the prediction error, which leads to undesirable results if the minority class is of interest (e.g., in an OVA or OVO model). Consequently, we upsampled all profiles to contain at least the number of patients of the largest profile p in terms of sample size *(maxNp)*. Upsampled patients were selected randomly from each profile and across features Gaussian noise was added to the observations (+/- 1 SD). Upsampling with Gaussian noise was shown to be especially suitable for clinical data sets (Beinecke and Heider, 2021). As a result, no class imbalance was present for multi-class and OVO. For OVA, the class imbalance was still present due to the OVA design. As upsampling would require upsampling for several magnitudes of the original profile size, and downsampling would discard

too much valuable information, a different technique was applied. In addition to upsampling to *maxNp*, we used a weighted random forest model using cost-sensitive learning. Thus, weights were introduced, which more severely punished for the misclassification of the minority class over the majority class (Thai-Nghe et al., 2010). The issue of the tendency toward majority predictions was, therefore, addressed also for the OVA binarization strategy.

Further, we compared two different CV schemes for optimal model tuning, namely, leave-one-out CV (LOOCV) and 10-fold CV repeated 10 times (RepCV). LOOCV is a special case of CV, in which the validation set consists of only one observation; RepCV splits the training set randomly into 10-folds, which is then repeated 10 times. LOOCV provides advantages for small data sets, as models are trained on larger sample size as compared to RepCV. However, in return, predictions may have high variance, as the variation in training sets is small. RepCV, in contrast, has lower variance due to differing training sets, but may be biased due to smaller sample size (Hastie et al., 2009).

Lastly, we compared different evaluation metrics which optimize classifiers to different aspects of predictive performance. The main measures to evaluate the performance of a classifier are accuracy, sensitivity, specificity, and precision. Accuracy defines the ratio between correctly classified instances and the total sample size. Sensitivity (also called recall) and specificity are evaluation metrics for binary classification problems, but can be easily extended toward multi-class classification problems by employing an OVA binarization of the classification problem. This, however, again introduces an imbalance in the data regarding the evaluation. Sensitivity refers to correctly classifying all classes of interest as positive, whereas specificity refers to the ability to correctly classify all remaining classes as negative. The precision of a classifier, in contrast, determines the preciseness of a classifier. That means precision is high if no other class was misclassified as the class of interest (Hicks et al., 2022). The four evaluation metrics we compared in the current study, namely, Cohen's kappa, balanced accuracy, F1-score, and the area under the precision–recall curve (AUPRC) differently weight aspects of accuracy, sensitivity, specificity, and precision. Cohen's kappa is inherently capable of evaluating multi-class problems, by comparing the accuracy to the baseline accuracy obtained by chance (Cohen, 1960). Balanced accuracy weights sensitivity with specificity, and is consequently less able to handle multi-class problems, since specificity increases with imbalanced data sets. The F1-score addresses this issue by calculating the harmonic mean between sen-

sitivity and precision, instead of sensitivity and specificity. Likewise, the AUPRC has shown to be especially suitable for imbalanced data (Sofaer et al., 2019). To determine the optimal classifier, it is important to select an adequate evaluation metric, suitable for the class distribution in the data set. Since we have different class distributions across our four classification strategies (multi-class, OVA, OVO, OVAOVO), we compared different evaluation metrics.

2.2.3.3 Model selection and evaluation To select the optimal classification model, we evaluated the four different classification strategies (multi-class, OVA, OVO, OVAOVO) on the training data set with respect to the different metrics (Kappa, balanced accuracy, F1-score, and AUPRC) and cross-validation procedures (repCV, LOOCV). To compare the performance of the models that were optimized with the different evaluation metrics, after training, a general post-hoc performance measure is needed. Here, we chose the F1-score as it summarizes both sensitivity and precision, and can adequately describe the performance of a classifier in case of imbalance. Accordingly, we determined the model leading to the highest F1-score by averaging the F1-scores across profiles and then selected it as the best performing classification model. Lastly, to evaluate the predictive performance of the selected classification model and its generalizability to new data, we evaluated the model on the test data set. Here, instead of the F1-score, we used both sensitivity and precision to provide a more thorough assessment of the classifiers' performance for the distinct auditory profiles.

2.3 Results

2.3.1 Generation of profiles

2.3.1.1 Estimation of profile number and covariance parameters To generate auditory profiles which characterize a diverse range of patient patterns across measures, the number of separable patient groups and the covariance parameter were determined. Figure 2.2 depicts the distribution of estimated cluster numbers across the 1,000 bootstrapped samples. Across bootstrapped samples, 11–19 profiles were estimated as an optimal model with a maximum of 13 clusters. Further, the covariance parameterization "VEI" was selected across all 1,000 subsamples. VEI (variable volume, equal shape, coordinate axes orientation) is a rather parsimonious model as it restricts both the shape and axis alignment of the clusters and requires a diagonal cluster distribution. The sizes of the clusters,

however, may vary. Hence, 13 clusters with the covariance parameter "VEI" are estimated to represent the data structure best.

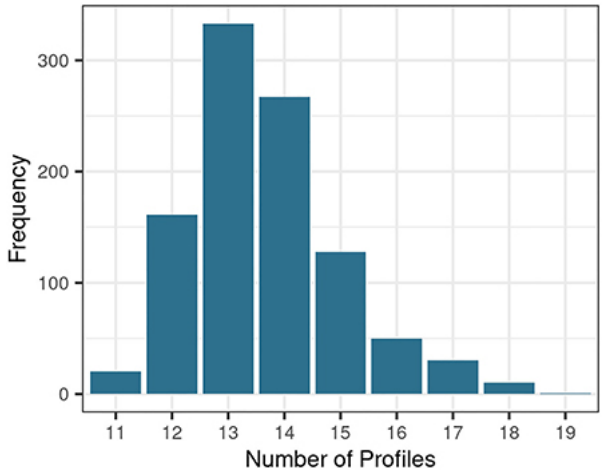

Figure 2.2: Distribution of optimal profile numbers across bootstrapped samples.

Subsequently, the above-defined parameterization ($k = 13$, "VEI") was used to generate profiles on all 20 completed data sets of the original data set. The completed data set showing the highest overlap with the remaining completed data sets regarding patient allocation into the profiles ($max\_similarity = 0.794$) was selected to base the auditory profiles on. Mean classification similarity across all 20 completed data sets was 0.75 ($SD = 0.032$).

2.3.1.2 Profile ranges across audiological measures Figure 2.3 shows the profile ranges of the generated auditory profiles and Table 2.2 contains the number of patients contained in each profile. The profiles cover a large range across audiological measures and show profile-based differences in patient presentation of the contained measures. All profiles can be distinguished from each other based on at least one audiological feature. The speech test results (Figure 2.3, blue box) regarding GOESA and the DTT are generally comparable. The profiles cover different extents of impairments, ranging from normal hearing (profile 1) to strong difficulties in understanding speech in noise (profile 13), as indicated by the increasing SRT. Likewise, the slope of the GOESA decreases with increasing SRT. Within the SRT range of -5 to 0 dB SNR, most of the profiles are contained. Here, the different profiles show similarities regarding SRT ranges, and the difference between the profiles can be found via other measures. Audiogram results (Figure 2.3, green box) indicate the existence of normal hearing (profile 1), moderately (profiles 2, 3, 6, 7, 8, 9, 11, 13), and rather steeply

Table 2.2: Number of patients contained in each auditory profile.

Profile	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
N	27	76	19	24	77	33	6	44	68	51	42	79	39

sloping (profiles 4, 5, 10, 12) patterns. Generally, we observe a trend of increasing thresholds on the audiogram together with increasing SRTs. There are, however, also exceptions. Profile 11 displays the highest thresholds across frequencies and profiles, but does not show the strongest impairment on the GOESA. Instead, it includes patients with an air–bone gap and asymmetric hearing loss, as indicated by the asymmetry score. Profiles can also be distinguished based on the ACALOS (Figure 2.3, loudness scaling—yellow box) and the UCL. With increasing SRTs, we can observe an increase in the UCL, as well as a decrease in the dynamic range, as shown by the difference between L35 and L15 for both 1.5 and 4 kHz. In spite of this, differences exist across profiles unrelated to the increasing SRT. Profiles 4 and 5, for instance, show overlapping ranges regarding the SRT, but differ with respect to the UCL. Across cognitive measures (Figure 2.3, cognitive measures—orange box), no clear distinctions across profiles were found. Likewise, ranges for the age of patients and the socio-economic status (Figure 2.3, anamnesis—gray box) overlap across profiles, with the exception of profile 1 containing younger patients.
To summarize, similarities exist to varying extents between profiles. Some profiles can be easily distinguished. For instance, profiles 1 and 2 can be easily distinguished from profiles 11, 12, and 13 across audiogram, GOESA, and loudness scaling data. In contrast, other profiles only differ on certain measures. Profiles 2 and 3, for instance, show overlapping ranges on both the audiogram and the GOESA, but different average loudness curves and distinct distributions regarding the UCL.

2.3.2 Classification into profiles

2.3.2.1 Model selection To allow for a classification of new patients into the auditory profiles based on a reduced set of measures widely available in clinical practice, classification models were built using random forests. Different parameterizations (optimization metrics, binarization strategies, and CV procedures) were compared with the aim to provide the classification model best suited for the auditory profiles. The *mtry* parameter was inherently determined within each model.

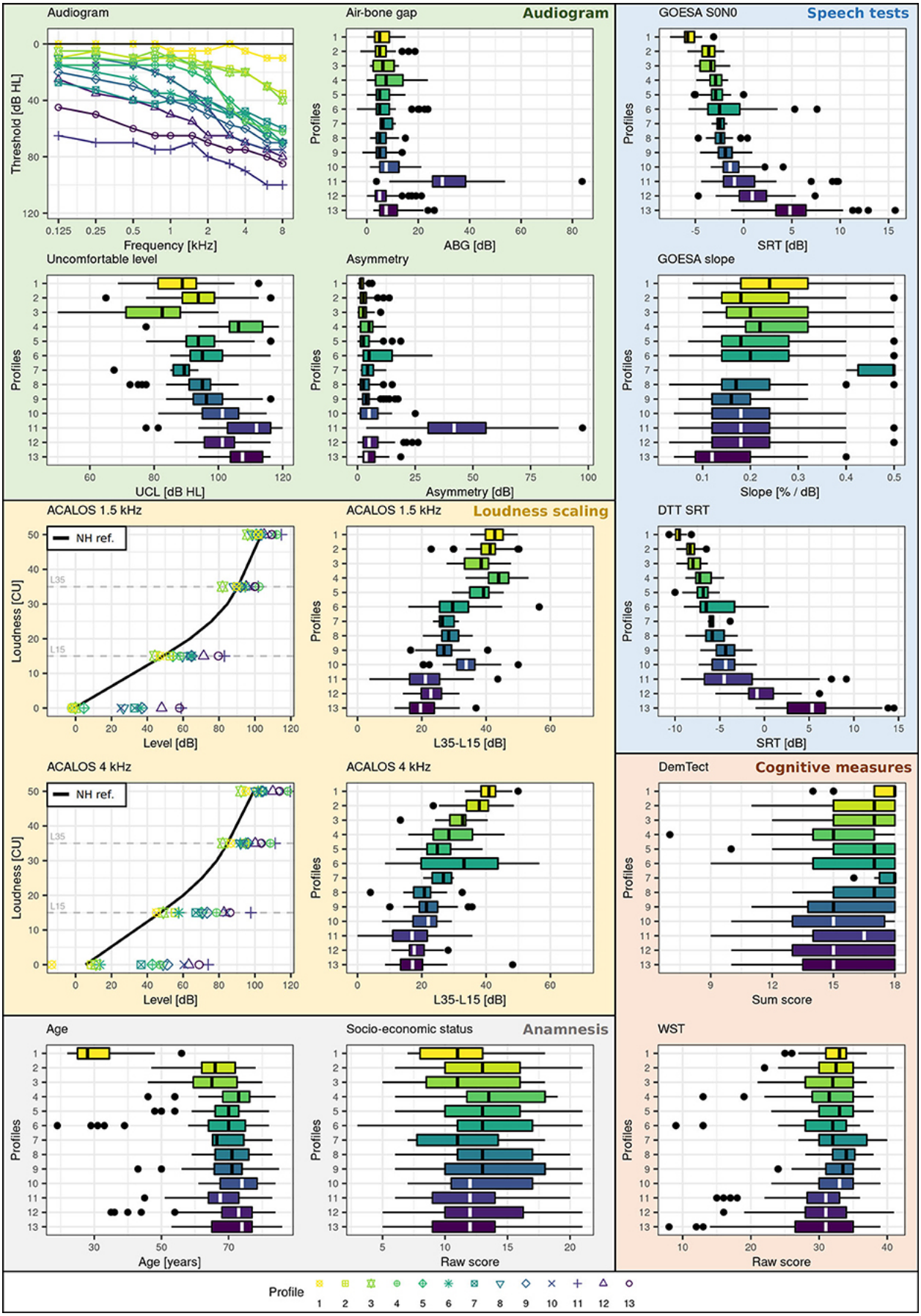

Figure 2.3: Profile ranges across measures. Plot backgrounds are colored according to underlying domains. Blue corresponds to the speech domain, green to the audiogram, yellow to the loudness domain, orange to the cognitive domain, and gray to the anamnesis. Profiles are color-coded (yellow to violet) and numbered (1-13) with respect to increasing SRT (impairment) on the GOESA.

Figure 2.4 displays the results of the comparative performance with respect to the binarization strategies, optimization metric, and cross-validation procedure on the training data set. Model performances with respect to the F1-scores were averaged across profiles to result in an overall F1-score. This allowed for a selection of the best model parameterization. Profile 7 was not selected for averaging, as the number of patients contained in the profile (N = 6) is not large enough to lead to reliable results and interpretations.

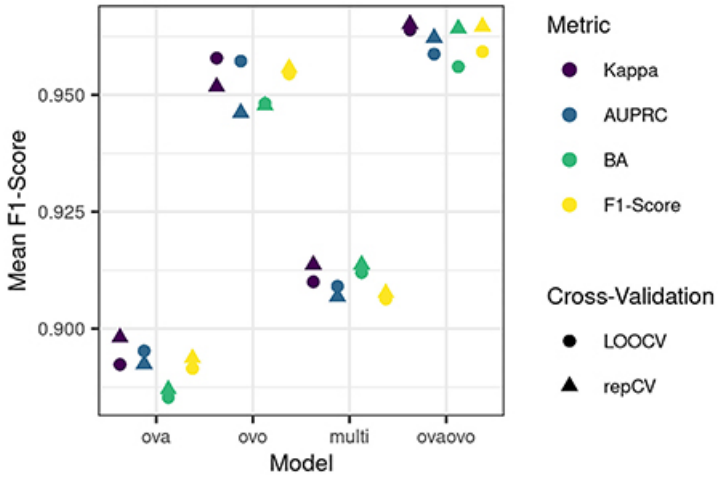

Figure 2.4: Performance of different models on the training data set. The mean F1-score was calculated as the mean of F1-scores across profiles 1–6 and 8–13. Metrics and cross-validation schemes can be distinguished by color and shape, respectively. BA refers to balanced accuracy. LOOCV refers to leave-one-out cross-validation; repCV to repeated 10-fold cross-validation.

All models perform well in predicting profile classes, as indicated by the overall small and high range of mean F1-scores. The highest F1-score was obtained by the OVAOVO model using the kappa evaluation metric and repeated 10-fold CV. Consequently, the OVAOVO (kappa, repCV) model is selected as the classification model to allow for a prediction of patients into profiles. Across models, the kappa metric provided the best results, whereas optimal CV procedures differed across binarization strategies, with the exception of the OVAOVO model in which repCV provided the best results for all evaluation metrics.

2.3.2.2 Model evaluation The previously selected optimal model (OVAOVO, repCV) was selected based on its performance on the training data set (75% of the patients). To investigate the generalizability of the classification model to new patients, its performance was subsequently evaluated on the test data set (25% of the patients). Figure 2.5 displays the performance results with respect

to the sensitivity and precision across all profiles.

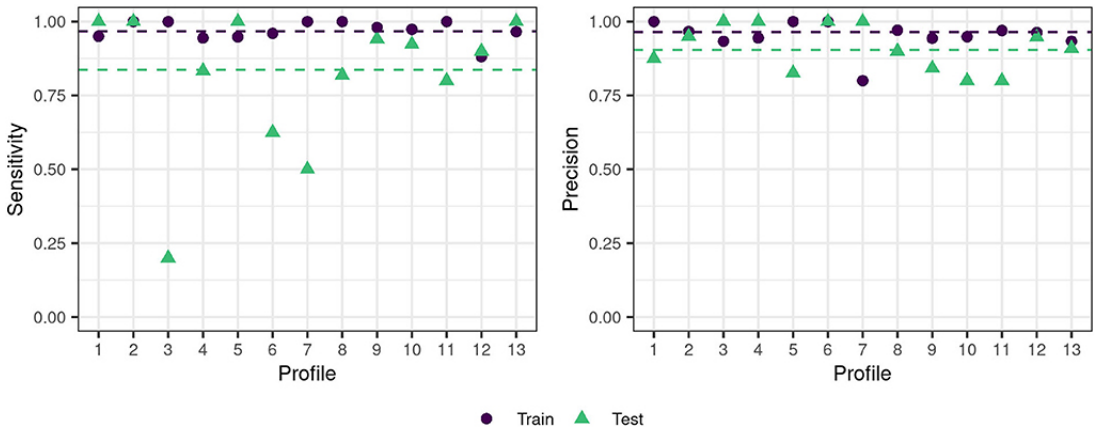

Figure 2.5: Train-test data set performance for the OVAOVO (kappa, repCV model) for both sensitivity and precision. The dashed lines indicate the mean across profiles 1–6 and 8–13 for the respective condition.

Generally, the classifier's performance is adequate regarding achieved sensitivity and precision on the test data set. Across profiles 1–6 and 8–13, average precision and sensitivity on the test data set are 0.9 and 0.84, respectively. Results for profile 7 were plotted for completeness, however, are unreliable due to the small sample size, since the test data set only consisted of two patients. Overall test performance is only slightly lower than training performance for most profiles, except for profiles 3, 6, and 7. For these profiles, the generalization of the learned classification approach toward unseen data is limited. Profile 3 and profile 6 show low levels of sensitivity, but high levels of precision. Thus, not all cases of the two profiles are detected, however, if the two profiles are predicted one can be highly certain that the patient does, indeed, belong to profile 3 or profile 6.

2.4 Discussion

The aim of this study was to propose a flexible and data-driven approach to patient stratification in the field of audiology that allows for a detailed investigation into the combination of hearing deficits across audiological measures. Our results demonstrate the feasibility and efficiency of our proposed profiling pipeline in characterizing hearing deficits in the form of patient groups, namely, auditory profiles. The proposed 13 auditory profiles separate patients with respect to ranges on audiological tests. Further, to ensure the applicability of the auditory profiles in clinical practice with only a basic set of audiological tests, classification models were built that allow for an adequate classification of the auditory profiles

given such a reduced set of audiological measures.

2.4.1 Generation of profiles

The proposed profiles aim to represent the underlying patterns of the current data set best. Hence, the profiles describe the patterns across measures for the available patients and etiologies, rather than aiming to cover all generally existent patient groups with the current set of auditory profiles. Additionally, the number of profiles that can be generated is variable and dependent on the underlying data. This becomes evident when inspecting the distribution of optimal profile numbers in Figure 2.2. Across bootstrapped data sets different profile numbers were suggested. This may in part be due to the applied method. Different subsets of the bootstrapped data may miss extreme patient patterns, and thus, lead to a reduction or increase in suggested profile numbers. This, next to the added uncertainty that stems from imputing missings, may explain the variability in suggested profile numbers across bootstrapped samples. By using a bootstrapping approach, where the optimal number of profiles is defined as the most frequently proposed profile number, it can be assumed, however, that the effects of imputations and extreme patient patterns on the generated profile number were minimized.

The number of profiles may further be influenced by the employed model restrictions. Since the covariance parameterization "VEI" restricts both the shape and axis alignment and requires diagonal cluster distributions, a parsimonious model was selected as describing the underlying data structure best. The number of profiles, therefore, may be large in order to characterize the data structure best with the given restrictions (Bouveyron et al., 2019). It would be of interest to apply the modeling approach to a larger dataset that allows for a less restrictive model in order to investigate if the resultant number of profiles would decrease. A more parsimonious model that leads to a larger number of profiles, however, is in line with our aim of detecting all plausible differences between patient groups.

2.4.2 Interpretation of profiles

The profiles, generally, cover a large range of different types and extents of hearing deficits and appear audiologically plausible. All profiles can be distinguished from each other by at least one audiological feature and can, thus, be considered as

distinct patient groups regarding audiological measures **(RQ1)**. The relevance of the distinction has to be evaluated with respect to the outcome measure of interest. Certain distinctions are, for instance, not necessarily relevant for diagnostic purposes. It can be assumed that profiles 4 and 5 would be categorized as bilateral sensorineural hearing loss (ICD code h90.3) (World Health Organization, 2004) and could, thus, for purposes of coarse diagnostic classification be combined. Profile 5, however, shows a lower range of UCL levels, indicating that loudness would need to be compensated differently in a hearing aid for patients within profile 5 as compared to profile 4. The distinctions regarding loudness perception could influence the benefit that patients within the separate profiles may experience from hearing aids, if the same hearing aid parameters are applied to both groups. This highlights our motivation for flexible profiles that can be combined or separately considered given different outcome measures. The exact number of profiles may, therefore, change with the inclusion of further datasets and also depend on the targeted outcome measure. The proposed auditory profiles, however, enable a detailed investigation into differences that exist between patient groups.

Most of the profiles can be assumed to be caused by symmetrical sensorineural hearing loss. Profile 11, however, also contains an asymmetric conductive hearing loss, as indicated by the presence of both an asymmetry between the ears and an air–bone gap in the group (Isaacson and Vora, 2003). For the remainder of the profiles, however, we can interpret the profiles in the consideration of the four-factor model for sensorineural hearing loss by Kollmeier (Kollmeier, 1999). The current profiles contain measures that allow for an estimation of the first two factors (attenuation and compression loss), but not binaural and central loss. The audiogram can provide an indirect indication for the attenuation loss, which is defined as the required amplification for each frequency to obtain an intermediate loudness perception (L25), whereas the ACALOS can indicate a compression loss via a reduced dynamic range (Kollmeier, 1999). Overall, we can observe differences in both the audiogram shapes and the dynamic ranges across profiles. Most importantly, similar audiogram shapes (e.g., profiles 2 and 3) do not necessarily lead to a similar compression loss and our profiles are able to detect these differences, which is in line with the assumption of the four-factor model, that the audiogram alone cannot explain all underlying characteristics of sensorineural hearing loss. We, therefore, conclude that the 13 auditory profiles provide meaningful information regarding two important factors of hearing deficits, i.e., attenuation and compression loss **(RQ1)**, and that the profiling pipeline has the

potential for the detection of patient group differences also for further datasets, if suitable measures are included.

In general, the interrelation across speech tests, loudness scaling, and audiogram data lead to a separation of patients into profiles. For instance, profiles 2 and 3 contain patients with both similar SRTs and audiogram thresholds. Profile 3, however, shows a reduced dynamic range with its uncomfortable loudness level (UCL) thresholds derived from the audiogram and the range between soft (L15) and loud (L35) sounds on the ACALOS reduced, which indicates recruitment. This, in turn, has implications for hearing aid fitting. It can be assumed that patients within the two profiles require different compression settings, in spite of similar audiograms (Dreschler et al., 2008; Launer et al., 2016). In contrast, the main difference for profiles 8 and 9 lies within their thresholds on the audiogram, with profile 9 showing about 10 dB higher thresholds, while showing similar SRT and loudness curve ranges. The relevance of a distinction between these two profiles, for both diagnostics and hearing aid fitting, thus, needs to be further investigated. For other profiles, differences are more strongly pronounced and they can well be separated.

Certain profiles also align well with the proposed phenotypes by Dubno et al. (2013). Profiles 6, 7, and 9 are consistent with the metabolic phenotype, and profiles 2 and 3 appear to be in between the pre-metabolic and metabolic phenotype with respect to the ranges on the audiogram. Profile 4 can be described in terms of the sensory phenotype and profiles 5 and 10 as the metabolic + sensory phenotype. However, the auditory profiles also contain different patterns, with either more severe presentations as described by the phenotypes (profiles 11 and 13), or different slopes in the lower frequency range of the audiogram (profiles 8 and 12). Further, instead of an older normal hearing profile to match the older normal hearing phenotype, only a young normal hearing profile is included. Regardless, certain probable etiologies can be inferred for the respective profiles, exemplifying how alternate stratification approaches could be connected to the auditory profiles proposed in this study. Since more than one profile can be matched to sensory and metabolic phenotypes, however, it can, again, be assumed that further contributors regarding individual presentations of hearing deficits exist, which are not assessed via the pure-tone audiogram.

No distinctions across profiles regarding the cognitive measures were found (WST, DemTect). Even though hearing deficits and cognitive impairments have been

widely associated (Lin, 2011), the precise causal relationship remains unclear and some studies did not find significant relations (Fulton et al., 2015). With the profiles, a slight trend toward increasing impairment on the DemTect with increasing SRT can be observed; however, the ranges across profiles overlap substantially. On the one hand, this may indicate, that none of the present profiles is significantly influenced by cognitive abilities and that the observed patterns of hearing deficits may occur for both cognitively impaired and non-impaired patients. This would require further investigations and the inclusion of patients with more severe cognitive impairments. On the other hand, the DemTect, as a screening instrument, may not be sensitive enough for detecting a further association between cognitive impairment and hearing deficits. For the auditory profiles, this indicates that cognitive differences are not well-represented, such that patients' cognitive abilities would need to be assessed via further cognitive measures that are currently not included in the database.

The currently available profiles naturally only provide a picture of the contained measures. It can be assumed that the inclusion of further measures will enhance the precision of patient characterization. Of the specified eight domains relevant for characterizing hearing deficits, defined by Van Esch et al. (2013), currently, four are contained in the defined profiles (pure-tone audiogram, loudness perception, speech perception in noise, and cognitive abilities). Spatial contributors, i.e., the intelligibility level difference (ILD) and binaural intelligibility level difference (BILD) measures, were - unfortunately - not included in the original database so no relation to the profiles given here can be provided. However, it can be assumed that they could provide an enhanced characterization of patients' hearing status, as well as prove valuable for hearing aid fitting. Similarly, measures describing the central factor of hearing loss could be incorporated if available in a data set, to comply with all four factors as suggested by Kollmeier (1999). Consequently, future studies should work toward incorporating these measures into the profiles.

2.4.3 Classification into profiles

By building classification models to match patients into the auditory profiles using only features from the air-conduction audiogram, loudness scaling, and GOESA, we aimed for the applicability of the profiles in a variety of settings. First, in clinical routine, both the audiogram and a speech test, measuring the SRT, are the current standard in hearing aid fitting (Hoppe and Hesse, 2017), and in Germany, the GOESA is included in the German guideline for hearing aid fitting

(Gemeinsamer Bundesauschuss, 2021). In addition, loudness scaling has proved promising for hearing aid adjustments (Kiessling, 2001). The three measures are, therefore, often available for hearing professionals and do not extend the testing time of patients and physicians. If fewer measures are available, e.g., only the audiogram and the GOESA, or a different set of measures, the classification models would have to be retrained for this purpose. We believe, however, that loudness scaling provides valuable information for hearing aid fitting and should, thus, be included in the fitting process. Second, to use the profiles in further research and clinical data sets, it is important to include measures that are frequently measured and available. Thus, even though further measures may be contained in the data sets, it is necessary to provide classification models containing measures widely available across data sets.

The present results indicate the feasibility of classifying patients into most of the profiles. The OVAOVO model with the kappa loss function and 10-fold repeated CV reached the highest F1-score and was, therefore, selected as the optimal classification model for the analyzed dataset. With the model test set, sensitivity was >75% for all profiles but profiles 3, 6, and 7 **(RQ2)**. For profile 7, this can be explained by the small sample size of the profile as only six patients were classified into the profile. Consequently, the training of a classifier for profile 7 does not lead to reliable results, and its generalizability is not assured. In spite of that, we included the results for profile 7 for completeness, since it may provide further separation from the remaining profiles for the multi-class classifier, by including counter-examples of patients. Profile 7, however, cannot yet reliably be used to classify new patients into it. Further information from databases is needed to investigate whether this profile represents rare cases or whether this profile was not represented enough in the present data set to provide a large enough sample size for classification purposes. Profile ranges for profile 6 are generally broader than for other profiles; therefore, misclassifications may have occurred more frequently, thus, reducing the sensitivity for profile 6.

The current classification model naturally only covers patient populations that were also contained in the analyzed dataset. Given the adequate classification performance of the classifier, it can be assumed that new patients with similar characteristics to the patients within the dataset would also be adequately predicted into the auditory profiles. At the same time, random forests allow for an estimation of the classification uncertainty when classifying patients into the profiles. This uncertainty estimation refers to how often a patient was predicted

into a given profile across the decision trees of the random forest as compared to the remainder of the profiles. For certain predictions, there is a high amount of agreement of the random forest, whereas for uncertain predictions there is a lower amount of agreement of the random forest. New patients are, therefore, classified into a given profile with an estimate of uncertainty, which, in turn, could also indicate if none of the profiles adequately represents the given patient. This could then reveal a rare patient case or a patient belonging to an additional profile that has not yet been defined. Generally, patients would always be allocated to a profile based on all measures that are contained in the classification model (i.e., audiogram, ACALOS, age, GOESA) and no single feature would determine the classification. For instance, the analyzed dataset contains mainly elderly hearing impaired patients and younger normal hearing patients. Children and younger individuals may, however, also experience hearing deficits. A classification based solely on the feature age would lead to a misclassification into the normal hearing profile 1. The generated classification model, in contrast, would also consider information from the audiogram, ACALOS, and GOESA and in that way avoid misclassification into the normal hearing profile 1.

It can be argued that predictive performance would have been improved by including all measures in the classification models. However, we aimed at providing classification models that can be readily used with measures available across clinics in Germany, such that no additional testing is required and time constraints of physicians are met. Consequently, we decided on a reduced set of measures and aimed at predicting profiles with widely available measures. In future, it may be of interest to provide classification models for all combinations of measures, such that if, e.g., bone-conduction thresholds or more specific psychoacoustic tests are also available in clinical settings, they can be used to increase predictive performance with regard to, e.g., the "binaural" and "central noise" factor (Kollmeier, 1999) involved in characterizing the individual hearing problem.

One limitation of the present classification is the number of patients contained in each profile. For further validation larger and more balanced data sets that also contain more severe patients are required, which can also be assumed to lead to improvements in the predictive performance. An increase in the size of the training set will support the training of the classifier, whereas an increase in the test set will improve the certainty of the predictions. Currently, test performance may have been artificially high for some profiles due to the small sample size in the test set. However, further reducing the training size would also not be

desirable, as it would increase the bias of the classification models. Thus, further evaluations on additional data sets containing further patients are required.

2.4.4 Properties of the profiling approach and comparison to existing approaches

The current data-driven approach toward generating auditory profiles to characterize patient groups is not aimed at being contradictory with hitherto available profiling approaches but aims at providing a more detailed account of existing patient groups and offers several advantages.

First, its flexibility in the definition of profiles derived via purely data-driven clustering allows extending and refining the profiles, if in further data sets more extreme patient representations are contained. More specifically, it can be assumed that applying the profiling approach to additional data sets containing both similar and more extreme patient presentations will result in a set of auditory profiles that show overlap to herein proposed profiles, but also contain additional profiles. The new set of profiles could then be used to update the current set of auditory profiles. As a result, the total number of auditory profiles is not fixed and instead remains flexible to include further profiles. Likewise, the presented profiling pipeline can be applied to additional data sets with varying measures. In case of differing measures across data sets, measures not used for clustering purposes could serve as descriptive features and allow for inference, if these features occur more frequently in certain profiles. The flexibility in terms of derived profiles and contained measures could, in future, aid in comparing patients across data sets. Appropriate means to combine profiles generated on different data sets, however, need to be defined. For this purpose, a profile similarity index based on, e.g., overlapping densities (Pastore and Calcagnì, 2019) could provide a cut-off score on when to combine or extend profiles.

Second, profiles are not tailored toward a certain outcome such as diagnostics or hearing aid fitting. This may, in part, explain the rather large number of generated profiles, since profiles may differ with respect to measurement ranges but not with respect to audiological findings, diagnoses, or treatment recommendations. By tailoring our analyses toward certain outcomes, we could have possibly reduced the number of generated profiles. Our aim, however, was to generate as many profiles as plausibly contained within the data set such that all differences between patient groups can be caught. More specifically, by using Bisgaard stan-

dard audiograms also as a feature for clustering, patients were already separated into 10 distinct audiogram ranges. Combining 10 separate audiogram ranges with different loudness curves and SRT ranges already leads to a larger amount of profiles, if these patterns across measures and patients (i.e., profiles) occur frequently and are well-distinguishable from other profiles. At the same time, the flexibility of the profiles by their definition directly on measurement ranges allows reducing the number of profiles if only certain outcomes are of interest. For instance, if, in future, profiles are connected to diagnostic information from further data sets, profiles leading to a distinction with respect to a diagnosis could be separated or merged. Similarly, if profiles are used for hearing aid fitting, only those profiles leading to separable groups with respect to aided parameters could be retained.

Third, all patients can be grouped into auditory profiles. In contrast, in Dubno et al. (2013), around 80% of audiogram shapes were categorized as non-exemplar and could not be matched into one of the phenotypes, whereas in Sanchez-Lopez et al. (2020), an "uncategorizable" category in addition to the four profiles exists.

A fourth advantage of the flexibility of our auditory profiles pertains to its ability to provide complementary knowledge compared to other profiling approaches, which allows analyzing the same data sets from different perspectives and potentially learning more about the inherent patterns. To exemplify, the profiling approach by Sanchez-Lopez et al. (2020) is applicable to different audiological data sets as well and also comprises the two steps of profile generation and classification. Both approaches are data-driven; however, the approach by (Sanchez-Lopez et al., 2020) is based on the hypothesis of two distortion types which limits the number of profiles to four. In contrast, our approach is purely data-driven, that is, the obtained number of profiles directly depends on the available combinations of measurement ranges in the respective data set, in order to detect all existing differences between patients. Each of our profiles (estimated by model-based clustering) characterizes the group of included patients in terms of underlying measurement data, while the profiles of (Sanchez-Lopez et al., 2020) are characterized by one respective extreme prototypical patient (due to archetypal analysis) and all other patients classified into a respective profile show less extreme results on the variables identified by principal component analysis. The profiles of (Sanchez-Lopez et al., 2020) are interpretable due to the hypothesis of two distortion types and the variables related to each distortion type; however, the obtained interpretation depends on the available measures in the dataset. That means that it needs to be ensured to employ an appropriate database, as was

achieved in Sanchez-Lopez et al. (2020) with the BEAR test battery (Sanchez-Lopez et al., 2021), following the findings of Sanchez Lopez et al. (2018) where the choice of data led to different, not completely plausible interpretations based on the two different analyzed datasets. In contrast, our profiling approach does not include explicit interpretability of every profile yet, but instead, interpretability needs to be added as an additional step. This can be done by relating the profiles to the literature as discussed above, or by including expert knowledge to label the different profiles. In addition, the type of interpretability required for different outcome measures considered in future analyses may be different, and can then be chosen appropriately.

For associating the profiles obtained by the two approaches, in a first step, the distributions of patient data grouped to profiles can be manually compared, for instance regarding audiogram and SRT ranges in Figure 6 of (Sanchez-Lopez et al., 2020) and in our Figure 2.3. However, this comparison is limited as only a small subset of measures is common in the BEAR test battery and our dataset, as well as due to methodological differences as discussed above. Instead, it would be interesting to apply the two profiling approaches to the respective other datasets. As we have GOESA and ACALOS available to characterize speech intelligibility and loudness perception, it would be interesting if the profiling approach of Sanchez-Lopez et al. (2020) also estimates speech intelligibility and loudness as the two distortion dimensions based on our data. Vice versa, the application of our approach to the BEAR test battery would generate a certain number of profiles, which could be compared to the profiles obtained in this study (and thereby to a comparison and potential combination of datasets), as well as reveal measurement combinations leading to sub-classes of the four auditory profiles of Sanchez-Lopez et al. (2020).

2.4.5 Limitations of the profiling approach

Despite the advantages of our purely data-driven profiling approach, certain limitations persist. At the current stage, the profiling approach can detect plausible patient subgroups in data sets. This property generalizes also to further data sets containing different sets of measures and a different patient population. A restriction in the application of the current profiling pipeline to additional databases is the current requirement for continuous or at least ordinal features. Relevant audiological measures may, however, also be categorical with no inherent ordering. Thus, to also incorporate these measures, the current pipeline would need to be

adjusted to also allow for categorical features.

The ability to detect differences in patient groups also depends on the sample size, the contained measures, as well as the presence of distinctive patient groups within the data set. If sample sizes are small, a smaller number of patient groups may be detected in the data sets, which in turn, would be defined by broader ranges across measures. At the same time, this could result in an increase in profiles, each containing only a few patients. This, however, would indicate that the underlying data set is not suitable for the herein proposed profiling approach, as nearly no similarities between patients could be detected. In such a case, it would not be certain whether a profile corresponds to a patient group that could also be identified in larger datasets, or whether it corresponds to outliers in the analyzed data set. Likewise, if only a few measures are contained in new data sets, not all existent distinctions between patients may be detected. Instead, only distinctions regarding the included measures would be available. Combining profiles generated on further data sets with the current profiles may, thus, prove difficult. An estimate of profile "conciseness" could tackle this challenge. This estimate could refer to the average similarity of patients within a profile regarding relevant measures. The similarity between patients with broader profiles will be smaller than the similarity between profiles with smaller ranges across audiological measures. As a result, the conciseness estimate could indicate if the generated profiles on the new data set only result in a coarse grouping of patients. It could then be analyzed, whether the coarse grouping could be explained by a mixture of already available auditory profiles. This would, however, require an overlap between audiological measures across the profiles. If the profiling pipeline is applied to a data set with low overlap regarding measures, the generated profiles would have to be interpreted separately from the current set of profiles, until a relation between measures has been established. This could either occur via available knowledge or by analyzing a data set that contains an overlap between the measures of interest. Regardless, newly generated profiles on further data sets would first need to be analyzed in terms of general audiological plausibility.

At the same time, the relevance of the distinctions between patient groups, in general, and for clinical practice needs further evaluation. This could either comprise asking experts to rate the plausibility and clinical applicability of the distinctions between the profiles or incorporating expert knowledge from other approaches toward patient characterization. The Common Audiological Functional Parameters (CAFPAs) by Buhl (2022), for instance, provide an expert-based concept

of describing patient characteristics; and in Saak et al. (2020), regression models were built to predict CAFPAs based on features that are also available for the current auditory profiles. Hence, the predicted CAFPAs would be available as additional descriptive information for the profiles generated in this study, and a consistency check to previous CAFPA classification (Buhl et al., 2021) could be obtained by analyzing the same data set from different perspectives (i.e., analysis tools). In that way, both approaches provide complementary insights, and both contribute to future combined analysis of different audiological databases. As a result, physicians' trust toward applications (e.g., clinical decision-support systems) using the auditory profiles could be enhanced, which has shown to be a relevant factor in the adoption of such systems in clinical routine (Shibl et al., 2013). Additionally, it can be assumed that the inclusion of more severe patient cases, e.g., with indications for a cochlear implant, could enhance the current profiles toward more extreme profile representations. Currently, profiles can be mostly assigned to mild to moderate hearing loss. With the inclusion of further data sets, containing a higher prevalence of severe patient cases, this aspect could be addressed.

2.4.6 Application and outlook

The herein proposed profiling approach serves as a starting point for uncovering patient groups and patient presentations across audiological measures for the increasingly available amount of larger data sets. Consequently, the proposed profiling approach needs to be applied to additional data sets, which include more severe and diverse patient populations, as well as additional audiological measures to cover further important factors of hearing loss (e.g., binaural and central components). The set of auditory profiles would need to be updated after the inclusion of every further data set by either merging similar generated profiles or adding new profiles. In that way, it would conclude in a final set of auditory profiles, if generated profiles converge. This means that generated profiles on new datasets are already contained in the set of defined auditory profiles and no new information is added, thus, resulting in a final set of auditory profiles describing the audiological patient population.

If the generated auditory profiles describe the audiological patient population, they could be used in a variety of applications due to their flexibility. The profiles could efficiently summarize patient information for a clinical decision-support system. Likewise, they could also support mobile assessments of patients, in e.g.,

a "virtual hearing clinic." If patients are tested on the measures used for the classification models (or appropriate mobile implementations of those measures, ensuring that measurements near the hearing threshold are feasible in realistic environments) they could be classified into a profile. In a clinical decision-support system, physicians could then be provided with statistical insights into patients' hearing statuses, whereas in a virtual hearing clinic patients themselves could receive information regarding their hearing statuses. To also provide diagnostic decision-support as well as aided benefit predictions, however, data from additional data sets containing these measures need to be incorporated into the current profiles. A metric allowing for the combination or separation of profiles, if new profiles are generated on additional data sets, hence, needs to be defined.

After the final set of auditory profiles has been defined, it would also be of interest to define a minimum set of tests that allow for adequate classification of patients into the profiles across data sets. This could highlight the audiological measures that are most relevant across all profiles. Likewise, the profiles could contribute to the selection of the next to-be-performed measures for characterizing the patients. If classification models are available for all measurement combinations, measures leading to the best discriminatory performance across profiles could be selected next. This, in turn, could reduce the testing time of the patients, as well as support the derivation of test batteries covering all relevant aspects of hearing deficits, as in (Van Esch and Dreschler, 2015; Van Esch et al., 2013), by highlighting the most important measures.

2.5 Conclusion

The proposed data-driven profiling approach resulted in 13 distinct and plausible auditory profiles and allows for efficiently characterizing patients based on the interrelations of audiological measures. All patients are characterized and patient groups with certain characteristics, such as asymmetry, are not excluded. Due to the profiles' flexibility by being defined on the contained patients' measurement ranges, profiles could be added or refined, given insights derived from applying the profiling approach to additional data sets. The profiles concur with other profiling approaches but are able to detect differences in patient groups regarding measurement ranges in more detail than hitherto available approaches.

New patients can be adequately classified into the auditory profiles for 12 of the 13 auditory profiles. For 10 profiles, both high precision and sensitivity were

achieved (>0.75), and for two profiles, low to medium sensitivity and high precision were achieved, and for one profile no classification could be achieved due to the profiles' small sample size. Since the classification model was based on a reduced set of measures often available in clinical practice in Germany (GOESA, ACALOS, air-conduction audiogram, and age), clinicians could use the auditory profiles even without performing a complete audiological test battery, if a quick classification with less clinical detail is required. Likewise, all measures required for classifying patients into the auditory profiles are potentially available also on mobile devices, facilitating mobile assessments of the patient.

The proposed profiling approach depends on the underlying data set in terms of the number of profiles or the covered range of patients. Its properties such as flexibility, not being tailored toward a specific outcome, or ability to handle incomplete patient data, however, generalize to other data sets including additional measures. Appropriate means to combine profiles generated across data sets need to be defined.

Future research should extend the profiling toward integrating different data sets with more severe and diverse patient cases. In addition, binaural measures should be included, as well as aided data to investigate hearing device benefits with the profiles.

Data availability statement

The data analyzed in this study was obtained from Hörzentrum Oldenburg gGmbH, the following licenses/restrictions apply: According to the Data Usage Agreement of the authors, the datasets analyzed in this study can only be shared upon motivated request. Requests to access these datasets should be directed to MB, mareike.buhl@uni-oldenburg.de and SS, samira.saak@uni-oldenburg.de. The analyses scripts can be found here: Zenodo, https://zenodo.org/, https://doi.org/10.5281/zenodo.6604135.

Author contributions

SS conducted the data analysis which was continuously discussed with all authors and drafted the manuscript. All authors conceptualized, designed the study, and contributed to the editing of the manuscript.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy – EXC 2177/1 – Project ID 390895286.

Acknowledgments

We thank Hörzentrum Oldenburg gGmbH for the provision of the patient data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Bibliography

Adnan, M. N. and Islam, M. Z. (2015). One-vs-all binarization technique in the context of random forest. In *ESANN*.

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49.

Banerjee, A. and Shan, H. (2010). *Encyclopedia of Machine Learning*, chapter Model-Based Clustering, pages 686–689. Springer US.

Beinecke, J. and Heider, D. (2021). Gaussian noise up-sampling is better suited than smote and adasyn for clinical decision making. *BioData Mining*, 14(1):49.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25:197–227.

Bisgaard, N., Vlaming, M. S., and Dahlquist, M. (2010). Standard audiograms for the iec 60118-15 measurement procedure. *Trends in amplification*, 14(2):113–120.

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press.

Brand, T. and Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. *The Journal of the Acoustical Society of America*, 112(4):1597–1604.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Buhl, M. (2022). Interpretable clinical decision support system for audiology based on predicted common audiological functional parameters (cafpas). *Diagnostics*, 12(2):463.

Buhl, M., Warzybok, A., Schädler, M. R., and Kollmeier, B. (2021). Sensitivity and specificity of automatic audiological classification using expert-labelled audiological data and common audiological functional parameters. *International Journal of Audiology*, 60(1):16–26.

Buhl, M., Warzybok, A., Schädler, M. R., Lenarz, T., Majdani, O., and Kollmeier, B. (2019). Common audiological functional parameters (cafpas): statistical and compact representation of rehabilitative audiological classification based on expert knowledge. *International journal of audiology*, 58(4):231–245.

Buhl, M., Warzybok, A., Schädler, M. R., Majdani, O., and Kollmeier, B. (2020). Common audiological functional parameters (cafpas) for single patient cases: Deriving statistical models from an expert-labelled data set. *International Journal of Audiology*, 59(7):534–547.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Dörfler, C., Hocke, T., Hast, A., and Hoppe, U. (2020). Speech recognition with hearing aids for 10 standard audiograms: English version. *Hno*, 68(Suppl 2):93.

Dreschler, W. A., Esch Van, T. E., Larsby, B., Hallgren, M., Lutman, M. E., Lyzenga, J., Vormann, M., and Kollmeier, B. (2008). Charactering the individual ear by the" auditory profile". *Journal of the Acoustical Society of America*, 123(5):3714.

Dubno, J. R., Eckert, M. A., Lee, F.-S., Matthews, L. J., and Schmiedt, R. A. (2013). Classifying human audiometric phenotypes of age-related hearing loss

from animal models. *Journal of the Association for Research in Otolaryngology*, 14:687–701.

Fang, Y. and Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis*, 56(3):468–477.

Folkeard, P., Bagatto, M., and Scollie, S. (2020). Evaluation of hearing aid manufacturers' software-derived fittings to dsl v5. 0 pediatric targets. *Journal of the American Academy of Audiology*, 31(05):354–362.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631.

Fraley, C. and Raftery, A. E. (2003). Enhanced model-based clustering, density estimation, and discriminant analysis software: Mclust. *Journal of Classification*, 20(2):263–286.

Fulton, S. E., Lister, J. J., Bush, A. L. H., Edwards, J. D., and Andel, R. (2015). Mechanisms of the hearing–cognition relationship. In *Seminars in hearing*, volume 36, pages 140–149. Thieme Medical Publishers.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., and Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, 44(8):1761–1776.

Gemeinsamer Bundesauschuss (2021). Richtlinie des gemeinsamen bundesausschusses über die verordnung von hilfsmitteln in der vertragsärztlichen versorgung. *Bundesanzeiger (BAnz AT 15.04.2021 B3)*.

Gieseler, A., Tahden, M. A., Thiel, C. M., Wagener, K. C., Meis, M., and Colonius, H. (2017). Auditory and non-auditory contributions for unaided speech recognition in noise as a function of hearing aid use. *Frontiers in psychology*, 8:219.

Greve, B., Pigeot, I., Huybrechts, I., Pala, V., and Börnhorst, C. (2016). A comparison of heuristic and model-based clustering methods for dietary pattern analysis. *Public health nutrition*, 19(2):255–264.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., and Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, 12(1):5979.

Hoppe, U. and Hesse, G. (2017). Hearing aids: indications, technology, adaptation, and quality control. *GMS current topics in otorhinolaryngology, Head and Neck Surgery*, 16.

Houtgast, T. and Festen, J. M. (2008). On the auditory and cognitive functions that may explain an individual's elevation of the speech reception threshold in noise. *International Journal of Audiology*, 47(6):287–295.

Humes, L. E. (2021). Factors underlying individual differences in speech-recognition threshold (srt) in noise among older adults. *Frontiers in Aging Neuroscience*, 13:702739.

Isaacson, J. E. and Vora, N. M. (2003). Differential diagnosis and treatment of hearing loss. *American family physician*, 68(6):1125–1132.

Kalbe, E., Kessler, J., Calabrese, P., Smith, R., Passmore, A., Brand, M. a., and Bullock, R. (2004). Demtect: a new, sensitive cognitive screening test to support the diagnosis of mild cognitive impairment and early dementia. *International journal of geriatric psychiatry*, 19(2):136–143.

Kates, J. M., Arehart, K. H., Anderson, M. C., Muralimanohar, R. K., and Harvey Jr, L. O. (2018). Using objective metrics to measure hearing aid performance. *Ear and hearing*, 39(6):1165–1175.

Kiessling, J. (2001). Hearing aid fitting procedures-state-of-the-art and current issues. *Scandinavian Audiology*, 30(1):57–59.

Kollmeier, B. (1999). On the four factors involved in sensorineural hearing loss. In *Psychophysics, physiology and models of hearing*, pages 211–218. World Scientific.

Kollmeier, B. and Wesselkamp, M. (1997). Development and evaluation of a german sentence test for objective and subjective speech intelligibility assessment. *The Journal of the Acoustical Society of America*, 102(4):2412–2421.

Launer, S., Zakis, J. A., and Moore, B. C. (2016). Hearing aid signal processing. *Hearing aids*, pages 93–130.

Lin, F. R. (2011). Hearing loss and cognition among older adults in the united states. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 66(10):1131–1136.

Lopez, R. S., Nielsen, S. G., Cañete, O., Fereczkowski, M., Wu, M., Neher, T., Dau, T., and Santurette, S. (2019). A clinical test battery for better hearing rehabilitation (bear): Towards the prediction of individual auditory deficits and hearing-aid benefit. In *23rd International Congress on Acoustics*, pages 3841–3848. Deutsche Gesellschaft für Akustik eV.

Lopez-Poveda, E. A., Johannesen, P. T., Pérez-González, P., Blanco, J. L., Kalluri, S., and Edwards, B. (2017). Predictors of hearing-aid outcomes. *Trends in hearing*, 21:2331216517730526.

Musiek, F. E., Shinn, J., Chermak, G. D., and Bamiou, D.-E. (2017). Perspectives on the pure-tone audiogram. *Journal of the American Academy of Audiology*, 28(07):655–671.

Oetting, D., Hohmann, V., Appell, J.-E., Kollmeier, B., and Ewert, S. D. (2018). Restoring perceived loudness for listeners with hearing loss. *Ear and hearing*, 39(4):664–678.

Pastore, M. and Calcagnì, A. (2019). Measuring distribution similarities between samples: a distribution-free overlapping index. *Frontiers in psychology*, 10:1089.

Saak, S. K., Hildebrandt, A., Kollmeier, B., and Buhl, M. (2020). Predicting common audiological functional parameters (cafpas) as interpretable intermediate representation in a clinical decision-support system for audiology. *Frontiers in Digital Health*, 2:596433.

Sanchez Lopez, R., Bianchi, F., Fereczkowski, M., Santurette, S., and Dau, T. (2018). Data-driven approach for auditory profiling and characterization of individual hearing loss. *Trends in hearing*, 22:2331216518807400.

Sanchez-Lopez, R., Fereczkowski, M., Neher, T., Santurette, S., and Dau, T. (2020). Robust data-driven auditory profiling towards precision audiology. *Trends in hearing*, 24:2331216520973539.

Sanchez-Lopez, R., Nielsen, S. G., El-Haj-Ali, M., Bianchi, F., Fereczkowski, M., Cañete, O. M., Wu, M., Neher, T., Dau, T., and Santurette, S. (2021). Auditory tests for characterizing hearing deficits in listeners with various hearing abilities: The bear test battery. *Frontiers in neuroscience*, 15:724007.

Schmidt, K. and Metzler, P. (1992). Wst-wortschatztest. *Göttingen: Beltz Test*, 16.

Schoof, T. and Rosen, S. (2014). The role of auditory and cognitive factors in understanding speech in noise by normal-hearing older listeners. *Frontiers in aging neuroscience*, 6:307.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Shibl, R., Lawley, M., and Debuse, J. (2013). Factors influencing decision support system acceptance. *Decision Support Systems*, 54(2):953–961.

Smits, C., Kapteyn, T. S., and Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *International journal of audiology*, 43(1):15–28.

Sofaer, H. R., Hoeting, J. A., and Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, 10(4):565–577.

Thai-Nghe, N., Gantner, Z., and Schmidt-Thieme, L. (2010). Cost-sensitive learning methods for imbalanced data. In *The 2010 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Van Esch, T. and Dreschler, W. (2015). Relations between the intelligibility of speech in noise and psychophysical measures of hearing measured in four languages using the auditory profile test battery. *Trends in Hearing*, 19:2331216515618902.

Van Esch, T. E., Kollmeier, B., Vormann, M., Lyzenga, J., Houtgast, T., Hällgren, M., Larsby, B., Athalye, S. P., Lutman, M. E., and Dreschler, W. A. (2013). Evaluation of the preliminary auditory profile test battery in an international multi-centre study. *International journal of audiology*, 52(5):305–321.

Von Luxburg, U. et al. (2010). Clustering stability: an overview. *Foundations and Trends® in Machine Learning*, 2(3):235–274.

Winkler, J. and Stolzenberg, H. (2009). *Adjustierung des Sozialen-Schicht-Index für die Anwendung im Kinder-und Jugendgesundheitssurvey (KiGGS)*. Number 07/2009. Wismarer Diskussionspapiere.

World Health Organization (2004). *International Statistical Classification of Diseases and related health problems: Alphabetical index*, volume 3. World Health Organization.

3 Integrating audiological datasets via federated merging of auditory profiles

Bibliographic reference

This chapter is a formatted reprint of the following paper, which has been submitted to Trends in Hearing (01.08.2024). It is identical in content.

Saak, S., Oetting, D., Kollmeier, B., & Buhl, M. Integrating audiological datasets via federated merging of Auditory Profiles

A preprint version can be found here: https://doi.org/10.48550/arXiv.2407.20765

Author contribution

S. Saak, B. Kollmeier, and M. Buhl conceptualized and designed the study. D. Oetting helped organizing and structuring the second dataset to make it available for the data analyses and provided knowledge and insights regarding the dataset and general audiological practice. S. Saak developed the methodology, implemented the algorithms, performed the analyses, and drafted the manuscript. All authors contributed to editing the manuscript.

Prof. Dr. Dr. Birger Kollmeier

Abstract

Audiological datasets contain valuable knowledge about hearing loss in patients, which can be uncovered using data-driven, federated learning techniques. Our previous approach summarized patient information from one audiological dataset into distinct Auditory Profiles (APs). To cover the complete audiological patient population, however, patient patterns must be analyzed across multiple, separated datasets, and finally, be integrated into a combined set of APs.

This study aimed at extending the existing profile generation pipeline with an AP merging step, enabling the combination of APs from different datasets based on their similarity across audiological measures. The 13 previously generated APs ($N_A = 595$) were merged with 31 newly generated APs from a second dataset ($N_B = 1272$) using a similarity score derived from the overlapping densities of common features across the two datasets. To ensure clinical applicability, random forest models were created for various scenarios, encompassing different combinations of audiological measures.

A new set with 13 combined APs is proposed, providing well-separable profiles, which still capture detailed patient information from various test outcome combinations. The classification performance across these profiles is satisfactory. The best performance was achieved using a combination of loudness scaling, audiogram and speech test information, while single measures performed worst.

The enhanced profile generation pipeline demonstrates the feasibility of combining APs across datasets, which should generalize to all datasets and could lead to an interpretable population-based profile set in the future. The classification models maintain clinical applicability. Hence, even if only smartphone-based measures are available, a given patient can be classified into an appropriate AP.

Keywords: auditory profiles, audiology, big data, data mining, machine learning

3.1 Introduction

Audiological datasets contain valuable knowledge about patients with hearing loss that can be exploited to learn about patterns in the data, for instance, for identifying patient groups that exhibit similar combinations of audiological test outcomes and may therefore benefit from a similar treatment. Data-driven techniques allow assessing these feature combinations without prior knowledge

about audiological findings or diagnostic information. Performing such analyses on large-scale datasets has received increasing attention across medical fields as well as by policy-makers, due to its potential for precision medicine, the identification of risk factors for diseases, and health care quality control ((Sinkala et al., 2020; Stöver et al., 2023; European Commision and Food Safety, 2016), to name a few). In the domain of audiology, available data for large-scale analyses is currently spread across institutions. While research institutes often have smaller datasets available and may share them openly, larger datasets are available in clinics, but these are restricted in terms of access. Further, clinical or research institutions generally obtain data from patient groups in line with the respective institutions purpose or purpose of the respective study which may lead to considerably different patient populations and audiological measures across data sets. For example, a cochlear implant clinic will have a larger proportion of severe hearing deficits in its datasets than an ambulant audiological center, which will be frequented more often by hearing aid candidates. Even more variability across datasets comes into play if online remote testing with smartphones is included, allowing for data collection outside labs and clinics with a population exhibiting a mild-to-moderate hearing loss. Hence, a comprehensive data analysis describing the complete audiological patient population requires the combination of available, but distributed data across institutions without having the access to all databases simultaneously, since the principle of federated learning has been shown to overcome the problem of distributed data ownership (McMahan et al., 2017). This study introduces the « federated merging of Auditory Profiles » by demonstrating the feasibility of sequentially extracting information from two different databases and then merging the resulting information in a second step.

Varying approaches exist to describe the audiological patient population. These approaches range from epidemiological studies regarding the prevalence of hearing loss in combination with different demographic factors (do Carmo et al., 2008; Roth et al., 2011), to developing optimal test batteries to characterize hearing deficits (Sanchez-Lopez et al., 2021; Van Esch et al., 2013), and advanced analyses into audiological groupings (Bisgaard et al., 2010; Saak et al., 2022; Sanchez-Lopez et al., 2020). Bisgaard et al. (2010) grouped audiograms into ten standard audiogram patterns. However, as no additional measures are considered, a detailed description based on multiple domains of hearing, including, e.g., suprathreshold auditory processing deficits like Plomp's (Plomp, 1986) "D-component", is not possible. Sanchez-Lopez et al. (2020) applied principal component analysis and archetypal analyses to multiple audiological measures

to characterize patients into four distinct auditory profiles. By defining the auditory profiles based on the hypothesis of two underlying distortion types they limit the number of profiles to four. Saak et al. (2022) proposed a profile generation pipeline which resulted in a first set of 13 Auditory Profiles (APs). Again, APs were generated based on multiple audiological measures (e.g. speech testing, loudness scaling, audiogram). In contrast to Sanchez-Lopez et al. (2020), the number of APs is not chosen based on a hypothesis, but dependent on the patient population of the underlying dataset, hence, completely data-driven. APs are described in terms of data distributions across measures, which has the potential to integrate additional profiles into a combined set of APs. If sufficient datasets are accounted for, such a combined set of APs could ultimately lead to robust, comprehensive AP set that can describe the complete audiological patient population independent of a respective audiological dataset that is either present or absent in the underlying profile generation process.

To create such a comprehensive AP set that accurately describes the patient population, it is necessary to compare and integrate multiple datasets, leveraging the growing amount of available data, including various subtypes of hearing deficits as well as different information collected in the respective dataset. Hence, a suitable tool must be capable of integrating APs from different datasets, while ensuring interpretability, flexibility, applicability to clinical practice, as well as complying with data privacy and protection regulations by implementing the federated learning principle (McMahan et al., 2017).

First, the process to combine profiles generated across datasets, as well as the resulting profiles need to remain interpretable to ensure the plausibility and applicability of the profiles in clinical practice. In clinical practice, for instance, if patients are classified into the APs, the APs could serve as a look-up table to obtain an estimation of the results on further audiological tests included in the AP but not measured by the practitioner. Only if their generation is interpretable and the resulting combined profiles are plausible, valuable insights can be obtained from these profiles. Finding such interpretable subgroups or subtypes of diseases via big data analyses has proven valuable and transformative for medical research, and precision medicine (Grant et al., 2020; Parimbelli et al., 2018). For example, molecular profiling has allowed to identify two pancreatic cancer subtypes with subtype-specific biomarkers. The two subtypes have different clinical outcomes and may be more responsive to different subtype-specific treatments (Sinkala et al., 2020). In a similar way, APs could serve to identify

subgroup-specific causes and treatment recommendations, for instance, by highlighting patient groups that exhibit a similar audiogram but may benefit from different fitting rationales for the hearing aids or by prompting sub-group specific research.

Second, it is important to achieve a flexible process for combining APs, such that they can be used in a variety of settings. For instance, for research purposes certain distinctions across audiological measures may be of interest, whereas for clinical applications a coarser separation may suffice. More detailed profiles would represent subtypes of coarser profiles and could, thus, inform on potential subgroups of individual profiles. Hence, providing a method that can make profiles more detailed or coarser based on the individual use case helps in the applicability of the profiles in practice.

Third, for APs to be useful in audiological/clinical practice, the identified subgroup-specific causes and treatment recommendations need to be accessible to practitioners, such as clinicians, hearing care professionals, and researchers. Hence, different classification models are required that are based on different subsets of audiological measurements (features) corresponding to the needs of the respective clinicians or researchers. To exemplify, hearing care professionals in Germany, thus far, mainly use the audiogram and the Freiburger speech test in quiet in their hearing aid provision process, as these are required for a hearing aid indication and hearing aid approval for the reimbursement of the health care insurance (Gemeinsamer Bundesausschuss, 2021), even though research indicates that loudness scaling can be beneficial in hearing aid fitting (Oetting et al., 2018). Smartphone apps, in contrast, can easily perform loudness scaling, but measuring bone conduction thresholds without special equipment is not as straightforward, even though conductive hearing loss can be estimated with the antiphasic digits-in-noise test (DIN, De Sousa et al. (2022)).

Finally, the process to combine and compare profiles needs to be able to also handle sensitive data, such that clinical datasets can be used that underlie data restrictions (European Parliament, 2016). Here, federated learning (McMahan et al., 2017; Pfitzner, 2021) could serve as a solution for data privacy restrictions, which is closely related to distributed computing. Distributed computing originally stems from distributing computing tasks across connected computers to achieve better performance (Hajibaba and Gorgin, 2014), but has also found its applications in machine learning and cloud computing for health care (Beyan

et al., 2020; Ehwerhemuepha et al., 2020). Federated learning makes use of the decentralized approach of distributed computing by aggregating results of locally trained models. This tackles data privacy issues, as data-sensitive computing could occur at the server of the sensitive data location without the need of sharing sensitive data. That means, APs could be generated at a specific institution, and the AP information could then be shared to update the already existing APs. In other words, datasets could be integrated, and their insights extracted via the APs without having to merge the respective datasets.

The present study aims to develop a flexible and interpretable approach for combining auditory profiles (APs) of Saak et al. (2022) that complies with data protection standards and can be used in clinical practice. That means, we aim to investigate how APs can be merged from different datasets in a federated way to allow for dataset integration via APs and take the next steps towards a population-based estimate of APs. To achieve this, we (1) aim at comparing AP generated across datasets with respect to their profile similarity. We hypothesize that some APs will be similar across datasets, while the inclusion of further datasets will also result in new AP patterns, as a broader range of audiological patients is covered. Further, we (2) aim to analyze the feature importance of the distinct APs, such that APs can be easily identified by their specific patterns across audiological measures. Finally, we (3) aim to make the APs applicable by providing classification models for various settings, including research, hearing aid fitting, and potential smartphone applications. The current study, thus, aims to answer the following research questions:
RQ1: How can we obtain a combined set of auditory profiles from two audiological datasets?

RQ1.1: How can auditory profiles, generated on different datasets, be merged, such that a privacy-preserving combination of profiles from different datasets can be achieved?

RQ1.2: Are the previously generated 13 auditory profiles also represented in the second audiological dataset?

RQ1.3: How can we ensure flexibility of the proposed merging approach towards different use cases and how does flexibility relate to feature importance in different merging steps?

RQ2: How well can we classify patients into the generated profiles and which features are most important for this?

RQ3: Are the same audiological features important for merging and for classifying patients into the auditory profiles?

3.2 Method

3.2.1 Datasets

We used two separate datasets (dataset A and B) in our analysis to generate two separate AP sets (profile set A and B). Both datasets were provided by the Hörzentrum Oldenburg gGmbH (Germany) and contain information with respect to a broad range of measures. They include measures contained in both datasets, i.e., common measures, as well as different additional measures. Common measures for both datasets are age, the Goettingen Sentence Test (GOESA, Kollmeier and Wesselkamp (1997)), the audiogram for air- and bone-conduction, and the Adaptive Categorical Loudness Scaling (ACALOS, Brand and Hohmann (2002)). Both datasets use narrowband noise for ACALOS for the left and right ear measured with headphones. For dataset A results for the frequencies 1.5 and 4 kHz are available and for dataset B for 1 and 4 kHz. For audiogram and ACALOS measures, we only used the worse ear. In the following, we will refer to these as measures, and refer to features both as specific sub-results of the measures, and as summarizing features.

To exemplify, a common feature between the two datasets of the measure GOESA is the speech recognition threshold (SRT) for the collocated S0N0 condition (specific sub-result). For the audiogram the common features we used are summarizing features, derived from specific sub-results. These include the pure tone average for air (AC PTA) and bone (BC PTA) conduction, an asymmetry score between left and right ear (ASYM), the air-bone-gap (ABG), the Bisgaard class to characterize the shape of the audiogram, and the pure tone average (0.5,1,2,4 kHz) of the uncomfortable level (UCL PTA). For ACALOS we used summarizing features that can characterize both the lower and upper part of the loudness curve (L15, L35, i.e., level corresponding to categorical loudness of 15 CU and 35 CU, respectively), and the difference between L15 and L35 (as a feature representing the dynamic range), for both available frequencies (i.e., 1_diff and 4_diff). More information on common features between the two datasets is given in Table 3.1. A description of ranges across common and additional features of the two datasets

can be found in the supplementary material.

3.2.1.1 Original dataset A The original dataset A refers to the dataset and the profile set used in Saak et al. (2022). The dataset was collected for research purposes (Gieseler et al., 2017) and consists of 595 listeners (*mean age* = 67.6, *SD* = 11.9, *female* = 44%) with normal to impaired hearing. Next to the common measures and features, additional information is contained. In the speech test domain, the SRT for the digit triplet test (Zokoll et al., 2012) and the slope for the GOESA S0N0 condition is available. Further, information regarding the socio-economic status and two cognitive measures are included, namely the Demtect (Kalbe et al., 2004) and the Vocabulary test (Schmidt and Metzler, 1992). All these measures were used for the generation of the profiles. Detailed information about this dataset can be found in Gieseler et al. (2017) and Saak et al. (2022).

3.2.1.2 New dataset B The new dataset B was collected for diagnostic purposes and consists of 1401 listeners. The main measures overlap with dataset A. However, no cognitive measures are available for dataset B. Further, additional GOESA conditions are available, namely the S0N90 binaural and monaural conditions. Due to the diagnostic purpose of the dataset, information on the tympanogram (Type A, As, Ad, B, C, D, tympanic membrane perforation, not measurable), Valsalva and otoscopy (not impaired, moderately impaired, impaired) is also available for both ears. All these measures were used for the generation of the profiles. Only patients with data for at least two measures (from the audiogram, speech test, and loudness scaling) were included to ensure sufficient information for the clustering. This resulted in 1272 patients (*mean age* = 63.74, *SD* = 13.22, *female* = 42.26%) with normal to impaired hearing. In addition, different outcome parameters are available, such as International Statistical Classification of Diseases and related Health Problems (ICD) codes, and information on a potential hearing aid supply and the respective aided performance. Outcome parameters were not used for the generation of the profiles.

3.2.2 Generation of Profiles

For dataset A, APs were already available and obtained from Saak et al. (2022). They contain 13 distinct APs, and will be referred to as profile set A. Each profile consists of distributions across different features from audiological measures (audiogram, ACALOS, GOESA, ...). These ranges provide an estimate of plausible values for individuals that belong to a certain profile. To generate the profiles,

both the common and additional features from dataset A were used.

For dataset B, APs (profile set B) were generated according to the profile generation pipeline from Saak et al. (2022). Features used for profile generation include the common and additional features from dataset B. No outcome measures, such as the ICD codes were included to cluster patients into the APs. The available categorical features from the tympanogram, otoscopy, and Valsalva were transformed to continuous features using multiple correspondence analysis (MCA, Lê et al. (2008). MCA is a dimension reduction method, similar to principal component analyses. It quantifies the relationship between categorical variables in the form of principal components, and in in that way transforms the categorical features into continuous components. We used the first three resulting components instead of the original three categorical variables as features for the subsequent data analyses. With this approach we retained some information regarding the three categorical variables for the generation of profiles. The resulting components, however, only represent the relationship across the categorical variables and some information will be missing. Hence, in the future a different approach to tackle the difficulty of handling mixed data may become preferable.

Model-based clustering was used to generate the profiles. Model-based clustering assumes that an underlying model generated the data, and the clustering aims at recovering the model (Banerjee and Shan, 2017). The model is a combination of data clusters that describe the patterns in the data. These clusters serve as the generated APs. The profile generation pipeline (Saak et al., 2022) using model-based clustering consists of two steps:

The robust learning step results in a robust estimate of the underlying model parameters for model-based clustering, namely the number of profiles present in the dataset and the respective covariance parameterization. To achieve this, bootstrapping without replacement was used to generate 1000 datasets, each using 90% of the data. Next, missing data in each bootstrapped dataset was imputed. Here, we made a small adjustment to the original pipeline. We replaced Multivariate Imputation by Chained Equations (MICE, Van Buuren and Groothuis-Oudshoorn (2011)) by imputation based on factorial analysis for mixed data (FAMD, Audigier et al. (2016)). FAMD is a competitive imputation technique based on principal components that reduces the computational complexity of the profile generation pipeline. This means, that instead of multiple imputed datasets only a single imputed dataset is generated for each bootstrapped dataset,

which simplifies the following computations for cluster analyses. Internal simulation analyses show the equivalence of FAMD to MICE for the current dataset. Next, features were scaled using the min-max scaling, which transforms values to range from 0 to 1. We then applied model-based clustering to each dataset using different parameter combinations (number of profiles, covariance parameterization). The possible number of profiles was set to 1 to 40 profiles to cover a broad potential range of profiles. The covariance parameterizations determine the shape, volume and orientation of the profiles (see Fraley and Raftery (2003) for all possible covariance parameterizations). The model describing the underlying data best was then selected using the Bayesian information criterion (Schwarz, 1978). Finally, the most frequently occurring number of profiles and covariance matrix across all bootstrapped datasets was selected as the model parameter solution fitting the data best.

The profile generation step uses the original complete dataset and, again, imputes missing data with FAMD. Model-based clustering is then applied to the scaled features with the learned model parameter solution to result in the APs of profile set B. To allow for later merging with anonymized data distributions from sensitive datasets, all profile data was transformed to count data with 100 equidistant steps. More details regarding the generation of profiles can be found in Saak et al. (2022).

3.2.3 Dataset combination via Auditory Profiles

To compare and combine profiles generated across multiple datasets, they need to share common features that allow to investigate how similar different profiles are. We, therefore, selected the common features of the two datasets across the domains of speech intelligibility, audiogram, loudness scaling, and anamnesis information, namely the age of the individual. The columns of Table 3.1 display all features used to combine the two datasets.

3.2.3.1 Overlapping density index A measure of similarity is required that allows for an estimation of similarity between two profiles. For this purpose, we use the overlapping index Pastore and Calcagnì (2019). The overlapping index is a distribution-free index that calculates the overlapping area of two probability density functions. To obtain the similarity between two respective profiles we calculate the overlapping index for each feature of the two profiles and use the mean as the final similarity score for the respective profiles. Sample size imbalances are accounted for by using normalized density distributions.

3.2.3.2 Merging procedure Starting from all profiles generated on datasets A and B, profiles were merged iteratively, and the procedure is depicted in Figure 3.1. The similarity score was calculated for all profile combinations, and the two profiles with the highest similarity score were merged. The SRT was used 6 times when calculating the similarity score to balance the effect of the speech test to the number of available features from the audiogram and ACALOS. Then, for the new set of profiles containing the merged profile, the similarity score was calculated for all profile combinations, and the two profiles with the highest similarity score were merged. This procedure continued until only two profiles remained. Hence, we pre-generated all potential profile solutions and then derived a stopping criterion to result in the final profile set, as described in section 3.2.3.3.

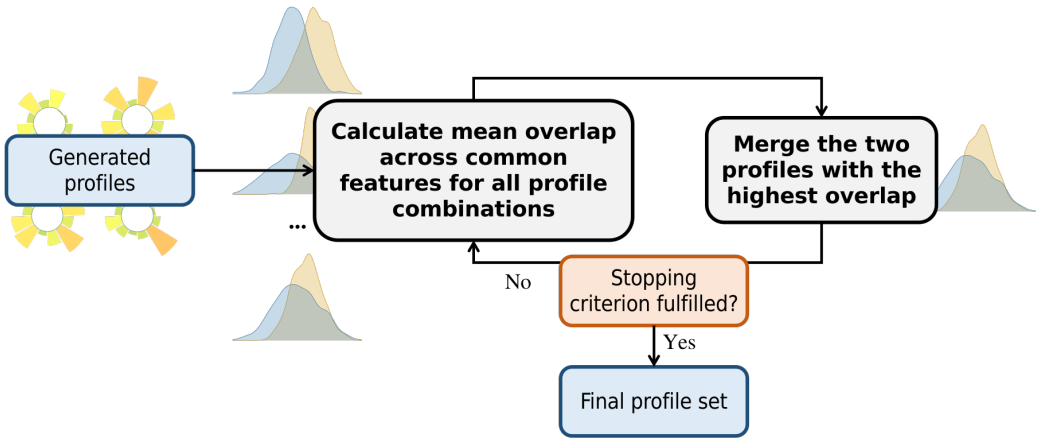

Figure 3.1: Merging procedure. Profiles are merged iteratively until a stopping criterion is fulfilled. The two profiles with the highest overlap are merged at each iteration step.

3.2.3.3 Selection of merged profile set After each merge a potential profile set is generated. That means the merging pipeline results in N -1 potential profile sets, ranging from all N available profiles from the two datasets to the last two remaining profiles. Depending on the intended use case in practice, a plausible profile set needs to be selected. For instance, the last two to three remaining profiles may be useful for a coarse classification into mild, moderate, and severe hearing loss. It does not, however, provide a detailed differentiation across features that may be needed for research purposes or detailed patient characterization. In contrast, a profile set containing a larger number of profiles may contain redundant information for clinical classification but may aid in investigating specific differences between patient groups. For that reason, we aimed at proposing a basic set of APs that remains detailed enough, whilst reducing

the number of profiles and combining similar profiles contained in both datasets. More specifically, we aimed at reducing redundancy across profiles for a proposed general set of profiles, such that the differences between the profiles are maximized. To achieve this, we used a combination of two steps to manually select a cutoff based on the overlapping density (see Figure 3.3A in the results section), corresponding to a number of obtained profiles at a certain number of merging iterations.

In the first step, we selected the cutoff based on two parameters, namely the slope of the maximum overlapping density, and the variance of the median overlapping density. The first parameter describes the slope over merging iterations for the two profiles with the highest overlap in each merging iteration. A steeper slope and lower overlap, thus, indicates when profiles are merged that are less similar than in previous steps. We selected the parameter such that the slope decrease is relatively larger after the cutoff, as compared to prior to the cutoff. The second parameter describes the variance over iterations for the median overlap of all profiles. If the variance changes too much over iterations, it indicates that merges took place between two profiles that differ strongly from each other. Hence, we selected the cutoff such that the variance remains relatively stable prior to the cutoff, contrasting the variance after the cutoff.

In the second step, we compared the overlapping index across features before and after the previous cutoff. For an optimal cutoff, we expect a higher overlapping index prior to the cutoff, and a lower overlapping index after the cutoff. In that way, we can determine whether features were overall similar for the two profiles to be merged, or whether they differ substantially from each other, in which case a merge may not be advisable. Hence, it also allows to investigate which features are most important for the merging procedure and the profiles, and likewise, which features distinguish two profiles the most. We aimed for a cutoff that shows high overlap of features prior to it, and substantial difference between profiles after the cutoff.

3.2.3.4 Feature importance of the merges For an interpretable merging procedure and interpretable APs, it is highly relevant to investigate which audiological measures drive the profile merges (high overlap), and which audiological measures hinder the profile merges (low overlap). That is because features that drive the profile merges are less relevant, whereas features that hinder the profile merges are more relevant and able to discriminate better between profiles. We,

therefore, investigated which features had the least overlap across profiles, that is, were most discriminative between profiles. For that purpose, we defined different sets of iterations, for instance, iterations before the defined cutoff and after the defined cutoff. Iterations before the defined cutoff indicate which features were merged and which information may have been lost. Iterations after the defined cutoff indicate which features are most important for discriminating between profiles. To exemplify, if we start at the last profile set with two remaining profiles, we can investigate which measures are least similar across the two profiles and would therefore lead to a split into three profiles. Hence, these features would drive the split and can be considered as relevant features.

We further investigated which features were most relevant across different ranges of iterations (merging areas, see A-E in Figure 3.3). A merging area, for instance, can contain all merges from the selected profile set to the last two remaining profiles, or all merges that led up to the selected profile set. For this, we calculated the mean overlap for each feature within a merging area and compared this to the average mean overlap for all features for the respective merging area. Hence, we investigated whether a feature was more or less important than the average of all features within the merging area.

3.2.4 Classification models

We built classification models for the APs for two reasons. First, we aimed at enabling the classification of new patients into the profiles, such that the profiles can be used in practice. Second, the feature importance of the classification models allows to draw conclusions with respect to which features are most relevant for classification into the respective profiles.

3.2.4.1 Feature sets and labels To simulate different use cases, enable applicability of the profiles in practice, and extract insights into feature importance, different feature sets (subsets of the common features used for merging) were used to build the classification models. The feature sets belong to three general categories, namely, use cases, combined, and single. The feature combinations “ALL”, “APP” (smartphone app), and “HA” (hearing aid fitting) belong to the category use case, as they are combinations that have use cases in practice. For instance, “HA” defines features generally available for a hearing care professional, whereas “APP” defines features that could potentially be measured via a smartphone. “ALL”, in contrast, allows for an overall feature importance interpretation across all features. The combined feature group explores the performance of only using

Table 3.1: The different feature sets used in the analysis (see 3.2.1 for feature descriptions). "ALL" corresponds to all features common to both datasets that were also used for the profile generation; "APP" to measures potentially measurable via a smartphone; "HA" to measures generally available for hearing care professionals. The remaining feature groups describe feature combinations and the importance of single features.

		Use cases			Combined			Single		
		ALL	APP	HA	AG SRT	AG ACALOS	SRT ACALOS	AG	SRT	ACALOS
Speech test	S0N0 bin	✓	✓	✓	✓		✓		✓	
Audiogram	AC PTA	✓	✓	✓	✓	✓		✓		
	ASYM	✓	✓	✓	✓	✓		✓		
	BISGAARD	✓	✓	✓	✓	✓		✓		
	BC PTA	✓		✓						
	ABG	✓		✓						
	UCL PTA	✓		✓						
Anamnesis	Age	✓	✓	✓	✓	✓	✓	✓	✓	✓
ACALOS	1_L15	✓	✓			✓	✓			✓
	1_L35	✓	✓			✓	✓			✓
	1_diff	✓	✓			✓	✓			✓
	4_L15	✓	✓			✓	✓			✓
	4_L35	✓	✓			✓	✓			✓
	4_diff	✓	✓			✓	✓			✓

two of the three main features. In that way the classification models could be used also for datasets only containing two of the three measures. The final feature group (single) investigates performance with each measure separately. All sets are displayed in Table 3.1. While these feature groups allow for broader applicability of the classification models, they also allow for feature importance interpretations, as the performance across feature sets can be compared. Profile numbers of the final combined profile set were used as labels for classification.

3.2.4.2 Classification model In Saak et al. (2022), a random forest (RF, Breiman (2001)) classification model was built to classify new patients into the profiles, which resulted in adequate performance. RF models work well with smaller sample sizes and provide an inherent feature importance index. We, therefore, also selected RF for the current analyses. RF uses an ensemble of trees, where the results of independent trees are aggregated, and the most frequently predicted label is used as the final prediction. Within each tree, only a subset of the features is used to split the predictor space into smaller regions,

which effectively decorrelates the trees.

To build the RF model we used the one-vs-all (OVA) design, as it provided good performance in Saak et al. (2022), and eases the interpretation with respect to feature importance. The OVA design splits a multiclass model into k binary models, where k is equal to the number of profiles. That means each model predicts whether an instance belongs to a specific profile or not. The feature importance therefore always highlights features that are most important for distinguishing the profile of interest from all remaining profiles. Data imbalance naturally exists with OVA design. To counter the imbalance, we upsampled labels using Gaussian noise to the average amount of labels available for each profile, which means all profiles have at least the average amount of patients in each profile, while profiles with patient numbers above the average retained their larger sample size. The rationale behind this was that we wanted to keep a balance between upsampled and original data. The remaining imbalance was addressed by using weights for the labels in the training. In that way, mistakes for the label of interest are more costly in terms of prediction errors.

3.2.4.3 Train, validation, and test set The complete dataset was split into a training (80%) and a test set (20%). The training set (containing 80% of the data) was then further split into a training (80%) and validation set (20%). The training set was used for training. For the training set we used 10 times repeated 10-fold cross-validation to get a better estimate of the prediction error. The validation set was then used to evaluate the performances of the models on cases that were not used in the model training. After the model was specified with the training and validation set, the final model performance was evaluated with the test set.

3.2.4.4 Classification performance evaluation Each classification model aims at reducing the prediction error, which is quantified by a specified evaluation metric. Evaluation metrics have different properties, making them useful for different prediction problems. For instance, accuracy is an evaluation metric that is easily interpretable, but does not perform well for imbalanced classification performances. We chose Cohen's kappa (Cohen, 1960) as the evaluation metric for two reasons. First, Cohen's kappa takes imbalances into account, by comparing the accuracy to the baseline accuracy that could be achieved by chance. Second, it proved to be the best evaluation metric among three others (balanced accuracy, Area under the precision recall curve, F1-score) for the classification model for

profile set A in Saak et al. (2022). Hence, in the model training process, we used Cohen's Kappa to tune the number of considered features at each split.

To evaluate the general performance of the trained classification models for each feature set, we used two distinct but complementary metrics, namely sensitivity and precision. Sensitivity describes the proportion of correctly classified cases for the class of interest, whereas precision describes the proportion of misclassifications for the class of interest. Both sensitivity and precision were compared across feature sets and further, for the overall profile classification and single profiles.

Finally, we build a dummy prediction model, which does not include any features and predicts profile labels based on stratified sampling, that is, it reflects the relative frequency of patients contained in different profiles. That way, we could estimate the benefit of our prediction models as compared to the baseline dummy model.

3.2.4.5 Feature importance of the classification model To estimate the feature importance of the RF models we used the inherent feature importance metric, namely the gini importance or rather the mean gini decrease (Breiman, 2001). The mean gini decrease is used in the training process to estimate how well a feature can split the labels across nodes in the ensemble of trees. A good split results in pure nodes where no misclassification occurs, whereas a bad split does not aid in separating the labels for the classification and leads to "impure" nodes. The mean gini decrease estimates how much a feature decreases the impurity on average across all nodes in the ensemble of trees, where a feature can be used multiple times in the same tree.
Overall feature importance across all profiles and feature importance specific for each profile was calculated and compared via a feature importance plot. We transformed the mean gini decrease to percentages (with the total mean gini decrease of a profile equaling 100 %) to show the contribution of each feature for each profile separately.

3.3 Results

3.3.1 Generation of Profiles

For dataset A, the optimal profile number of 13 was obtained from Saak et al. (2022). For dataset B, the optimal profile number was determined according to the profile generation pipeline described in the present paper. The distribution

of estimated optimal profile numbers across the bootstrapped datasets can be found in Figure 3.2. Both 31 and 32 were most frequently selected across the bootstrapped datasets and there is only a marginal difference between these two profile sets in terms of frequencies. In contrast, estimated profile numbers higher or lower than 31/32 were selected less frequently for the bootstrapped datasets. Since profile number 31 was slightly more frequently estimated than profile number 32, we selected 31 profiles as optimal for dataset B. Just as in Saak et al. (2022), the best covariance parameterization for dataset B was "VEI". "VEI" refers to variable volume, equal shape, and coordinate axis orientation. It therefore allows clusters to be of different size but restricts them in terms of shape and axis alignment.

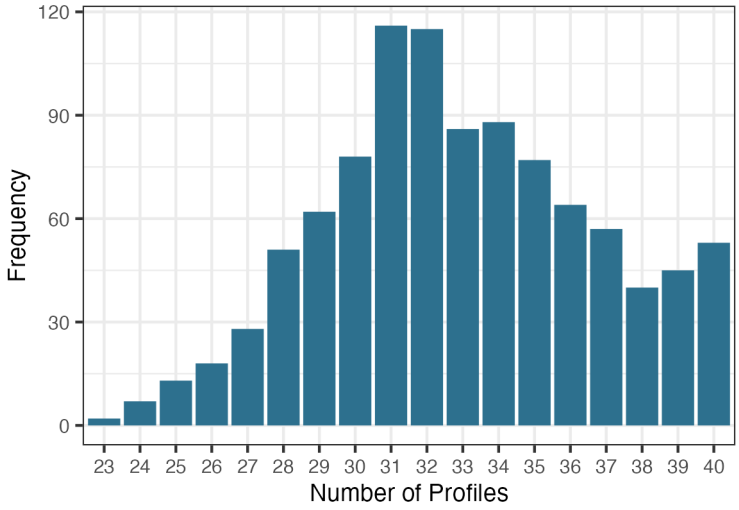

Figure 3.2: Distribution of optimal profile numbers across the bootstrapped datasets.

3.3.2 Merging Profiles

The profiles of the two datasets were merged using the mean overlapping density. This requires the definition of criteria to stop the merging process and result in the final profile set. Figure 3.3 displays the two criteria that we used to select the proposed combined profile set.

Figure 3.3A displays the highest overlap (i.e., profiles to be merged) next to the median overlap across all profiles for each iteration of the merging pipeline. Two cutoffs are depicted that show two potential profile sets. The first cutoff at 20 iterations is characterized by a steepening decrease of the highest overlap, next to a reduction in variance. The second cutoff at 32 iterations precedes an even steeper slope decrease and high variations in variance. The high variations in variance in-

dicate that merges occurred where profile ranges became much broader such that the similarity between profiles could increase again. Since this is undesirable, a cutoff should occur prior to high variations in the variance.

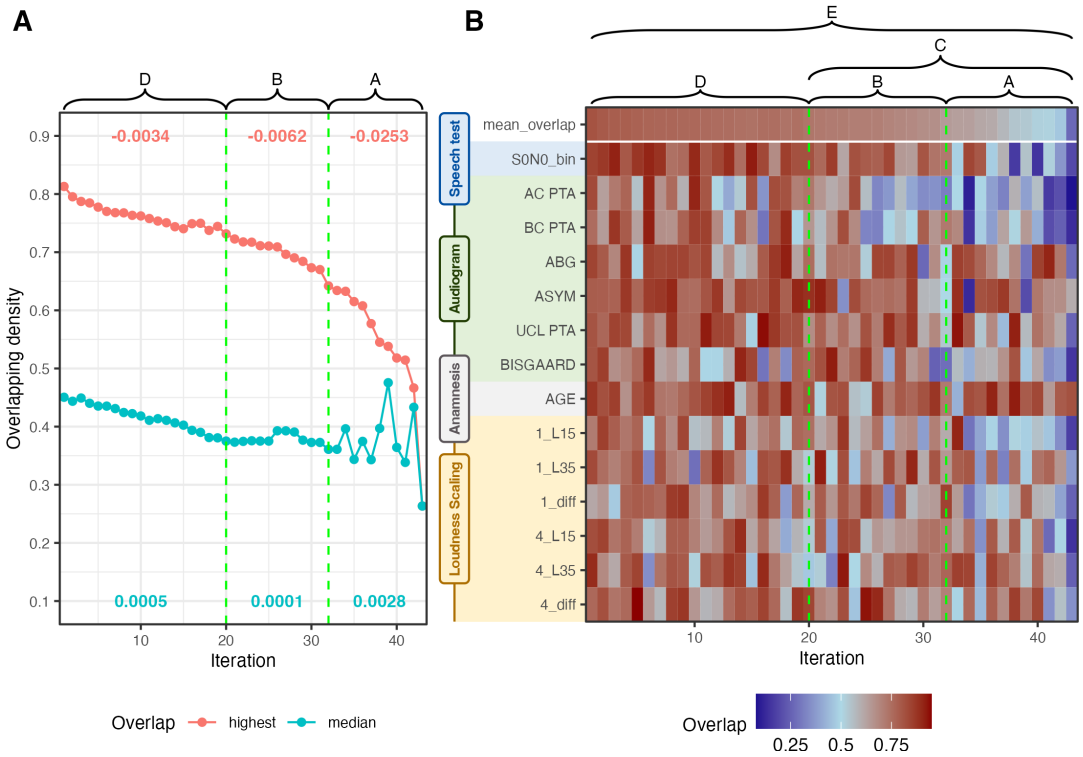

Figure 3.3: Merging iterations. Vertical green lines indicate cutoff candidates (A) Median and highest overlapping density across the merge iterations. Red numbers indicate the slope of highest overlapping density for the respective merging area; turquois numbers indicate the variance of median overlapping density of the respective merging area. (B) Overlapping densities across features for the two profiles to be merged at the given iteration. A, B, C, and D indicate different iteration sets.

Figure 3.3B displays the overlapping densities of the two profiles to be merged at each merging iteration. In that way it corresponds to the highest overlap from Figure 3.3A. The two cutoffs are indicated with the dashed green lines. We can observe a general decrease in overlap with increasing merging iteration, that corresponds to the results from Figure 3.3A but provide us with insights into which features drove the decrease in overlap. We can observe that especially after the cutoff at 32 iterations the mean overlap decreases substantially across multiple features, which indicates that profiles would be merged that can be distinguished. Hence, merging the profiles beyond 32 iterations would result in a substantial loss of information. This widespread decrease in overlap is not as pronounced for the cutoff at iteration 20. Hence, for the following analyses the profile set at 32 iterations was selected, which leads to a proposed combined profile set with thirteen APs. The corresponding analyses for the profile set at

20 iterations can be found in the supplementary material.

3.3.3 Feature importance of the merges

To ensure that merges are based on plausible features we investigated the feature importance of the merges. In other words, we estimated which features were most responsible for merging profiles. Features important for merging indicate they are less relevant for profile distinction, as they are similar across profiles. Conversely, features that are less important for merging can be interpreted as features relevant for profile separability. Hence, these features are determined as important features. Figure 3.4 displays the feature importance of the merging pipeline for different merging areas (A - E). A corresponds to the selected profile set at iteration 32; C to the profile set at iteration 20 (in more detail in the appendix); E visualizes profile importance across all iterations. B and D display feature importance for remainders of the iterations.

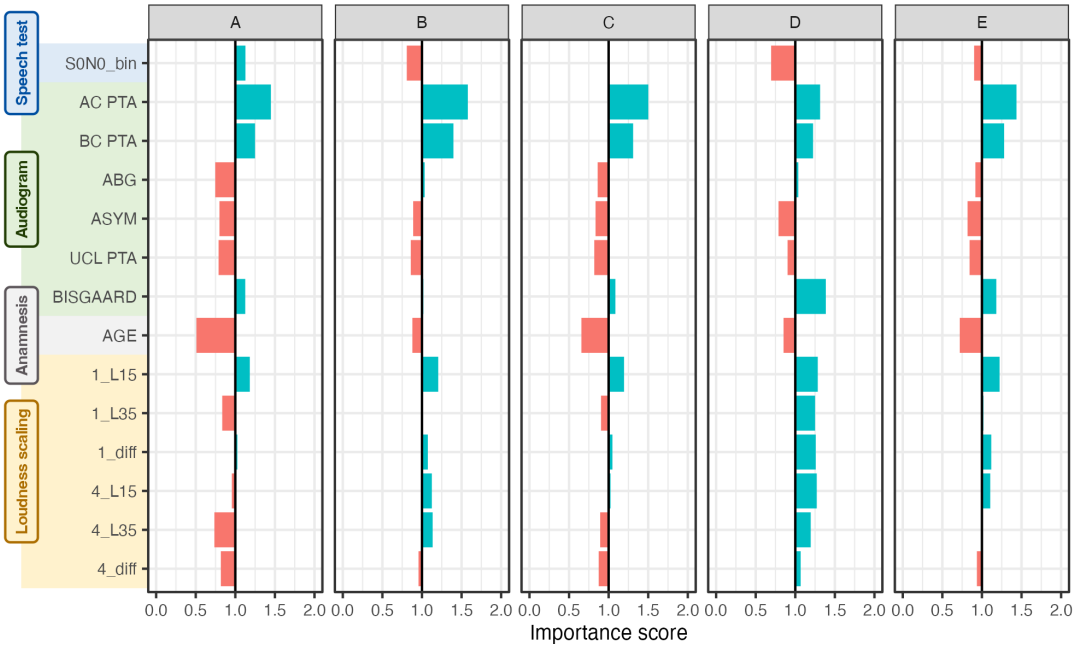

Figure 3.4: Feature importance for different merging areas. A, B, C, D and E correspond to the indicated areas in Figure 3.3. Mean feature overlap was calculated for each feature in each merging area and then subtracted from 1 to indicate higher importance with higher scores. Subsequently, scores are divided by the mean overlap of all features of the merging set. Scores, thus, indicate the higher or lower importance than the average. Turquoise bars with values above 1 indicate higher importance; red bard with values below 1 indicate lower importance.

Feature importance, or profile separability for the 13 profiles of profile set 32 (merging area A) is mostly based on air- and bone-conduction PTA, the Bisgaard class, the SRT, and L15 and the difference of L35-L15 of the ACALOS for 1

kHz. For the complete set (E) the SRT becomes less important, and the L35-L15 difference becomes more relevant, next to L15 for 4 kHz. This occurs due to the higher importance of ACALOS features in merging area D. This means, that in the first profile merges in D (or later profile splits – if going from right to left iterations), loudness scaling based features are merged first (ACALOS information is lost), whereas in A, profile merging is driven more by audiogram-based features, the SRT, and L15 from the ACALOS. Vice versa, this also implies that if we would start from only 2 profiles and would split until iteration 1, in later split iterations (profile sets with a higher number of profiles) more detail with respect to loudness scaling is added to the profiles.

3.3.4 Proposed profile set

Figure 3.5A visualizes the proposed profile set with 13 APs. Both the profile ranges for each feature (A) and a proposal for single profile visualization are depicted (B). We can observe distinct patterns across APs. Profile 13 corresponds to a normal hearing profile and profile 1 has the highest SRTs. Overall, the profiles cover a large range of hearing deficits in terms of test measurement ranges. We can see a distinctiveness of the profiles based on the SRT, audiogram-related, and loudness scaling-related features. The age was not found important in the two datasets.

In the SRT range of -5 to 0 dB SNR differences between profiles are mainly driven by audiogram and loudness scaling based features. For instance, profiles 10 and 11 show similar SRTs, but differ regarding loudness perception and the AC PTA. Further, for profile 10 an asymmetry is present which could partly explain the higher AC PTAs. As expected, the presence of a higher asymmetry and a higher air-bone gap compensate for higher AC PTAs in terms of higher SRTs.

We can see a clear inverse trend of the dynamic ranges for the ACALOS 1 and 4 kHz to the SRT. That means, generally a higher SRT is accompanied by a reduced dynamic range. This trend fluctuates, when an ABG and asymmetry is also present in the profile.

The single profile visualization (Figure 3.5B) aids in visualizing the pattern for a single profile. Not all profiles are displayed, but the remainder can be found in the supplementary material. The polar plots depict the normalized median difference of each profile to the normal hearing profile (green circle). We can clearly see the impact that the presence of an ABG or asymmetry has on the relation between

SRT and AC PTA (Profile 1 vs. Profile 3), that is, a smaller SRT results from a higher AC PTA in profile 3 due to asymmetry and/or ABG that both mitigate the general SRT deterioration with increasing AC PTA.

Figure 3.6 depicts the distribution of profiles from the two profile sets (A & B) in the final 13 profiles. The 13 profiles from profile set A have been merged with the 31 profiles from profile set B. Hence, the previous profiles from Saak et al. (2022) are also included within the new profile set. However, since, dataset B contains patients with a larger variety and more severe deficits, the previous profiles are merged in favor of retaining a broader profile distribution with the new profile set. The number of patients per profile corresponds to the relative frequency of patients contained in the datasets. This is because model-based clustering, used for profile generation, does not impose any constraints on cluster size with the variable volume parameterization ("VEI").

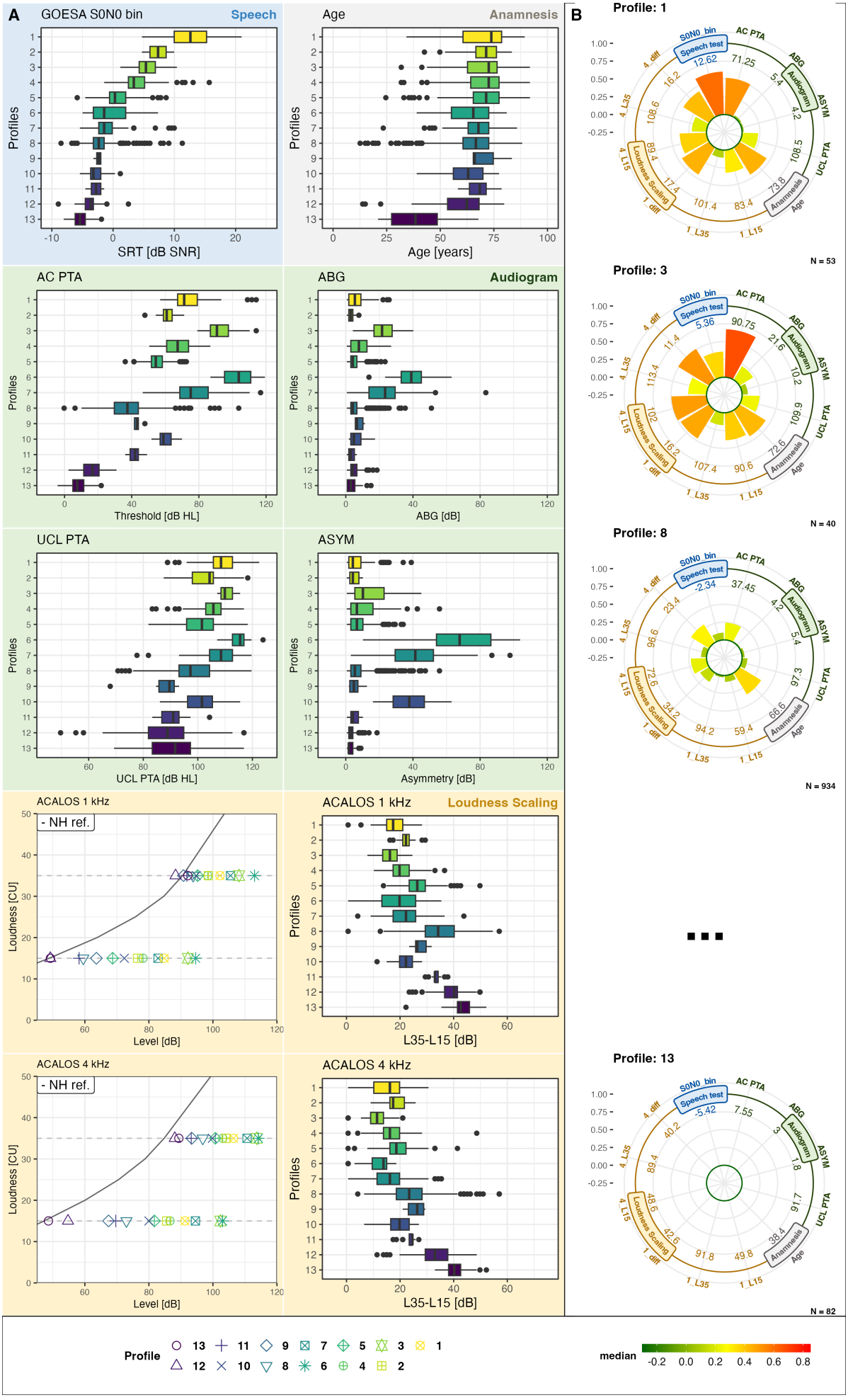
A
GOESA S0N0 bin
Speech
SRT [dB SNR]
Age
Anamnesis
Age [years]
AC PTA
Threshold [dB HL]
ABG
Audiogram
ABG [dB]
UCL PTA
UCL PTA [dB HL]
ASYM
Asymmetry [dB]
ACALOS 1 kHz
- NH ref.
Loudness [CU]
Level [dB]
ACALOS 1 kHz
Loudness Scaling
L35-L15 [dB]
ACALOS 4 kHz
- NH ref.
Loudness [CU]
Level [dB]
ACALOS 4 kHz
L35-L15 [dB]
Profiles
B
Profile: 1
N = 53
Profile: 3
N = 40
Profile: 8
N = 934
Profile: 13
N = 82
S0N0_bin
Speech test
AC PTA
ABG
Audiogram
ASYM
UCL PTA
Anamnesis
Age
1_L15
1_L35
4_L15
4_L35
Loudness Scaling
1_diff
4_diff
Profile
median

Figure 3.5: The 13 proposed APs across the speech, anamnesis, audiogram, and loudness scaling domain. (A) Profile ranges are depicted for all features and are ordered with respect to the increasing median SRT. (B) View of singular APs. Data is referenced to the NH profile (value 0). All bar values represent the median deviation from the NH profile, whereas the numbers indicate the true median value.

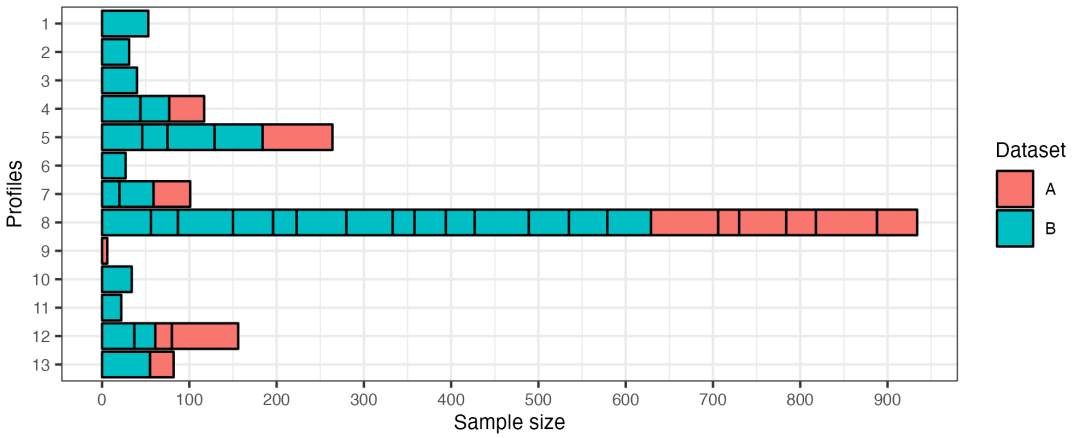

Figure 3.6: Distribution of profiles that were merged to result in the selected 13 auditory profiles. A corresponds to dataset A and presents the previous 13 profiles detailed in (1). B refers to the new dataset B and the 31 profiles that were defined as optimal by the profile generation pipeline. The x-axis depicts the sample size for each sub-profile, as well as the proposed 13 profiles.

3.3.5 Classification models

Classification performance across profiles for different feature sets are shown in Figure 3.7. The general classification performance is adequate for the "APP" and "ALL" feature sets (Figure 3.7A), with "APP" performing best among all feature sets. All feature sets performed better on average than the dummy model but for different profiles different benefits are achieved (Figure 3.7B). The higher performance of the dummy model for AP 8 can be explained by the larger sample size of this profile, which increases the chances of belonging to this specific AP. The "single" feature groups performed worst among the feature groups. This can, however, be expected, as less information is available to discriminate between APs with "single" sets. For the "single" feature group "AG" (audiogram), and for the "combined" feature group "AG ACALOS" achieved the best performance. The feature groups "SRT ACALOS" and "AG SRT" performed comparable on average and differences can be observed when comparing performances across profiles. For profiles with fewer deficits (higher profile number) "SRT ACALOS" could generally discriminate better than "AG SRT", and vice versa for profiles with higher deficits (lower profile number). The feature sets "APP" and "ALL"

both perform better than "HA" from the usecase feature group, which does not contain ACALOS information. The results demonstrate the importance of using audiological tests beyond the audiogram to adequately classify patients into APs. All three measures contribute to better discriminability into the distinct APs and the benefit of including ACALOS information for better discriminability is shown.

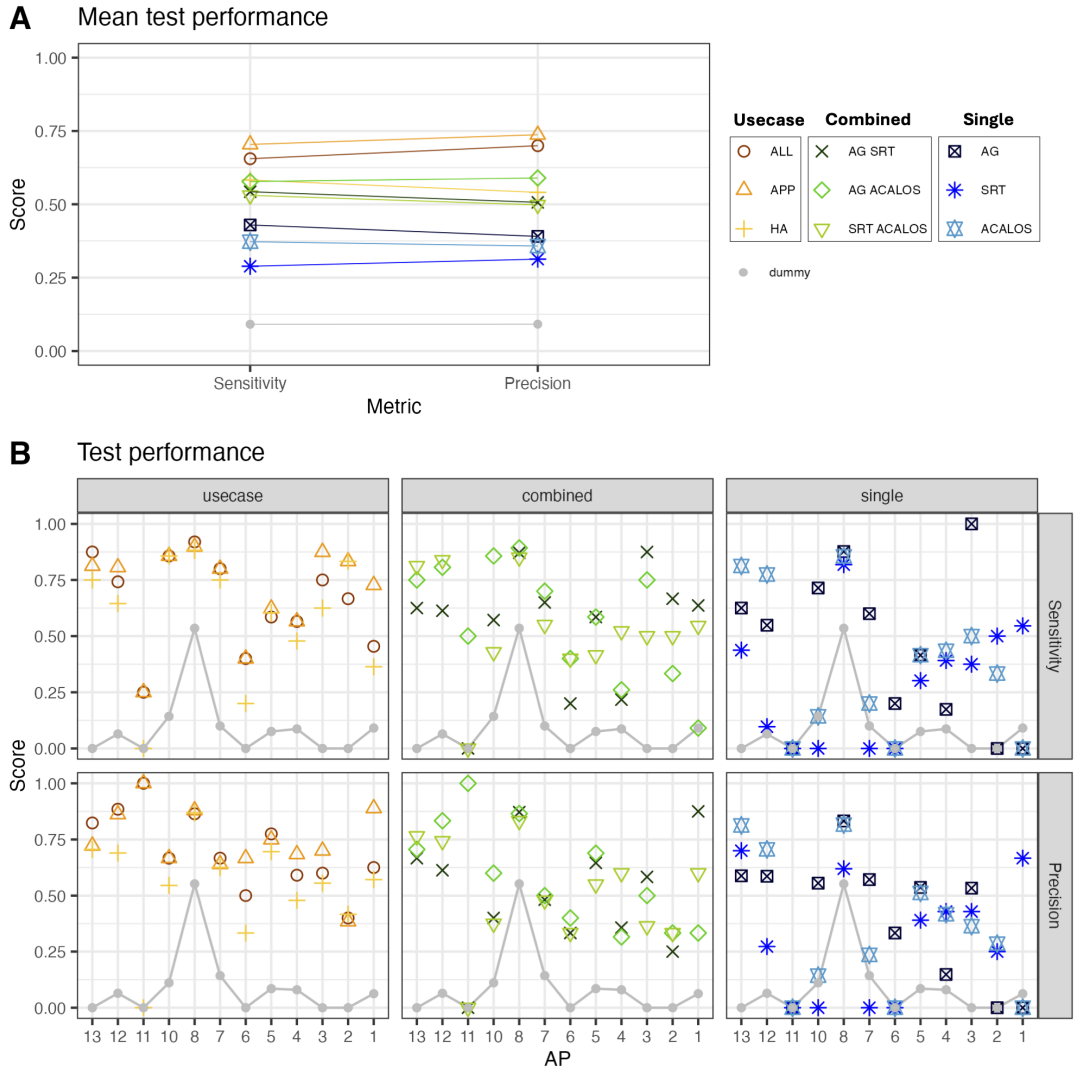

Figure 3.7: Classification performance across profiles for different feature sets. Feature groups are categorized into "usecase", "combined", "single". Profile 9 is not displayed, as the sample size is too small for adequate training. (A) shows the mean test performance across profiles and (B) show the test performance for each profile. Dummy indicates the performance of the dummy model that predicts profile labels based on stratified sampling.

To investigate the feature importance of the classification models, we selected the best-performing feature set, namely the "APP" feature set of the "use case" feature group (Figure 3.8). Hence, a reduced feature set performed best in

classifying patients into the profiles. The AC PTA was determined to be overall most important for classifying profiles. More specifically, this means that it is most important for distinguishing between profiles in the presence of the remaining features. This is seconded by the SRT. We can, generally, observe that for profiles with higher SRTs (lower profile number) the SRT becomes more important. For instance, for distinguishing profile 1 and 2 from the remaining profiles, the SRT is more important than the AC PTA. We can further observe the most important features by comparing them to the mean importance of all features (Figure 3.8, dashed line in upper panel). The most important features are the SRT, AC PTA, asymmetry, and L15 for 1 kHz, and the importance of these features varies for different profiles. Profiles with lower SRTs (higher profile numbers) show a slight trend for higher importance of ACALOS features for differentiating the profiles from all remaining profiles.

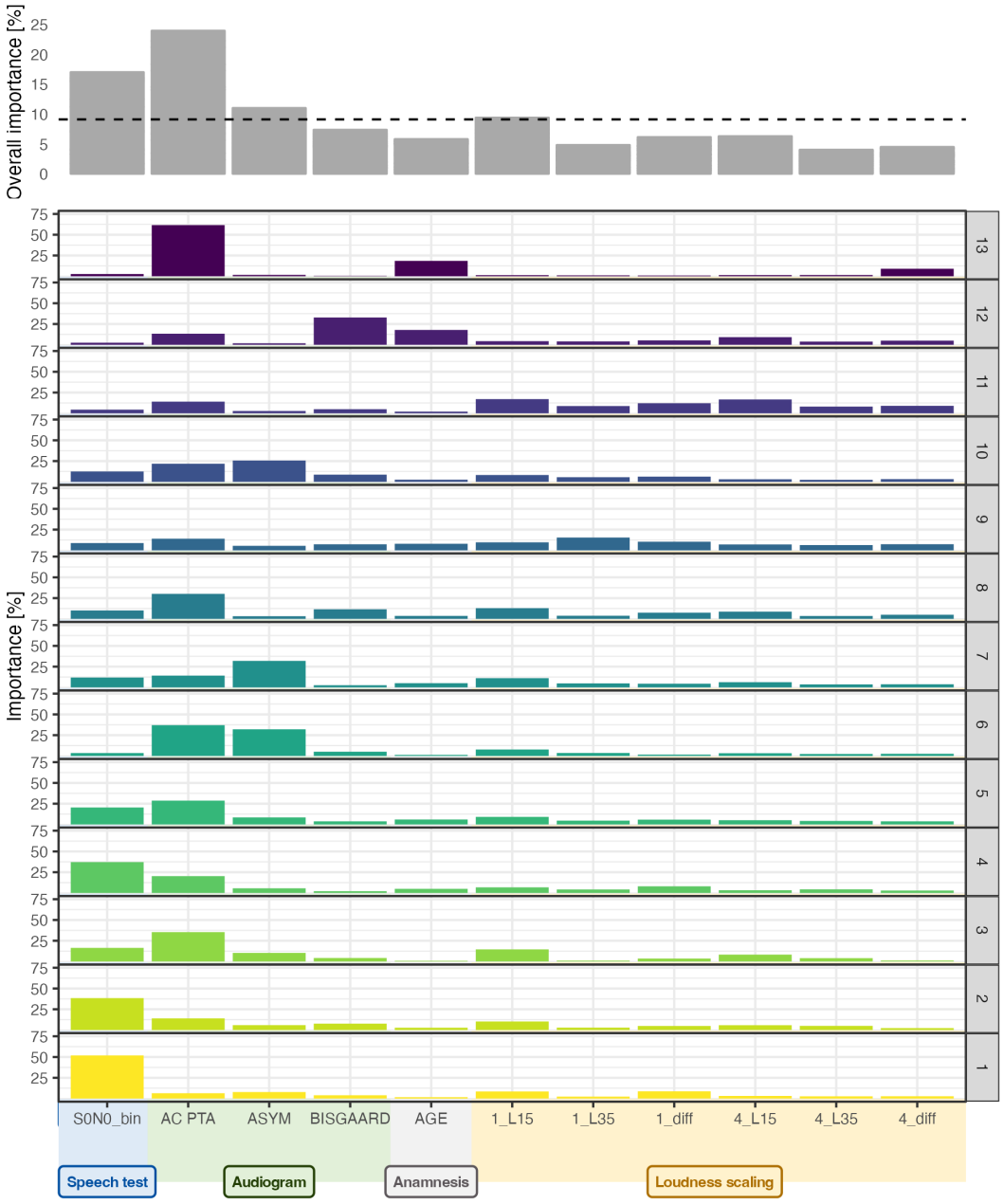

Figure 3.8: Feature importance for "APP" feature set. Feature importance refers to percentage distribution of the mean gini decrease for each profile separately. Profile specific importance is shown by the colors. Overall importance (transformed to percentage) across all profiles is depicted by the grey bars. The dashed line indicates the mean percentage across features

3.4 Discussion

With the present study we aimed at extending the existing profiling approach such that datasets can be integrated into the auditory profiles (APs) towards a population-based estimate of APs with a federated learning approach. Our results show that APs generated across datasets can be plausibly merged using

the mean density overlap of two profile distributions. The exact number of profiles is flexible and can be adjusted regarding the required detail of the profiles. We further trained classification models that allow for adequate classification of new patients into the APs using different feature combinations. Finally, we determined the importance of different features for both merging of and classification into the APs.

3.4.1 Auditory profile generation for dataset B

The optimal number of profiles generated for dataset B is 31. In comparison to the 13 APs from dataset A, this number can be considered rather high. However, one must consider the differences between the two datasets. First, dataset A is a research dataset where participants were recruited via a participant database of the Hörzentrum. In contrast, dataset B is a clinical dataset where participants approached the Hörzentrum themselves for the diagnostic support of an ENT-physician. As a result, dataset B contains a larger sample covering a broader range of hearing loss including patient patterns with more severe hearing loss, as evident from the higher SRTs in profiles 1-3, which were not merged with profiles from dataset A. Second, the two datasets vary in terms of additionally included audiological measures. For dataset A, the most prominent additional features to the common features are cognitive measures. Dataset B, in contrast, contains additional features such as the S0N90 condition, both monaurally and binaurally, and information from the tympanogram, valsalva, and otoscopy. As a result, these additional features, in combination with the larger variation in patient patterns can explain why dataset B resulted in a higher optimal profile number (see Section "Role of common and additional features for profile generation, merging, and classification").

3.4.2 Merging procedure and its flexibility

Our proposed iterative merging procedure, using the highest mean density overlap between two APs, respectively, enables the integration of different datasets via APs **(RQ1.1)**. The separately generated profiles from the two datasets can, thus, be combined to describe both datasets together. The newly proposed and merged/combined profile set covers a wide range of hearing deficits and extends the range of deficits to the APs of profile set A, which are described in detail in Saak et al. (2022). Both sensorineural and conductive/mixed hearing losses (APs 3, 6, and 7) are covered within the profiles. Integrating the two datasets (A and B) has extended the range of hearing deficits contained in the profiles

as compared to the profile set A. This is evident from the new profiles with higher SRTs (Figure 3.6, APs 1-3) covering patients with more severe hearing deficits. The APs now compactly describe varying hearing deficits patterns. AP 1 and 3, for instance, vary with respect to the impairments across included measures. AP 1 has a worse SRT score, but scores closer to the normal hearing reference for ACALOS and Audiogram features compared to AP 3. This can be explained by the larger ABG present in AP 3. The general prevalence of an asymmetry and an ABG in the APs highlights the importance of including these patient types in further research. For AP 5 we can observe a typical worsening of hearing deficits with higher frequencies. More specifically, we observe a more pronounced reduction in the dynamic range at higher frequencies (1_kHz diff vs. 4_kHz diff) compared to lower frequencies, which can aid in better speech intelligibility, as compared to a uniformly reduced dynamic range observed in AP 1.

The similarity score develops plausibly in a continuous way across merging iterations and features can be identified that drive the AP merges. In that way the merging procedure allows to plausibly combine, compare, and characterize the content of two datasets containing common features.

The newly proposed profile set contains information from the two profile sets (A & B). The previously generated 13 APs (profile set A) are also represented in the dataset B. They are both merged with profiles from dataset B, and with profiles from dataset A. That means, the profile set A is now represented by fewer and coarser profiles, while additional profiles with more severe hearing deficits were added to the new proposed profile set **(RQ1.2)**. This behavior can be explained by the continuous merges leading to a coarser profile separation, more severe cases being present in the dataset B, and that profiles from profile set A were more similar to themselves than profiles from profile set B.

The profiles, further, are flexible regarding their precise number of profiles following the merging. Depending on the use-case of the profiles, a more detailed or a coarser separation between profiles may be needed. Generally, selecting a cutoff in the later merging steps will lead to fewer profiles with broader ranges and a coarser separation. Conversely, selecting a cutoff in the earlier merging stages will lead to a higher number of profiles and allows to investigate smaller differences between profiles **(RQ1.3)**. For instance, for screening purposes, a small number of profiles may suffice in broadly estimating mild, moderate, or severe hearing deficits. To achieve this, profiles could be merged further than

shown in the current paper to result in fewer profiles. In contrast, a hearing care professional may need a more detailed separation of patients to potentially incorporate information from the profiles into the fitting procedure for hearing aids. Researchers may need an even more detailed separation of profiles to investigate relations and effects of certain audiological features. Here, a cutoff could be selected based on the loss of information for the feature of interest. For instance, if a lot of information regarding a certain feature is lost after a merge, one might select a cutoff prior to the information loss. An example for a more detailed profile set is shown in the Appendix (Figures S.3 and S.4).

The two datasets A and B contained a different set of features but were merged based on common features. While we only show the common features in the current study, the remaining features are also available, e.g., S0N90 for dataset B. This provides the possibility of estimating conditional probabilities of feature ranges given a respective profile, and maintains information provided by the additional features in a descriptive manner.

3.4.3 Classification model and its applications

With the "APP" feature set, we achieve an adequate classification into the 13 combined APs. The APs could, therefore, be predicted with a combination of audiogram, ACALOS, age, and speech test information. The majority of features were determined important, with the most important features being the SRT, AC PTA, ASYM, and L15 1 kHz of the ACALOS **(RQ2)**. As no bone conduction measures are included in this feature set, the current set of APs could also be measured via smartphone. Classification into APs could, thus, be performed on data collected via smartphones. For audiogram-based measures, a variety of implementations already exist (Chen et al., 2021), while implemented speech tests include, for instance, DIN tests (Van den Borre et al., 2021), word recognition tests ((Van Zyl et al., 2018), and the matrix sentence test ((Kollmeier et al., 2015; Saak et al., 2024). As speech tests differ, it would be necessary to consider appropriate ways how to achieve comparability between the available speech tests in different datasets before merging APs generated on the respective dataset.

Hearing care professionals currently measure the audiogram for hearing aid fitting. Depending on the respective regulations and tests available in each country, a speech test in quiet or noise is also used for hearing aid indication (Hoppe and Hesse, 2017). The first-fit of a hearing aid is, however, only based on the audiogram. Speech or loudness measurements are not considered, which appears

to be insufficient to cover all aspects involved for compensating a hearing loss (Kollmeier and Kiessling, 2018). This is also reflected in the features required to classify a patient with respect to the AP found here: While to some extent the classification into profiles also works with the "AG SRT" feature set, the benefit of including loudness scaling shows with the performance improvement of the "APP" feature set. Single measures (single feature group, such as, e.g., the audiogram), in contrast, did not perform well in classifying patients into the profiles. This demonstrates the inability of single measures to characterize the complete extent of hearing deficits sufficiently and shows that, in practice, a combination of measures is needed to adequately characterize audiological patients.

3.4.4 Overall feature importance and interpretability

The feature importance of the merging procedure and the classification models are generally comparable and appear audiologically plausible **(RQ3)**. For both merging and classification, the same common features were initially considered. For merging, we performed no further feature selection. However, for the classification models, the best performing model was selected. Consequently, only features from the best performing model "APP" were considered from the common feature set. This resulted in the AC PTA, SRT, and the L15 (1 kHz) of the ACALOS to be among the most important features. The most important features, thus, cover the combination of threshold information, speech intelligibility, and loudness perception for soft sounds at 1 kHz. Especially in the later merging iterations (Figure 3.3B & Figure 3.4 - merging area A), the speech intelligibility gains importance, which demonstrates the relevance of speech information for our combined profile set. We, therefore, conclude that profiles were merged plausibly, and the underlying procedure is explainable, as we can observe which feature information is lost in each merging iteration and which features are relevant. By investigating the feature importance of the merging procedure, we can observe, for instance, that some information of loudness scaling is lost in earlier iterations (merging area C). To capture all differences regarding loudness scaling, an earlier cutoff could be selected.

Our feature importance results are in line with existing results that highlight the importance of characterizing hearing deficits beyond the audiogram (Musiek et al., 2017). While the audiogram is often seen as the gold standard for characterizing hearing loss, it cannot characterize every aspect of existing deficits (Gieseler et al., 2017; Musiek et al., 2017; Sanchez-Lopez et al., 2021; Van Esch and Dreschler, 2015). Instead, a combination of threshold- and suprathreshold-

based methods is needed, which are covered by audiogram, loudness scaling and speech intelligibility in the present research. Loudness scaling provides additional information beyond the audiogram (Kollmeier and Kiessling, 2018; Launer et al., 1996; Oetting et al., 2016; van Beurden et al., 2021), which is confirmed by the better performance of the classification models if ACALOS is included ("HA" vs. "ALL" and "APP"). This effect is present even if the UCL PTA of the audiogram is included, hence proving the benefit of ACALOS beyond the UCL measure of the audiogram. One interesting finding is that the "AG SRT" and "SRT ACALOS" classification models perform comparable on average. However, profiles with lower SRTs (better speech intelligibility) are slightly better predicted by the SRT in combination with loudness information, and profiles with higher SRTs (worse speech intelligibility) are better predicted by the combination of audiogram information with the SRT. This trend can also be observed in Figure 3.8, where there is a slight trend for higher importance of ACALOS in profiles with lower SRTs as compared to profiles with higher SRTs. One potential explanation could be that L15 from the ACALOS is related to the threshold of the audiogram, as it measures sounds that were perceived as soft. In that way, it could partly cover audiogram-related information and additionally cover loudness-based information. Here, we can note that especially 4 kHz diff (L35-L15, measure of the dynamic range) is more important for profiles with lower SRTs than for profiles with higher SRTs. This is plausible, as it can be expected that individuals that are close to the normal hearing reference have a higher dynamic range and this dynamic range decreases more rapidly with increasing hearing loss for higher frequencies. Regardless, the combination of SRT and ACALOS with audiogram-related information remains necessary for adequate classification performance. Hence, it appears crucial to include threshold and suprathreshold information in characterizing hearing deficits.

3.4.5 Role of common and additional features for profile generation and merging

The available features vary between datasets, and consequently also between profile generation and merging steps. In the profile generation steps, common features (see Table 3.1) and additional features are available (see descriptions of the datasets). In the merging step, profiles can only be merged based on common features. Merged profiles, therefore, contain integrated information on the common features used in the merging process. Note that both datasets employed included information on the individual audiogram, speech recognition in noise and loudness scaling which has some impact on the resulting finding,

that all three information areas are relevant for classifying the individual patient. This would not have shown up if clinical datasets would have been employed with much less suprathreshold audiological information, which is both a strength and a potential pitfall of the current study. Additional features are not used to merge the APs, as they are not contained within both datasets, but were included for the generation of the profiles. The rationale behind this is that we aimed to make the first patient grouping based on the most informed choice using all available information - which includes the information contained in the additional features. That means the additional features can impact the initial grouping of the patients by allowing for finer distinctions, as compared to profile generation based solely on the common features. To exemplify, using only the AC PTA and the SRT would create coarser groupings and miss certain subgroups that are revealed when including features from the ACALOS. In later merging steps, it is then possible to obtain profile sets where the information provided by additional features is either maintained or cancelled out by the merging. It follows, that a certain number of common features should be available to adequately merge profiles generated from different datasets and to classify patients into a given profile. Given the provided feature importance analysis, we advocate for using a combination of threshold, loudness- and speech test information. That is because these measures were consistently estimated as being important for both the merging and the later classification into the profile. However, depending on the use case and data availability, profile generation and potential later merging may also be used on an exploratory basis to learn about the content of a dataset at hand.

Another property of the additional features is that they could be used to infer probabilities within a given profile where data for the additional profiles is available. Here, it could be of interest to investigate whether certain feature ranges occur more frequently in certain profiles – especially if profiles were merged with at least one further profile containing the additional profile.

3.4.6 Towards a population-based set of combined auditory profiles

The combined profile set is the next step towards a population-based estimate of APs. In the future, additional datasets need to be integrated to cover further hearing deficits, as well as information from different audiological measures. While the new set of APs does include new information, continuing to integrate additional datasets can improve profile definitions and classification models. Currently, one large profile (AP 8) describes most of the patients due to the

obviously high relative frequency of this profile (corresponding to a moderate hearing loss, potentially age-related hearing loss) in the two employed datasets. Therefore, including additional datasets can help in better defining the remaining profiles based on a higher, more representative number of patients. If representative data for the whole population were included, the relative frequencies of patients contained in different profiles would allow for a prevalence estimate of different profiles. In addition, the current datasets mainly cover hearing aid candidates, whereas cochlear implant candidates are rarely included. A population-based set should, however, include all degrees of hearing deficits and be as complete as possible.

Our AP merging approach presented in this study provides properties of a federated learning procedure that can be used to obtain a population-based set of APs in the future, without sharing individual patient data. The federated learning procedure includes the steps to generate and later merge profiles. One important property of the merging step is its ability to work with anonymized data. This is important to integrate datasets containing sensitive clinical data which often underly data sharing restriction (Gauthier, 2017). For these datasets, the profile generation could occur at the sensitive data location and only the anonymized count data would be shared. As overlapping density is calculated for each feature separately and then averaged, it also enables shuffling patient records for each feature to completely anonymize the patient data. This does not impact the distributional descriptions of each feature contained in the profiles and therefore enables integration of sensitive datasets to work towards a population-based estimate of APs. However, to reach such a population-based estimate, we need to continue to merge profiles generated on different datasets until profiles converge on a final set of APs. A convergence would mean, that no new APs are added to the existing set of APs when integrating additional datasets. This follows the common principle of convergence in optimization algorithms, where the optimal solution (here sets of APs) is selected when parameters converge (Hastie et al., 2009).

The next step to obtain a population-based set of APs is, therefore, to integrate further datasets. For this, data standards will be important. If data formats and required meta data are standardized across institutions, barriers for integrating datasets are reduced and could pave the way for big data analyses in the field of audiology.

Finally, with a sufficiently large AP set, in terms of included patients, and varying deficits, it could be interesting to investigate the feasibility of profile-based first-fits for hearing aids. Profile-based fits have the potential to reduce the time required to fine-tune hearing aids, by incorporating information from the audiogram, speech test results, and loudness scaling. Especially for the individual selection of parameters for additional signal processing parameters a manufacturer-independent profile-based recommendation would help hearing care professionals by better individualize the first-fit parameters. Furthermore, first-fits could already include information on both thresholds and loudness loss. The benefit of using loudness-based fitting has been shown by Kramer et al. (2020) and Oetting et al. (2018). Since they highlight the importance of binaural broadband loudness scaling, future profile generation and evaluation should also include binaural broadband categorical loudness scaling.

3.5 Conclusions

The current study demonstrates the feasibility of integrating datasets using the proposed profile generation and merging procedure, which qualifies as a federated learning approach for a combined characterization of the content of audiological datasets. Combining the two datasets yields a new combined set of APs, which consists of 13 APs. Profiles can generally be well characterized based on the three dimensions: audiogram-, speech test-, and loudness scaling features.

We further enable the classification of patients into the APs using random forest classification models. The best performing classification models include a combination of these three measures (audiogram, speech test, loudness scaling), excluding the bone-conduction audiogram. While classification models tailored for hearing care professionals, which use only the audiogram and a speech test, are also available, their performance can be improved if loudness scaling is included.

Audiogram-, speech test-, and loudness scaling-based measures provide complementary information that aid in characterizing patients and are consistently determined as important for both merging and classification. Even though this finding is based on the composition of the two underlying datasets that both include these measures, we nevertheless advocate for the inclusion of these measures for detailed patient characterization.

Towards a population-based set of APs, it is necessary to incorporate additional

datasets using the proposed method. These datasets should include a wider range of audiological subpopulations, such as cochlear implant users, thereby extending the APs contained. The privacy-preserving approach, employing federated learning, which involves the potential to share only data distributions from locally generated profiles, facilitates the integration of datasets that are subject to privacy restrictions.

Conflict of interest

The authors declare that the research was conducted in the absence of any conflicts of interest.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 – Project ID 390895286. MB was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 496819293, and the "Fondation Pour l'Audition" (FPA IDA10).

Data availability

The data analyzed in this study was obtained from Hörzentrum Oldenburg gGmbH, the following licenses/restrictions apply: According to the Data Usage Agreement of the authors, the datasets analyzed in this study can only be shared upon motivated request. Requests to access these datasets should be directed to DO, oetting@hz-ol.de. Note that dataset A will soon be published. The analyses scripts can be found here: Zenodo, https://doi.org/10.5281/zenodo.13132817.

Bibliography

Audigier, V., Husson, F., and Josse, J. (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, 10:5–26.

Banerjee, A. and Shan, H. (2017). *Encyclopedia of Machine Learning and Data Mining*, chapter Model-Based Clustering, pages 848–852. Springer US.

Beyan, O., Choudhury, A., Van Soest, J., Kohlbacher, O., Zimmermann, L., Stenzhorn, H., Karim, M. R., Dumontier, M., Decker, S., da Silva Santos,

L. O. B., et al. (2020). Distributed analytics on sensitive medical data: the personal health train. *Data Intelligence*, 2(1-2):96–107.

Bisgaard, N., Vlaming, M. S., and Dahlquist, M. (2010). Standard audiograms for the iec 60118-15 measurement procedure. *Trends in amplification*, 14(2):113–120.

Brand, T. and Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. *The Journal of the Acoustical Society of America*, 112(4):1597–1604.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Chen, C.-H., Lin, H.-Y. H., Wang, M.-C., Chu, Y.-C., Chang, C.-Y., Huang, C.-Y., Cheng, Y.-F., et al. (2021). Diagnostic accuracy of smartphone-based audiometry for hearing loss detection: meta-analysis. *JMIR mHealth and uHealth*, 9(9):e28378.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

De Sousa, K. C., Smits, C., Moore, D. R., Myburgh, H. C., and Swanepoel, D. W. (2022). Diotic and antiphasic digits-in-noise testing as a hearing screening and triage tool to classify type of hearing loss. *Ear and hearing*, 43(3):1037–1048.

do Carmo, L. C., da Silveira, J. A. M., Marone, S. A. M., D'Ottaviano, F. G., Zagati, L. L., and von Söhsten Lins, E. M. D. (2008). Audiological study of an elderly brazilian population. *Brazilian journal of otorhinolaryngology*, 74(3):342–349.

Ehwerhemuepha, L., Gasperino, G., Bischoff, N., Taraman, S., Chang, A., and Feaster, W. (2020). Healthedatalab–a cloud computing solution for data science and advanced analytics in healthcare with application to predicting multicenter pediatric readmissions. *BMC medical informatics and decision making*, 20:1–12.

European Commision, D.-G. f. H. and Food Safety, Renner A, B. J. L. A. (2016). Study on big data in public health, telemedicine and healthcare – final report. https://data.europa.eu/doi/10.2875/734795.

European Parliament, E. (2016). Regulation (eu) 2016/679 of the european parliament and of the council, of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of

such data, and repealing directive 95/46/ec (general data protection regulation). *Official Journal of the European Union*, L 119:1–88.

Fraley, C. and Raftery, A. E. (2003). Enhanced model-based clustering, density estimation, and discriminant analysis software: Mclust. *Journal of Classification*, 20(2):263–286.

Gauthier, C. (2017). The impact of the eu general data protection regulation on scientific research. *ecancermedicalscience*, 11.

Gemeinsamer Bundesausschuss (2021). Richtlinie des gemeinsamen bundesausschusses über die verordnung von hilfsmitteln in der vertragsärtzlichen versorgung. *Bundesanzeiger*, B3.

Gieseler, A., Tahden, M. A., Thiel, C. M., Wagener, K. C., Meis, M., and Colonius, H. (2017). Auditory and non-auditory contributions for unaided speech recognition in noise as a function of hearing aid use. *Frontiers in psychology*, 8:219.

Grant, R. W., McCloskey, J., Hatfield, M., Uratsu, C., Ralston, J. D., Bayliss, E., and Kennedy, C. J. (2020). Use of latent class analysis and k-means clustering to identify complex patient profiles. *JAMA network open*, 3(12):e2029068–e2029068.

Hajibaba, M. and Gorgin, S. (2014). A review on modern distributed computing paradigms: Cloud computing, jungle computing and fog computing. *Journal of computing and information technology*, 22(2):69–84.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hoppe, U. and Hesse, G. (2017). Hearing aids: indications, technology, adaptation, and quality control. *GMS current topics in otorhinolaryngology, Head and Neck Surgery*, 16.

Kalbe, E., Kessler, J., Calabrese, P., Smith, R., Passmore, A., Brand, M. a., and Bullock, R. (2004). Demtect: a new, sensitive cognitive screening test to support the diagnosis of mild cognitive impairment and early dementia. *International journal of geriatric psychiatry*, 19(2):136–143.

Kollmeier, B. and Kiessling, J. (2018). Functionality of hearing aids: State-of-the-art and future model-based solutions. *International journal of audiology*, 57(sup3):S3–S28.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., and Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International journal of audiology*, 54(sup2):3–16.

Kollmeier, B. and Wesselkamp, M. (1997). Development and evaluation of a german sentence test for objective and subjective speech intelligibility assessment. *The Journal of the Acoustical Society of America*, 102(4):2412–2421.

Kramer, F., Oetting, D., Schädler, M. R., Hohmann, V., and Warzybok, A. (2020). Speech recognition and loudness perception in normal-hearing and hearing-impaired listeners. In *Forum Acusticum*, pages 3493–3493.

Launer, S., Holube, I., Hohmann, V., and Kollmeier, B. (1996). Categorical loudness scaling in hearing-impaired listeners-can loudness growth bei predicted from the audiogram? *Audiological Acoustics*, 35:156–163.

Lê, S., Josse, J., and Husson, F. (2008). Factominer: an r package for multivariate analysis. *Journal of statistical software*, 25:1–18.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Musiek, F. E., Shinn, J., Chermak, G. D., and Bamiou, D.-E. (2017). Perspectives on the pure-tone audiogram. *Journal of the American Academy of Audiology*, 28(07):655–671.

Oetting, D., Hohmann, V., Appell, J.-E., Kollmeier, B., and Ewert, S. D. (2016). Spectral and binaural loudness summation for hearing-impaired listeners. *Hearing Research*, 335:179–192.

Oetting, D., Hohmann, V., Appell, J.-E., Kollmeier, B., and Ewert, S. D. (2018). Restoring perceived loudness for listeners with hearing loss. *Ear and hearing*, 39(4):664–678.

Parimbelli, E., Marini, S., Sacchi, L., and Bellazzi, R. (2018). Patient similarity for precision medicine: A systematic review. *Journal of biomedical informatics*, 83:87–96.

Pastore, M. and Calcagnì, A. (2019). Measuring distribution similarities between samples: a distribution-free overlapping index. *Frontiers in psychology*, 10:1089.

Pfitzner, B, S. N. A. B. (2021). Federated learning in a medical context: a systematic literature review. *ACM Transactions on Internet Technology (TOIT)*, 21(2):1–32.

Plomp, R. (1986). A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *Journal of Speech, Language, and Hearing Research*, 29(2):146–154.

Roth, T. N., Hanebuth, D., and Probst, R. (2011). Prevalence of age-related hearing loss in europe: a review. *European Archives of Oto-Rhino-Laryngology*, 268:1101–1107.

Saak, S., Huelsmeier, D., Kollmeier, B., and Buhl, M. (2022). A flexible data-driven audiological patient stratification method for deriving auditory profiles. *Frontiers in Neurology*, 13:959582.

Saak, S., Kothe, A., Buhl, M., and Kollmeier, B. (2024). Comparison of user interfaces for measuring the matrix sentence test on a smartphone. *International Journal of Audiology*, pages 1–13.

Sanchez-Lopez, R., Fereczkowski, M., Neher, T., Santurette, S., and Dau, T. (2020). Robust data-driven auditory profiling towards precision audiology. *Trends in hearing*, 24:2331216520973539.

Sanchez-Lopez, R., Nielsen, S. G., El-Haj-Ali, M., Bianchi, F., Fereczkowski, M., Cañete, O. M., Wu, M., Neher, T., Dau, T., and Santurette, S. (2021). Auditory tests for characterizing hearing deficits in listeners with various hearing abilities: The bear test battery. *Frontiers in neuroscience*, 15:724007.

Schmidt, K. and Metzler, P. (1992). Wst-wortschatztest. *Göttingen: Beltz Test*, 16.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Sinkala, M., Mulder, N., and Martin, D. (2020). Machine learning and network analyses reveal disease subtypes of pancreatic cancer and their molecular characteristics. *Scientific reports*, 10(1):1212.

Stöver, T., Plontke, S., Guntinas-Lichius, O., Welkoborsky, H., Zahnert, T., Delank, K., Deitmer, T., Esser, D., Dietz, A., Wienke, A., et al. (2023). Struktur und einrichtung des deutschen cochlea-implantat-registers (dcir). *Hno*, 71(12):767–778.

van Beurden, M., Boymans, M., van Geleuken, M., Oetting, D., Kollmeier, B., and Dreschler, W. A. (2021). Uni-and bilateral spectral loudness summation and binaural loudness summation with loudness matching and categorical loudness scaling. *International Journal of Audiology*, 60(5):350–358.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67.

Van den Borre, E., Denys, S., van Wieringen, A., and Wouters, J. (2021). The digit triplet test: a scoping review. *International journal of audiology*, 60(12):946–963.

Van Esch, T. and Dreschler, W. (2015). Relations between the intelligibility of speech in noise and psychophysical measures of hearing measured in four languages using the auditory profile test battery. *Trends in Hearing*, 19:2331216515618902.

Van Esch, T. E., Kollmeier, B., Vormann, M., Lyzenga, J., Houtgast, T., Hällgren, M., Larsby, B., Athalye, S. P., Lutman, M. E., and Dreschler, W. A. (2013). Evaluation of the preliminary auditory profile test battery in an international multi-centre study. *International journal of audiology*, 52(5):305–321.

Van Zyl, M., Swanepoel, D. W., and Myburgh, H. C. (2018). Modernising speech audiometry: using a smartphone application to test word recognition. *International Journal of Audiology*, 57(8):561–569.

Zokoll, M. A., Wagener, K. C., Brand, T., Buschermöhle, M., and Kollmeier, B. (2012). Internationally comparable screening tests for listening in noise in several european languages: The german digit triplet test as an optimization prototype. *International Journal of Audiology*, 51(9):697–707.

S.1 Supplementary Materials

Table S.1: Distribution of common and additional features (prior to imputation) for dataset A and B. Mean, mean, and standard deviation (SD) are shown. The hyphen indicates that the features was not available for the respective dataset.

		Dataset A			Dataset B		
		Median	**Mean**	**SD**	**Median**	**Mean**	**SD**
Speech test	GOESA S0N0 bin [dB SNR]	-2.2	-1.46	3.09	-2.1	-0.79	4.27
Audiogram	AC PTA [dB HL]	37.5	38.75	20.93	45	47.64	22.09
	ASYM [dB]	3.75	7.49	12.49	6.25	11.34	15.42
	BC PTA [dB HL]	30	29.29	18.07	38.75	38.43	15.95
	ABG [dB]	6.25	8.1	8.18	2.5	4.9	6.32
	UCL PTA [dB HL]	96.25	97.09	10.54	98.75	98.35	10.12
Anamnesis	Age [years]	70	67.6	11.89	65.8	63.74	13.22
ACALOS	1_L15 [dB]	61.93	62.45	12.26	62.44	64.88	14.41
	1_L35 [dB]	94.27	94.1	7.82	95.42	95.7	8.81
	1_diff [dB]	31.4	31.64	9.11	30.86	30.81	9.72
	4_L15 [dB]	73.69	72.17	15.42	76.37	76.21	15.84
	4_L35 [dB]	97.14	97.37	9.93	98.41	98.53	10.64
	4_diff [dB]	23.07	25.19	9.71	21.24	22.32	8.71
Additional features Dataset A	GOESA S0N0 slope [dB SNR]	0.18	0.2	0.11	-	-	-
	DIN S0N0 [dB SNR]	-5.83	-4.61	4.13	-	-	-
	Socio-economic status	12	12.89	3.94	-	-	-
	DemTect	17	15.82	2.31	-	-	-
	Vocabulary Test	32	31.44	4.93	-	-	-
Additional features Dataset B	GOESA S0N90 bin [dB SNR]	-	-	-	-5.7	-4.39	5.25
	GOESA S0N90 mon [dB SNR]	-	-	-	-3.7	-2.47	4.75
	GOESA ILD [dB]	-	-	-	3.7	3.6	1.99
	GOESA BILD [dB]	-	-	-	2.1	1.94	2.21
	ACALOS 2 kHz L15 [dB]	-	-	-	67.94	68.57	14.3
	ACALOS 2 kHz L35 [dB]	-	-	-	94.75	95.44	8.98
	ACALOS 2 kHz diff (L35-L15) [dB]	-	-	-	18.45	19	12.11

Table S.2: Number of patients in each category for Otoscopy and Valsalva (better and worse ear). Missing data was not imputed and therefore sample sizes may vary.

	Dataset B		
	Okay	**Not okay**	**Not completely okay**
Otocoscopy worse ear	901	94	226
Otoscopy better ear	968	63	191
Valsalva worse ear	1003	48	48
Valsalva better ear	1041	31	108

Table S.3: Number of patients in each category for Tympanogram (better and worse ear). Missing data was not imputed and therefore sample sizes may vary.

	Dataset B						
	A	**B**	**C**	**As**	**Ad**	**TM Perforation**	**Not measurable**
Tympanogram worse ear	987	25	29	24	112	4	32
Tympanogram better ear	1035	9	18	22	99	1	25

Table S.4: Number of patients in each Bisgaard class. Missing data was not imputed and therefore sample size may vary.

Bisgaard class	**Dataset A**	**Dataset B**
N1	126	141
N2	95	207
N3	136	287
N4	46	167
N5	21	79
N6	13	61
N7	8	38
S1	66	74
S2	51	119
S3	33	87

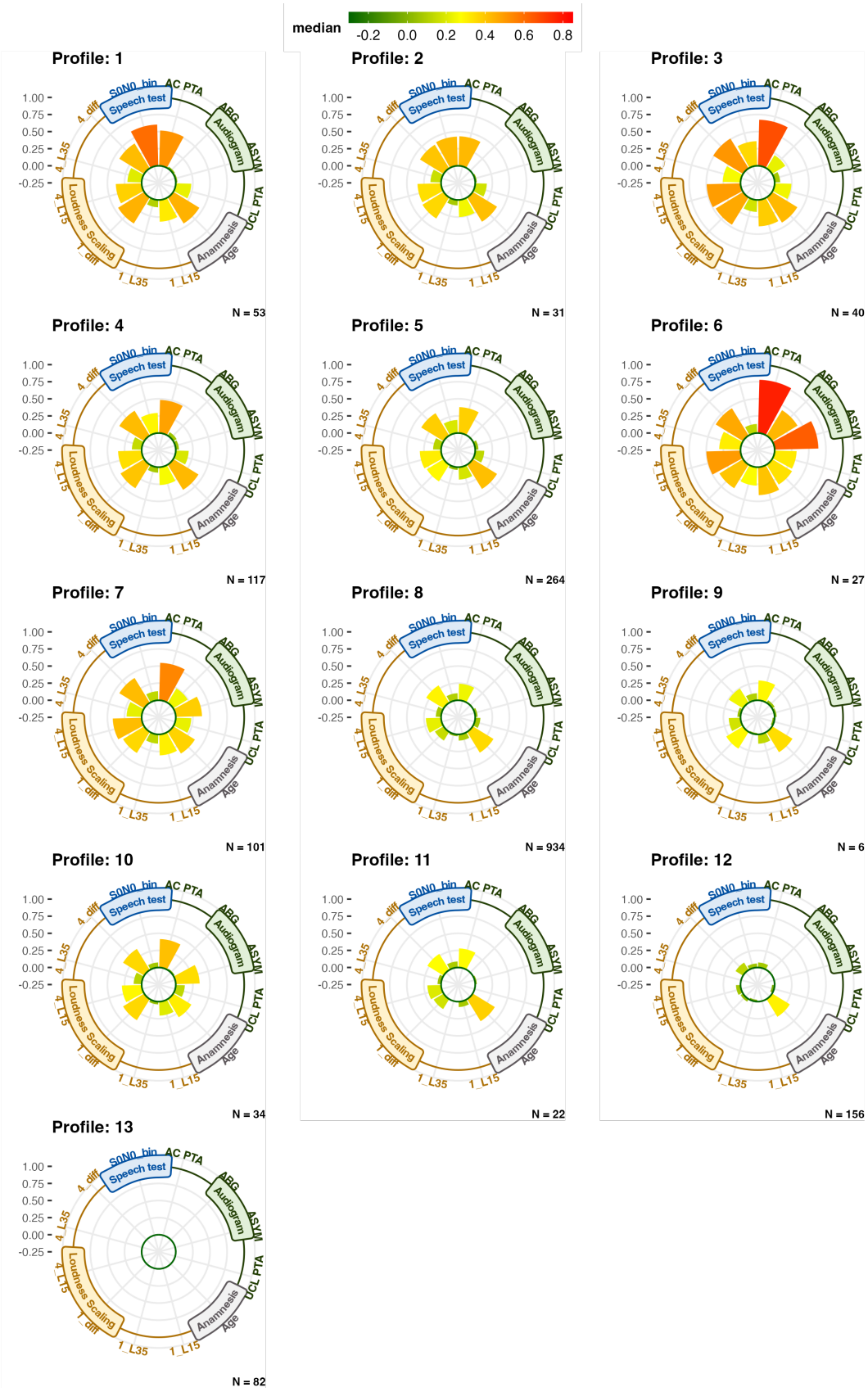

Figure S.1: Polar Profile plots for the 13 auditory profiles.

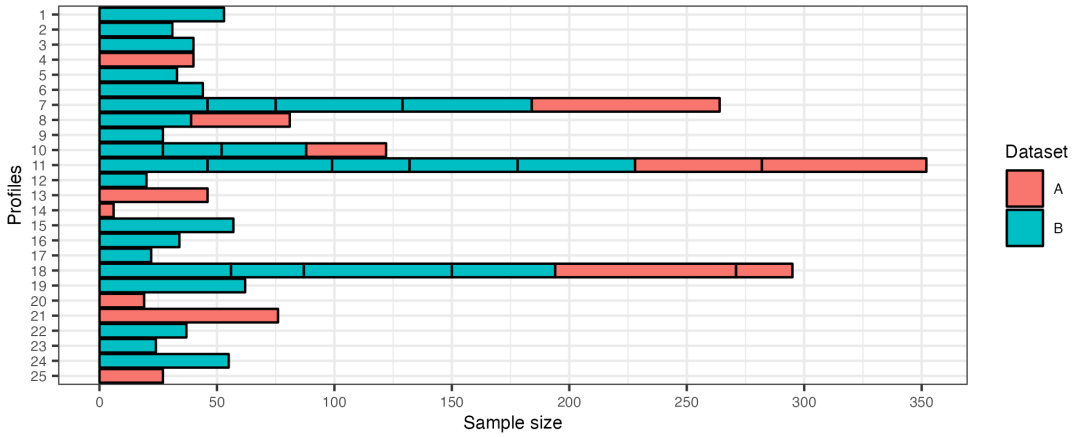

Figure S.2: Distribution of profiles that were merged to result in the larger profile set with 25 auditory profiles. A corresponds to dataset A and presents the previous 13 profiles detailed in (1). B refers to the new dataset B and the 31 profiles that were defined as optimal by the profile generation pipeline. The x-axis depicts the sample size for each sub-profile, as well as the proposed 25 profiles.

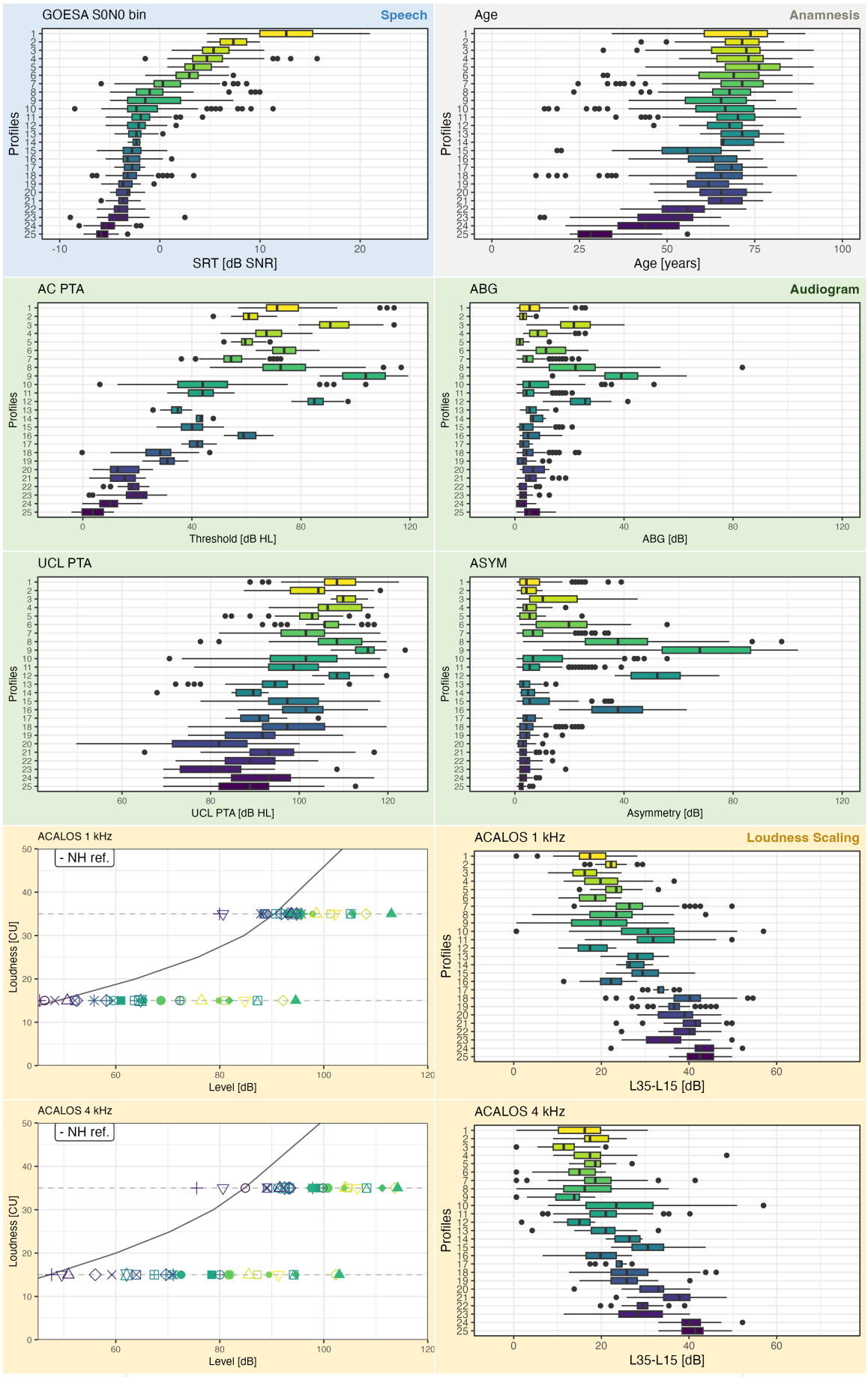
GOESA S0N0 bin
Speech
SRT [dB SNR]
Age
Anamnesis
Age [years]
AC PTA
Threshold [dB HL]
ABG
Audiogram
ABG [dB]
UCL PTA
UCL PTA [dB HL]
ASYM
Asymmetry [dB]
ACALOS 1 kHz
- NH ref.
Loudness [CU]
Level [dB]
ACALOS 1 kHz
Loudness Scaling
L35-L15 [dB]
ACALOS 4 kHz
- NH ref.
Loudness [CU]
Level [dB]
ACALOS 4 kHz
L35-L15 [dB]
Profiles
Profile
25 23 21 19 17 15 13 11 9 7 5 3 1
24 22 20 18 16 14 12 10 8 6 4 2

Figure S.3: Profile distribution for the larger profile set with 25 Auditory Profiles across the speech, anamnesis, audiogram, and loudness scaling domain. Profile ranges are depicted for all features and are ordered with respect to the increasing median SRT.

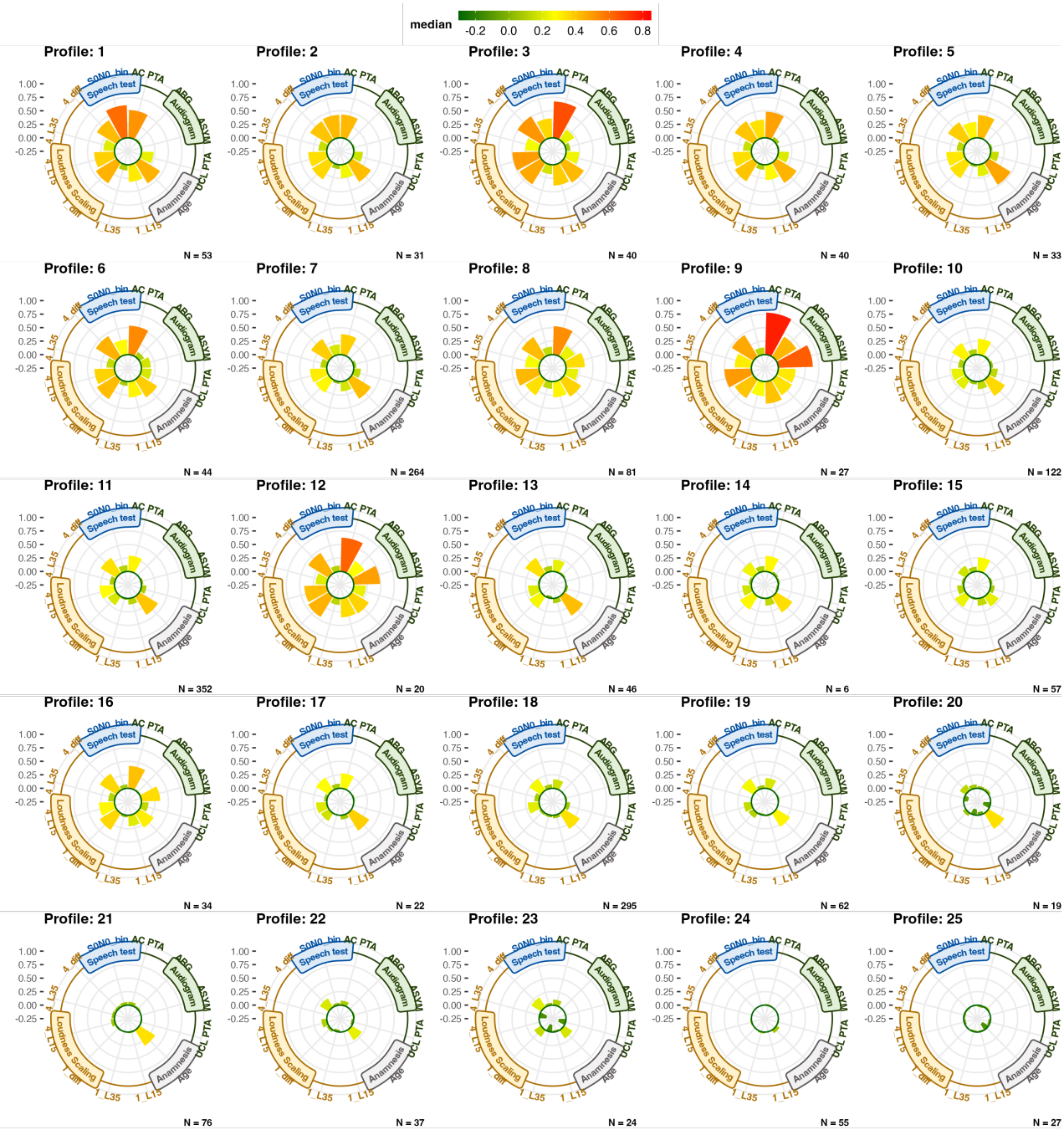

Figure S.4: Polar Profile plots for the 25 auditory profiles

4 Comparison of user interfaces for measuring the matrix sentence test on a smartphone

Bibliographic reference

This chapter is a formatted reprint of the following paper. It is identical in content and has been published in the International Journal of Audiology.

Saak, S., Kothe, A., Buhl, M., & Kollmeier, B. (2024). Comparison of user interfaces for measuring the matrix sentence test on a smartphone. International Journal of Audiology, 1-13. https://doi.org/10.1080/14992027.2024.2385551

Author contribution

S. Saak, M. Buhl, and B. Kollmeier conceptualized and designed the study. S. Saak developed the web application for the study and set up the experiment. S. Saak and A. Kothe conducted the measurements with the participants. S. Saak performed the statistical analyses and wrote the draft of the manuscript. All authors contributed to reviewing and editing the manuscript.

Prof. Dr. Dr. Birger Kollmeier

Abstract

Smartphone-based self-testing could facilitate large-scale data collection and remote diagnostics. For this purpose, the matrix sentence test (MST) is an ideal candidate due to its repeatability and accuracy. In clinical practice, the MST requires professional audiological equipment and supervision, which is infeasible for smartphone-based self-testing. Therefore, it is crucial to investigate the feasibility of self-administering the MST on smartphones, including the development of an appropriate user interface for the small screen size.

We compared the traditional closed matrix user interface (10x5 matrix) to three alternative, newly-developed interfaces (slide, type, wheel) regarding SRT consistency, user preference, and completion time. We included 15 younger normal hearing and 14 older hearing-impaired participants in our study.

The slide interface is most suitable for mobile implementation, providing consistent and fast SRTs and enabling all participants to perform the tasks effectively. While the traditional matrix interface works well for most participants, some participants experienced difficulties due to its small size on the screen.

We propose the newly-introduced slide interface as a plausible alternative for smartphone screens. This might be more attractive for elderly patients that may exhibit more challenges with dexterity and vision than our test subjects employed here.

Keywords: Speech test, matrix sentence test, mobile audiology, mobile self-testing, user interfaces

4.1 Introduction

Smartphones can facilitate remote data collection and diagnostics and can therefore enable the collection of large-scale datasets. Collecting large datasets to extract information is highly relevant to investigate the complex interdependencies between genetic, medical, environmental, and lifestyle aspects for varying diseases (Allen et al., 2012). In this context, it is important to include speech testing, as understanding speech is a crucial part of every-day life and needed for effective communication with others (World Health Organization, 2021). The matrix sentence test (MST) provides the possibility to accurately assess speech intelligibility with a high precision of ± 1 dB SNR (Wagener,

2004). The MST is used in research, and is used in clinical contexts in several countries, such as for hearing aid or cochlear implant benefit assessment. It measures speech intelligibility in noise using semantically unpredictable 5-word sentences, such as "Doris brought nineteen large sofas". As a result, it provides an ecologically more relevant estimate of an individual's hearing challenges in real life than a speech test in quiet, while being less dependent on a quiet test environment and an exact calibration of the test equipment (Kollmeier et al., 2015). Moreover, the (unaided) test outcomes are still grossly related to the average audiogram (Wardenga et al., 2018). A smartphone implementation, however, is currently missing to also allow for individual self-testing in a remote setting, allowing precise characterisation of speech intelligibility deficits and the assessment of hearing device benefit. The MST is an ideal candidate for remote diagnostics, as it is a repeatable speech test available in 20 languages covering 60% of the world's population (Hörzentrum Oldenburg gGmbH, n.d.; Kollmeier et al., 2015). If listeners conduct the MST by themselves (without experimenter), the current user interface (UI) of the MST presents all 50 words at once to the user in form of a 10 x 5 matrix. This may prove problematic for a smartphone implementation, given the restricted display size of smartphones. Current guidelines point out the necessity to simplify the user interface (UI) design and increase the size and distance between interactive controls for the elderly (Gomez-Hernandez et al., 2023). This is especially relevant, as the majority of potential users of such a mobile speech test implementation are above 65 years and could be affected by age-dependent declines in visual acuity, motor skills, and cognitive abilities, which could hinder the effective interaction with the UI (Farage et al., 2012; Salman et al., 2023; Wong et al., 2010).

First, a decline in visual acuity can lead to difficulties with detecting and discriminating details (Farage et al., 2012) on mobile interfaces/small screens. Second, age-related decline of fine motor control reduces the precision of arm, hand, and finger movements (Bowden and McNulty, 2013; Saunders et al., 2021) and in turn, also increase the required time for task completion (Seidler et al., 2010). Presenting all 50 words of the MST could, therefore, be problematic on the small screen of smartphones, as small fonts might pose a problem, buttons, and button spacing might be too small (Hwangbo et al., 2013; Lee and Kuo, 2007), and too much information might be presented at once (Wong et al., 2010). Third, an age-related decline in cognitive abilities can result in reduced working memory capacities, and processing speed, among others (Park and Schwarz, 2000). Reduced working memory capacities could affect the ability to remember

presented sentences, in that way adversely affecting task performance; reduced processing speeds could more generally increase completion time and perceived task demands (Borella et al., 2011; Murman, 2015).

For a successful smartphone-based implementation of the MST, it is therefore crucial to employ an interface that users can perform the tasks with in general, and also do so efficiently and satisfactorily (Lewis, 2014). In addition, the interface needs to consider potential difficulties that elderly may experience with the application. In other words, a potential interface for the MST needs to be usable by the elderly target group. Unfortunately, older users are often not considered in the development of technology applications (Chun and Patterson, 2012). As a result, usability for the elderly is often not optimal, fostering technological anxiety and inhibiting its uptake (Frishammar et al., 2023).

To reduce the cognitive complexity, the simplified MST could be used. The simplified MST is a reduced version of the MST, only contains 3-word sentences, e.g. “seven old windows”, and is currently available in 5 languages (gGmbH, nd; Wagener K, 2005). The simplified MST was first intended to be used with children, but later turned out to be useful for older patients as well (Buschermöhle et al., 2016). It could prove to be a useful speech test for smartphones, given that it also reduces the amount of information that needs to be displayed on the smartphone, and the number of physical interactions with the interface.

Next to fulfilling usability requirements for the elderly, an application needs to be usable within its environment. If laboratory or advanced audio equipment are required, most individuals will not have easy access to remote testing. In contrast, readily available equipment may increase the usage. The feasibility of audiometric screening via smartphones with inexpensive headphones has been shown (Hussein et al., 2016; Swanepoel et al., 2014). Likewise, the Digits-in-Noise (DIN) test has proven successful as a screening test via telephone and the internet (Smits et al., 2004; Zokoll et al., 2012). The speech material consists of spoken numbers, which can be selected from a telephone interface, and the DIN is fast and easy to conduct. It is therefore well suited for hearing screening and is also measured as part of the large-scale biomedical database of the UK Biobank (Sudlow et al., 2015). However, these tools have primarily focused on screening whereas the MST is focused on diagnostic evaluations. The MST takes longer to conduct in order to achieve its high reliability but requires training in order to compensate for the training effect which is measurable due

to the high efficiency of the test (Kollmeier et al., 2015). The MST also offers an advantage in that it represents every-day-language and its speech spectrum in the form of syntactically correct sentences. The MST is currently used for clinical evaluations and the fitting and benefit assessment of hearing aids and cochlear implants. Consequently, a mobile version of the MST could facilitate comprehensive speech recognition assessments in remote data collection and allow for the comparability with clinical applications.

Several implementations already exist for measuring the German MST (also called Oldenburg Sentence Test) outside the lab. Ooster et al. (2020) implemented a system for measuring the German MST via Amazon Alexa using automatic speech recognition (ASR). Likewise, Bruns et al. (2022) developed a voice-over-IP system for measuring the German MST via telephone, also using an ASR system. However, both implementations use technology without a graphical UI, which represents the open-set version of the MST that yields in most languages significantly different results from the clinically employed closed-set version where all word options are displayed to the user (Kollmeier et al., 2015), thus limiting the comparability to clinical data. Moreover, a mobile user interface with the closed version of the MST could further ease the training with the speech material as it enables the users to visually familiarise themselves with the speech material.

Implementing the MST on a smartphone and providing an appropriate graphical UI is, therefore, highly relevant to facilitate remote data collection. It is, however, not yet known to what extent elderly hearing-impaired (oHI) individuals can conduct the MST and simplified MST accurately on a smartphone, given the combination of a small screen size, and potential age-dependent declines in cognitive ability, visual acuity, and motor skills. With the present study, therefore, we aim at (1) exploring the general feasibility of measuring the MST on a smartphone. For that purpose, we investigate the performance of younger normal hearing (yNH) individuals on smartphone implementations of both the German MST and the German simplified MST. We hypothesise that the test will be feasible with a smartphone and household in-ear headphones. We further aim at (2) proposing a suitable interface that complies with design needs of elderly. To that end, we compare the yNH and the oHI group with respect to usability aspects, namely, their performance consistency, completion time and user preference, across four suggested potential interfaces for the MST, and the typical (matrix) interface of the simplified MST. The four interfaces differ

regarding layout and response options. Ultimately, we (3) propose an interface for the MST that both user groups can easily cope with, is time efficient, and provides accurate speech recognition threshold (SRT) results. Hence, our research should allow us to answer the following research questions:

RQ1: Are SRTs obtained for the MST and simplified MST comparable between (calibrated) smartphone measurements and controlled, typical lab measurements, for both yNH and oHI participants?

RQ2a: Is there a group effect (age and hearing loss; yNH vs.oHI) on SRT consistency, completion time, and user preference?

RQ2b: If a group effect is present, is it specific to certain interfaces?

RQ3: Which MST interface is most suitable for both oHI and yNH in terms of SRT consistency, time efficiency, and user preference?

4.2 Materials and methods

4.2.1 Participants

We recruited 32 participants for our study: one young, normal-hearing (yNH) group, and one older, hearing-impaired (oHI) group with mild to moderate hearing loss. The yNH group was recruited via posts at the University website, while participants for the oHI group were recruited via the Hörzentrum Oldenburg gGmbH. The sample size was chosen such that an SRT difference of 1 dB SNR can be statistically detected with a power of 80%. Calculations were performed using the simr package in R (Green and MacLeod, 2016). All participants gave written informed consent and were paid for their participation in the study. Inclusion criteria for the yNH group were a pure tone average (PTA, hearing thresholds averaged over for 500, 1000, 2000, and 4000 Hz) < 20 dB HL and age < 50 years; inclusion criteria for the oHI group were a symmetric hearing loss (PTA difference < 10 dB) with a PTA > 20 dB HL, and age ≥ 50 years. Fifteen participants qualified for the yNH group (*mean age* $= 23.67$, $SD = 2.32$, *female* $= 66.7\%$) and 14 qualified for the oHI group (*mean age* $= 73.77$, $SD = 6.73$, *female* $= 53.8\%$). The remaining three participants (oHI) were excluded, as they were not able to manipulate a smartphone sufficiently with their fingers, but instead required a pen. In the yNH group every participant owned a smartphone. 73.33% used it almost continuously during the day; 26.67% used it once or more during the day.

In contrast, for the oHI Group 10% used it almost continuously during the day, and the remaining 90% used it once or more during the day. Figure 4.1 depicts the audiogram ranges for the yNH and oHI groups.

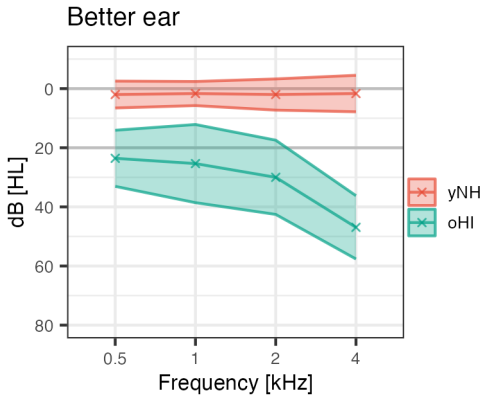

Figure 4.1: Audiogram ranges for the yNH and oHI groups for their respective better ear. Means are shown as well as ranges corresponding to one standard deviation for yNH and oHI.

4.2.2 Research design

We designed a crossover study to examine the overall feasibility of measuring both the MST (Wagener, 2004) and simplified MST (Wagener K, 2005) on a smartphone, and to investigate which interface is best for measuring the MST. The measurement language was German. The MST consists of a 10 x 5 matrix in which sentences are built by presenting word combinations in a fixed name-verb-number-adjective-object syntactical structure. For instance, "Peter has seven old windows". The simplified MST is a reduced version of the MST, consisting of a 10 x 3 matrix (number-adjective-object). Each participant conducted several measurement conditions of the MST and the simplified MST in both a reference laboratory control session and in the smartphone test session. The order of test and control session, and the order of the measurement conditions were randomised. All tests were conducted in a soundproof-booth and both sessions used calibrated setups to control the levels and to allow comparing the influence of equipment and UIs. Ethical approval was obtained from the Research Ethical Committee of the Universität Oldenburg [Drs. 71/2015].

4.2.2.1 Laboratory control session The laboratory control session was conducted using the software Oldenburg Measurement Application (OMA) by the Hörzentrum Oldenburg gGmbH with HDA200 headphones. By using this software, the exact same measurements are performed as in clinical practice.

Both the MST and simplified MST in OMA were measured in an open version to mimic the standard clinical procedure. That means, participants repeated the words they understood to the experimenter (open version), instead of selecting the respective words from a matrix displayed on a screen (closed version). OMA uses the adaptive procedure by Brand and Kollmeier (2002) to control levels of speech and noise. Calibration was performed using the test-specific noise of the German MST and an artificial ear.

4.2.2.2 Smartphone test session We developed a web-based implementation to measure both the matrix and the simplified matrix test via an interface optimised for smartphones. The underlying measurement procedures mimic the procedure used by OMA. The backend of the application was build using Python Flask, Octave, and Bash. Python Flask is a common framework for web-development. Octave was chosen for implementing the sentence tests to reuse available measurement scripts for the MST (Schädler, 2021). The communication between Python Flask and Octave was then enabled via the shell language Bash. The frontend and the interfaces were built using the common mixture of JavaScript, HTML, and CSS. A Linux laptop served as a server to host the application. The website was then opened via a browser on a OnePlus Nord N10 (Android operating system) smartphone with inexpensive in-ear Headphones (Sony MDR-XB50AP). For data security reasons, the current website was hosted within a secure network (eduroam). The adaptive procedure follows the adaptive procedure by Brand and Kollmeier (2002) with fixed decreasing step sizes (± 10, 6, 5, 3, 2, 1, 1.5, 0.5). The smartphone and in-ear headphones were calibrated across frequencies (125–10,000 Hz) using the KEMAR artificial head. The resulting noise level was also checked to match the control session.

4.2.2.3 Interfaces for the MST & simplified MST To investigate how to best present the MST on a smartphone, we developed and compared four potential interfaces (Figure 4.2). The *matrix* interface served as a reference interface and corresponds to the traditional closed interface of the MST. Here, the complete matrix is displayed at once. The traditional matrix interface, however, does not comply with design guidelines (Gomez-Hernandez et al., 2023) when implemented on the small screen of a smartphone. Hence, three alternative interfaces were developed. The *slide* interface displays the columns of the matrix sequentially. The next column is presented as soon as a word is selected, which makes the interface faster. The buttons and the font are larger than with the matrix interface, but the words have to be provided in the given order (name-verb-number-adjective-

objective). The *type* interface requires users to type in the words they understood in any order. Suggestions of words from the complete 50 words of the matrix are provided based on the already typed input. As such, this interface serves as a mixture between an open and a closed version of a MST. The *wheel* interface displays the columns as wheels that can be scrolled horizontally. All words can be selected in any given order and fonts and buttons are larger than with the matrix *interface*. The larger font and button sizes of the alternative interfaces aim to reduce the potential impact of visual and fine motor declines. Further, the *wheel* and *type* interface allow users to simultaneously see all selected words, whereas the slide interface immediately shows the next column. The *type* and *wheel* interface further assess different input modalities, namely horizontal scrolling vs. typing. The order of presented interfaces was randomised. The simplified MST was measured only with the matrix interface, as we assumed that the screen size of smartphones is sufficient to display a 3 x 7 matrix of words.

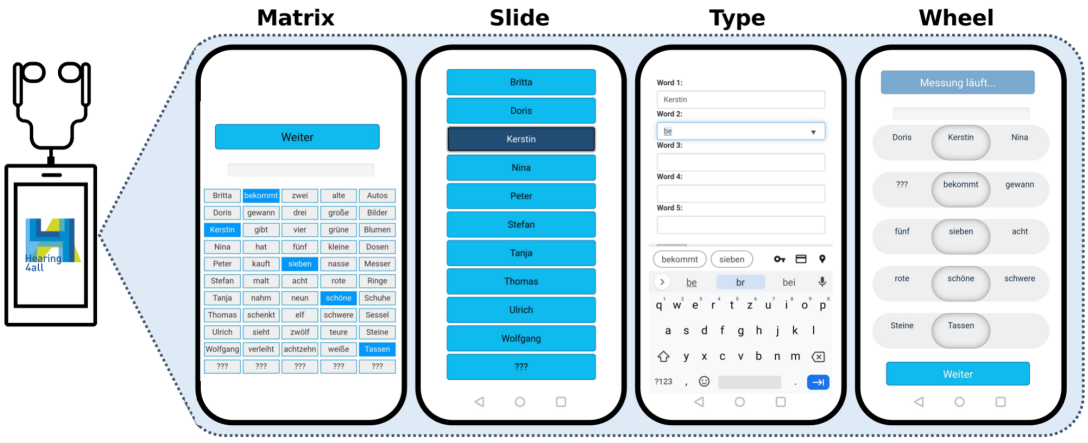

Figure 4.2: Matrix Sentence Test (MST) interfaces. The four interfaces under investigation. "Matrix" corresponds to the traditional closed MST UI. "Slide" presents the columns of the matrix sequentially. "Type" asks users to type in the words, while providing suggestions of words from the matrix. "Wheel" requires users to scroll the words horizontally.

4.2.2.4 MST and simplified MST conditions The MST was measured with ICRA1 (S0N0, S0N90 binaural & monaural) and ICRA5 noises, i.e. one stationary and one fluctuating noise signal (Dreschler et al., 2001). During the measurement, the noise level was fixed at 65 dB SPL and the speech level was adapted starting from an SNR of 0 dB. The SRT50 was measured, i.e. the threshold where 50% of the words are correctly understood. In the smartphone application, ICRA1 was measured across interfaces. ICRA1 S0N90 binaural & monaural, and ICRA5 were only measured with the slide interfaces to reduce

testing time. We chose the slide interface for this, as we expected all participants to be able to conduct the measurements with this interface, whereas the small fonts and buttons of the matrix interface could prove difficult for the oHI group. For the spatial conditions (S0N90 binaural & monaural) the noise direction (90° from left or right) was presented to the respective better ear according to PTA, even though differences between ears were small as participants were required to have asymmetrical hearing loss. For the monaural condition was only measured with the better ear. Binaural measures were included as they provide more realistic testing situations. They provide relevant information regarding binaural and spatial hearing that cannot be captured with the collocated S0N0 condition (Pastusiak et al., 2019). The simplified MST was also measured with ICRA1 and ICRA5 in S0N0. The MST test list consisted of 20 sentences, whereas the simplified MST was measured with 14 sentences. For a tabular overview of all conditions see column "Condition" of Table 4.2.

4.2.2.5 Questionnaire data Within the smartphone application, participants answered several short questions regarding age, gender, and general smartphone usage. In addition, users could provide open comments regarding the different interfaces. Finally, after completion of both the laboratory control and smartphone test sessions, participants were asked to rank the interfaces from highest to lowest preference, and could provide verbal comments regarding the different interfaces, which were noted by the experimenter.

4.2.3 Procedure

First, the pure-tone audiogram (500, 1000, 2000, 4000 Hz) was measured with OMA to control for the correct group allocation (yNH, oHI). Second, two training runs were conducted with the MST to counter potential training effects. The training lists were measured with the closed OMA version to familiarise the participants with the speech material. The first list was measured in quiet, the second list in the collocated S0N0 condition with test-specific noise (65 dB SPL). Next, participants were randomly allocated to either start with the laboratory control session (LAB), or with the smartphone test session (WEB). Within both sessions, the different conditions were measured in random order. The only exception was that the S0N0 ICRA1 slide condition was always measured prior to the remaining slide conditions. This was to ensure that the first encounter with each interface was measured in the same noise and spatial condition. As a final step, participants were asked to rank the interface according to their preference. They also could provide verbal comments to the experimenter regarding the interfaces.

4.2.4 Analyses

4.2.4.1 SRT consistency To test our hypothesis **(RQ1)** that both yNH and oHI can self-conduct the matrix and simplified matrix test, we compared paired SRT scores of participants for the smartphone test session to the laboratory control session. For this, we calculated both the root mean square error (RMSE) and bias for the different interface conditions (matrix, slide, type, wheel), spatial conditions (S0N90 monaural & binaural), and noise (ICRA1, ICRA5) conditions. The RMSE describes the general error between the WEB and LAB results, whereas the bias indicates whether a trend towards over- or underestimation of the SRTs exists (positive and negative bias, respectively). A high positive bias, for instance, could reveal a general trend towards overestimating SRTs with a specific interface. For the interfaces, we further tested for significant differences to the laboratory control session ($p<0.05$). To determine which interface is best in terms of SRT consistency **(RQ2a)**, we tested for significant differences of the interfaces compared to the control condition across the two groups (yNH, oHI). To assess potential interface effects on SRT consistency, we further tested for interaction effects. For this, we used a linear mixed model nested within participants, with the following formula:

$$SRT = intercept(LAB) + UI + GROUP + UI * GROUP + (1|Participant)$$

Where *SRT* is the response variable, *intercept* refers to the control condition, *UI* refers to the specific user interface (matrix, slide, wheel, type), *GROUP* refers to either the yNH or the oHI group, *UI*GROUP* tests for an interaction effect, and $(1|Participant)$ refers to the design of the model being nested within subjects.

4.2.4.2 Completion time We analysed the completion time of the four interfaces across the yNH and oHI groups via boxplots. This was done to find out which interface is fastest **(RQ2b)**. We further tested for significant differences in completion time across interfaces, separately for both groups, and for general differences between yNH and oHI, using the paired Wilcoxon signed-rank test.

4.2.4.3 User preferences To evaluate user preferences, we analysed the provided rankings of the participants. To that end, we calculated the preference ratios for the interfaces to determine the overall preferred interface. That means, we calculated how often (in percent) a given interface was ranked higher than each of the remaining interfaces, for both the yNH and oHI group. As a result, we could compare pairwise preference differences between the two groups for each

interface. Next, we analysed the verbal or written comment data using conceptional concept analysis (Hsieh and Shannon, 2005). For each interface, written comments obtained by the app and verbal comments of the participants were combined. They were then categorised into different concepts in an inductive manner. This means, concepts were generated during the coding, and updated if new concepts were detected in the comments. To ensure all contained concepts were covered, this process was reiterated several times. After a final set of concepts was determined, their occurrence across interfaces was counted. Concept analysis was performed in German, as participants provided their comments also in German. The concepts and examples were then translated to English via DeepL for an unbiased and standardised translation (see Table 4.3 for some examples and Supplementary Table S.1 for all concepts with explanations).

4.3 Results

4.3.1 SRT consistency

4.3.1.1 MST interfaces Figure 4.3 depicts the comparison of speech recognition thresholds (SRT) for the laboratory control session (LAB) and the smartphone test session (WEB). The yNH group was able to accurately measure the MST with a smartphone, as well as with all interfaces. The differences of the interfaces to the control session are not significant ($p > 0.05$, see Table 4.1) and the RMSE are around the test-retest error of 1 dB. For the type and wheel interfaces, the SRTs are slightly more biased towards higher SRTs than for the matrix (lowest) and the slide interface (Table 4.2). That means the wheel and type interfaces show a trend to overestimate SRTs.

The oHI group had the highest SRT consistency with the matrix interface, as indicated by (1) the low RMSE and bias scores (Table 4.2), and (2), the UI coefficient of the linear mixed model (Table 4.1). However, we excluded the visible outlier of the matrix interface (SRT = 5/-5 dB SNR) from the statistical analyses, as it would have strongly biased the analyses. This participant highlights that while most users would achieve accurate SRT scores with the matrix interface, some may not be capable of performing speech-in-noise tests on a smartphone with the matrix interface. A mobile implementation suitable for MST measurements should, however, be robust enough to avoid such outliers. For the other interfaces RMSE scores are higher, but also no extreme outliers are found. The differences between wheel and slide are marginal. While the wheel interface has slightly lower RMSE scores, it also results in slightly higher bias

scores and UI coefficients. The type interface consistently had the highest SRT elevation, as evident from Figure 4.3, the RMSE results, and biases.

Overall, the matrix interface is most accurate for both the yNH and oHI, with the exception of the outlier. We can observe an effect of the oHI group on the performance with the different interfaces. The lowest interface-specific effects can be observed with the wheel interface, indicating that the wheel interface was most stable across the two groups (yNH and oHI).

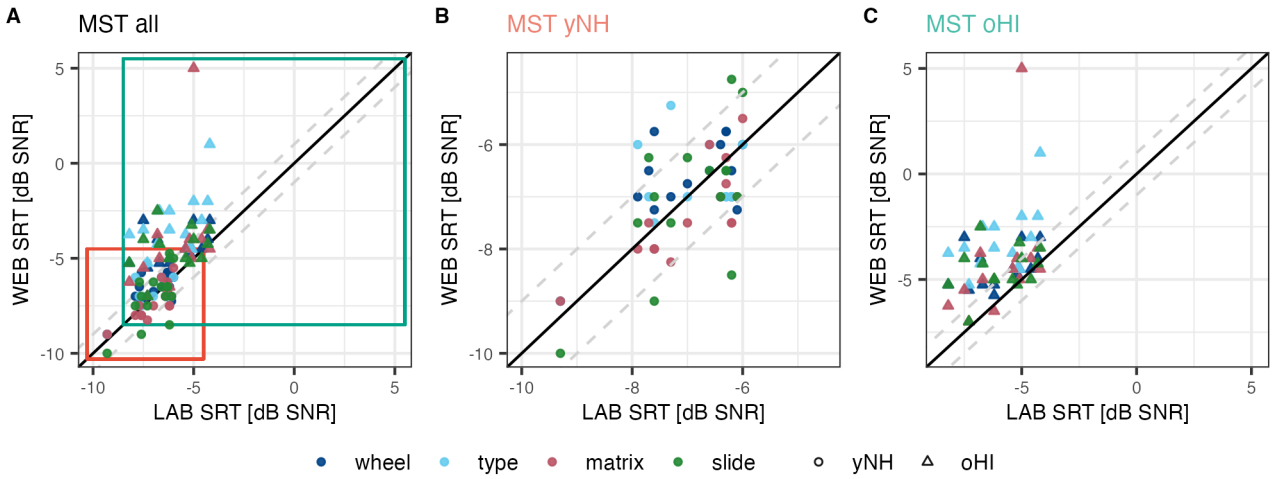

Figure 4.3: Comparison between SRTs obtained with the smartphone web-based interface (WEB, ordinate) vs. the reference laboratory implementation using OMA (LAB, abscissa). (A) SRT results for the younger normal hearing (yNH) group (circles) and the older hearing impaired (oHI) group (triangles) across all four interfaces for ICRA1 in the S0N0 session. The dashed line corresponds to a 1 dB deviation from the line of perfect agreement. Data points within the dashed lines are within the expected test-retest error. (B) SRT results for the yNH group (zoom from small box in A) (C) SRT results for the oHI group (zoom from the large box in A).

4.3.1.2 MST spatial conditions and fluctuating noise (slide interface)

The ICRA1 S0N90 binaural condition resulted in adequate test-retest values for the yNH group, while the SRTs with the ICRA1 S0N90 monaural condition yield an underestimation for the smartphone web-based version (Figure 4.4), most likely due to interaural crosstalk effects (see discussion). For the oHI group (N = 11 due to missing data) a slight bias can be observed in both conditions (see Table 4.2), similar to the bias observed with ICRA1 S0N0, but adequate consistencies between smartphone and laboratory UI are achieved. For the fluctuating noise condition ICRA5, we can, again, observe a bias similar to other slide interface conditions. The bias and RMSE are higher for the ICRA5 condition. Since the expected test-retest variability with ICRA5 is generally larger (2 dB for NH

Table 4.1: Main and interaction effects of two linear mixed models for UI on SRT score with either yNH or oHI as the reference category for the group $SRT = intercept(LAB) + UI + GROUP + UI * GROUP + (1|Participant)$. Interaction effects are only shown for the yNH reference model; for the oHI reference model the direction of the UI coefficients would be reversed. Laboratory control session (LAB) serves as the control (intercept). UI and group (yNH/oHI) are fixed effects nested within subjects. Bold values indicate that no significant difference to the intercept exists.

	Coefficients		Estimate	Conf. Interval	P-value
yNH	intercept	LAB	-6.967	(-7.515,-6.418)	0.0000
	UI	UImatrix	**-0.350**	(-0.899,0.199)	**0.2273**
		UIslide	**-0.050**	(-0.796,0.696)	**0.8626**
		UItype	**0.050**	(-0.499,0.599)	**0.8626**
		UIwheel	**0.300**	(-0.249,0.849)	**0.3003**
oHI	intercept	LAB	-5.977	(-6.612,-5.201)	0.0000
	UI	UImatrix	1.015	(0.885,2.429)	0.0014
		UIslide	1.419	(0.617,2.161)	0.0000
		UItype	2.746	(1.992,3.536)	0.0000
		UIwheel	1.419	(0.689,2.233)	0.0005
	GROUP	oHI	0.990	(0.184,1.795)	0.0264
	UI*GROUP	matrix*oHI	1.365	(0.560,2.171)	0.0017
		slide*oHI	1.469	(0.664,2.275)	0.0007
		type*oHI	2.696	(1.890,3.502)	0.0000
		wheel*oHI	1.119	(0.314,1.925)	0.0094

listeners, (Wagener, 2004)), this effect is in line with expectations.

4.3.1.3 Simplified MST Due to the higher test-retest variability of the simplified MST (Wagener K, 2005), the comparison between smartphone UI vs. laboratory control UI session is expected to result in larger RMSE values than for the MST. Since the S0N0 ICRA5 condition was only measured using the slide interface, only the results for the respective slide interface conditions are displayed in Figure 4.5 for the simplified MST and MST. For both the yNH and oHI groups the simplified MST resulted in higher RMSE (as expected) and bias than the MST with the matrix interface (Table 4.2). For ICRA5, we need to compare the simplified MST matrix interface to the MST slide interface, as ICRA5 was only measured with the slide interface for the MST. We again observe higher RMSE and bias values for simplified MST with the yNH group. In contrast, for the oHI group, we might observe an interface effect, as RMSE and bias scores are slightly lower for simplified MST. However, this may be caused by the comparison of the slide and matrix interface, where the slide interface generally resulted in higher scores.

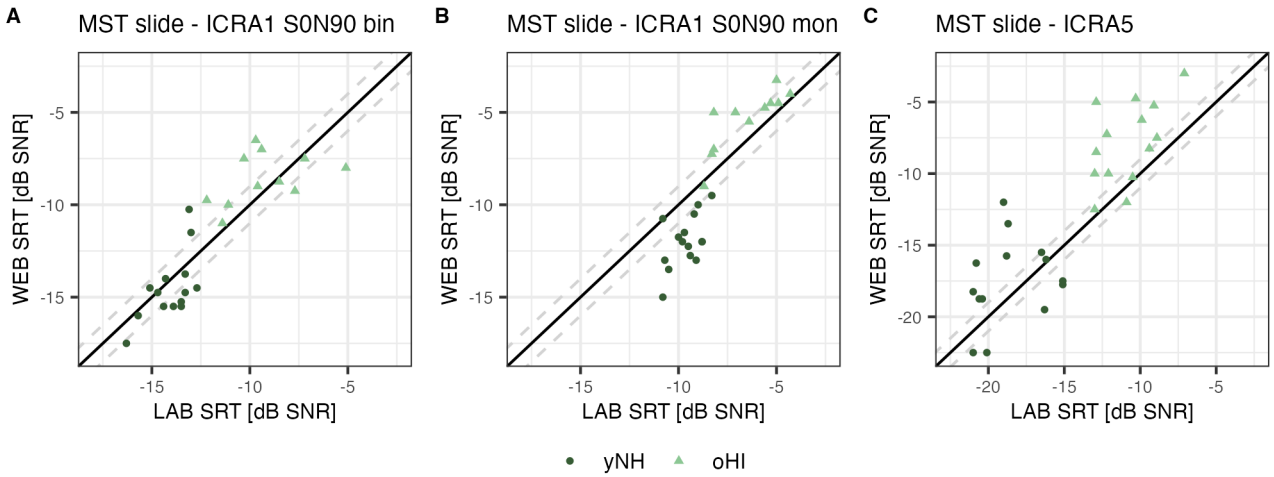

Figure 4.4: SRT results for slide condition. Dashed lines indicate expected test-retest variability (1 dB for ICRA1; 2 dB for ICRA5 (NH)).

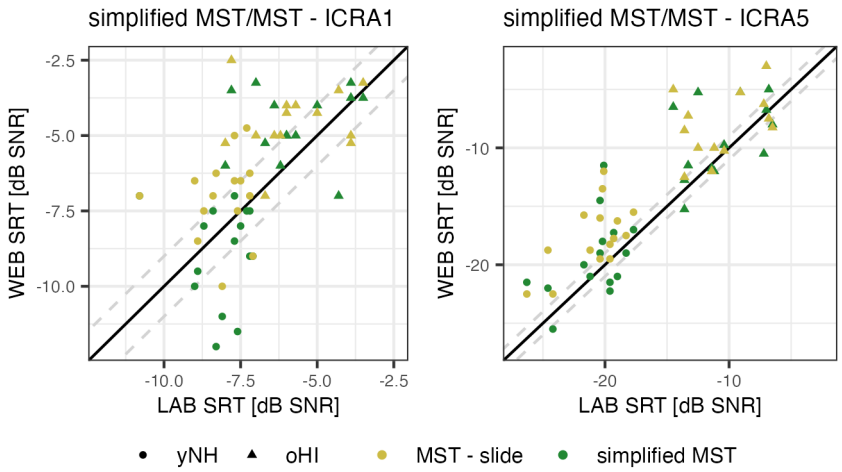

Figure 4.5: SRT results for the simplified MST and comparison to the MST slide interface for both the smartphone test session and the laboratory control session. Dashed lines indicate test-retest variability (1 dB for ICRA1; 2 dB for ICRA5).

4.3.2 Completion time

Figure 4.6 compares the absolute completion times for the different MST interfaces, as well as the simplified MST to infer completion times for practical implementation. The fastest test was the simplified MST. This was expected, as the simplified MST is a reduced version of the MST with only three words per sentence and 14 sentences in a test list. The matrix interface ranks second but is only marginally faster than the slide interface and this difference is not statistically significant. The type interface took the longest time for both groups. The yNH group was generally faster in completing the test with all interfaces, as compared to the oHI group ($p < 0.01$). This difference appears stable across all interfaces but the type interface. With the type interface, there appears to be

Table 4.2: Condition overview, RMSE and bias for smartphone (Web SRT) vs. laboratory tests (LAB SRT) for all conditions across all participants, and separately for the yNH and oHI group.

Condition				RMSE			Bias		
Test	**Spatial condition**	**Noise**	**Interface**	**All**	**yNH**	**oHI**	**All**	**yNH**	**oHI**
Matrix sentence test (MST)	S0N0	ICRA1	matrix	1.07	0.68	1.335	0.28	-0.35	0.95
			slide	1.54	1.01	1.96	0.64	-0.05	1.39
			type	2.24	0.91	3.08	1.36	0.05	2.76
			wheel	1.43	0.77	1.89	0.86	0.30	1.46
		ICRA5	slide	3.48	3.18	3.78	1.79	0.94	2.98
	S0N90 binaural	ICRA1	slide	1.57	1.38	1.74	0.20	-0.49	0.52
	S0N90 monaural	ICRA1	slide	2.07	2.60	1.28	-0.63	-2.34	0.81
Simplified MST	S0N0	ICRA1	simplified MST	2.02	2.05	2.00	-0.14	-0.77	1.12
		ICRA5	simplified MST	3.35	3.36	3.33	1.66	1.44	1.49

an interaction effect between the interface and the oHI group. That is, the oHI group took much longer in completing the test with this interface as compared to the remaining interfaces.

4.3.3 Interface ranking

Figure 4.7 compares the interface rankings between oHI and yNH as expressed in the ratio how often an interface (color) was preferred over another interface (shape). The line indicates perfect agreement between yNH and oHI with respect to the ranking of a given interface. Symbols in the upper triangle indicate that oHI preferred this interface over another interface to a greater extent than yNH. Higher values of the preference ratio indicate a higher percentage of cases where the respective interface was preferred over the other interfaces. Overall, between yNH and oHI only slight preference differences exist. The results indicate that on average the matrix interface (red symbols) was mostly preferred by both yNH and oHI, especially since the matrix UI was preferred in the direct comparison between slide and matrix UI by both groups. However, the matrix interface was generally more strongly preferred by the yNH. This becomes most visible by the matrix-wheel comparison (red square). The yNH group preferred the matrix interface over the wheel interface to a greater extent (red square). The resulting higher preference of the oHI for the wheel interface also transfers to the comparison with the type interface. Here, we can observe a slight preference

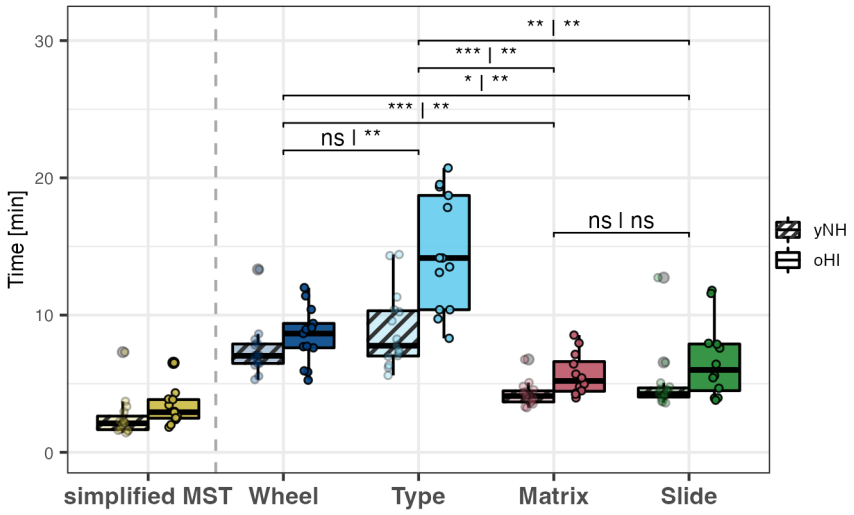

Figure 4.6: Completion time of the MST (20 sentences) measured with ICRA1 S0N0 for each interface, as well as simplified MST (14 sentences), separate for yNH and oHI individuals. Statistically significant differences are indicated by an asterisk ($* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$) for both groups (yNH | oHI) and nonsignificant differences by ns.

of the oHI for the wheel interface, and for the yNH for the type interface (blue circle, turquoise square).

The second highest ranking was achieved by the slide interface (green symbols). In comparison to the matrix interface, it is only slightly less preferred (green triangle, red diamond) while the yNH group showed a stronger relative preference for the matrix UI than the OHI group. Further, the slide UI consistently ranked higher than the type and wheel interface. Noteworthy is that for the oHI group the slide interface had a higher ranking than the matrix interface in comparison to the wheel interface (green square vs. red square).

4.3.4 Interface preferences

To investigate how participants perceived the different UIs of the MST on a smartphone, we analyzed the comments participants provided using concept analysis. Figure 4.8 visualizes the frequency of reported concepts across interfaces for both yNH and oHI. Examples for concepts, their explanation, and source comments can be found in Table 4.3. For a complete description of all concepts see Supplementary Table S.1. The results are clustered to highlight the relevant concepts across interfaces. The concepts are, thus, in different order for yNH and oHI.

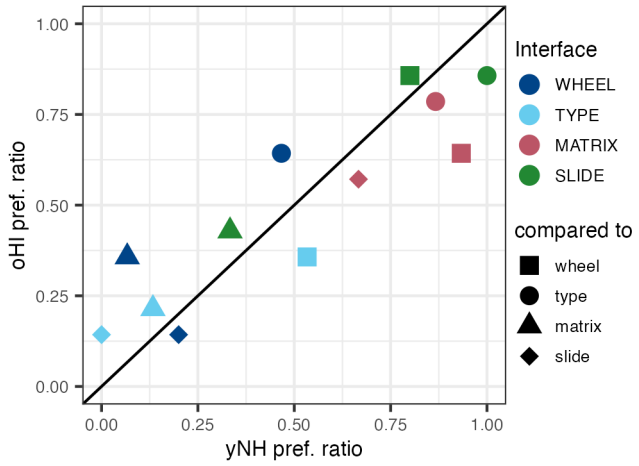

Figure 4.7: yNH vs. oHI preference ratios of the different interfaces. The ratio indicates how often an interface (color) was preferred over another interface (shape).

The analysis of the written and verbal comments of the participants revealed only slight differences across the two groups (yNH, oHI). The yNH provided more comments (mean = 7.13) than the oHI group (3.64). However, every participant provided at least one comment.

The type interface mainly received negative comments. For instance, it was reported by both yNH and oHI that the interface was cumbersome and difficult to handle. Similarly to the interface ranking, the wheel interface received rather negative comments by the yNH. The oHI noted that the wheel interface had a good size and was easy to use, but also that it was rather slow. The yNH also noted that it was slow, and indicated that it was annoying and cumbersome, among others. The oHI further noted that that they forgot the words while providing their response during a trial with the type and wheel interface.

The yNH group rated the matrix interface mostly positive. They perceived it as clear and easy, and noted positively that corrections were possible with this interface. For the slide interface, they instead negatively highlighted that no correction was possible, and errors occurred due to choosing a non-intended button. The oHI, in contrast, noted that the slide interface was easy and had good sizing. Though to a smaller amount, the oHI also pointed out that no correction was possible. In contrast to the slide interface, the oHI commented on the small size of the matrix interface.

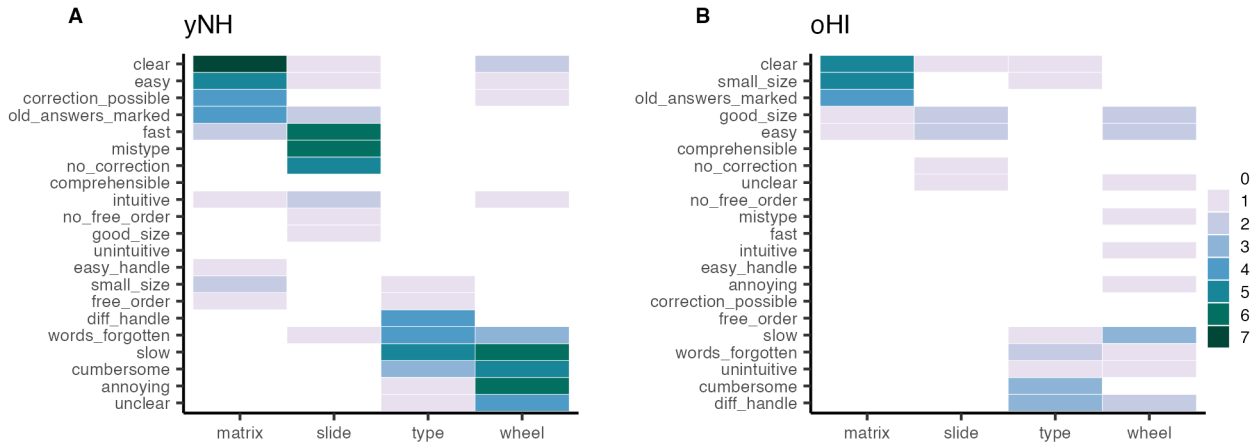

Figure 4.8: Prevalence of different concepts in the written and verbal comments of the participants. Concepts are clustered for visualization purposes. (A) displays the results for the yNH group; (B) the results for the oHI group.

4.4 Discussion

The present study proved the feasibility of measuring the German Matrix sentence test (MST) using a smartphone with household in-ear headphones for both yNH and oHI. We tested appropriate interfaces for a mobile implementation of the MST and compared them with respect to the following usability aspects: SRT consistency, required time, and user preferences. Finally, we discuss group effects (yNH vs. oHI) on interface usability and propose an interface for a mobile implementation.

4.4.1 Group effects on the performance of the different interfaces

4.4.1.1 General feasibility of smartphone-based MST measurements

The different interfaces resulted in different SRTs, completion times, and user preferences. The yNH results confirm that the MST could be tested in a valid way with all four interfaces. This means, there are no interface-specific boundaries that hinder the assessment. The oHI results, conversely, demonstrates that potential age-dependent declines and/or unfamiliarity with smartphones can affect the feasibility of smartphone-based measurements.

Generally, we observed an increase in RMSEs with the oHI group. For all interfaces this increase was significant ($p < 0.05$). The increase for the matrix, slide, and wheel interface was, however, rather small. We therefore conclude that it is generally feasible to conduct the MST on a smartphone with inexpensive headphones, but there may be a slight SRT elevation with elderly hearing im-

Table 4.3: Exemplary concept names and explanations are shown, next to examples from provided comments (translated from German to English via DeepL for an unbiased and standardized translation). Underlined words highlight what caused a comment to be counted in the respective concept.

Concept	Concept explanation	Example comment parts
easy	Performing the tasks with this interface was easy or simple	- Super easy to use. Self-explanatory
Mistype	Mistyping occurred / wrong selection of a word	- similar, bad: quickly move on if mistyped; if last word tapped: immediately move on to next sentence - I couldn't correct if I had clicked the wrong way.
no_correction	It was not possible to correct one's input	- In case of mistyping, is it possible to come back? - I couldn't correct if I had clicked the wrong way.
Slow	It took time to complete the task with this interface, it was not fast	- time-intensive - It took quite a bit of time to write it out, so I wasn't sure about the last few words.
words_forgotten	Sentence sound lost / Words forgotten / You had to remember the words for too long	- As I said, somewhat cumbersome and sometimes annoying, because you did not know in which direction the searched word was and so forgot the other words. An overview of the words available for selection would have been nice.

paired individuals **(RQ1)**. Further, we confirm that an interface effect is present within the oHI group (also evident from the interaction effects). In other words, the interfaces influence the consistency of the measurement for older adults with hearing loss **(RQ2)**. This highlights the necessity for including elderly in the development of smartphone-based MSTs and mobile health applications. Otherwise, applications may provide inaccurate results, or may not be easily usable and lead to frustration by the users (Kalimullah and Sushmitha, 2017).

4.4.1.2 SRT consistency and preference The matrix interface resulted in the most consistent SRT scores, was the fastest MST interface, and was overall preferred both by yNH and oHI. For the yNH this could be expected,

as the complete overview of the matrix is provided, the handling is fast, and smaller buttons are not expected to be an issue. Conversely, it is surprising that the matrix interface was mostly positively received by the oHI and generally provided the most consistent results. Regardless, we did observe one participant (outlier) to be unable to achieve consistent results with the matrix interface and later noting the small size of the interface.

It could have been expected that the comparatively small button size and spacing would lead to larger difficulties with more participants of the oHI group, due to potentially reduced fine motor skills in the oHI group. However, several explanations are possible for the general SRT consistency and positive perception of the oHI group with the matrix interface.

First, the cognitive complexity of the matrix interface is smaller than with, for instance, the wheel and type interface. The Cognitive Complexity Theory (Kieras and Polson, 1985) in the context of UIs can be described as the number of different production rules to be learned. Production rules, in turn, are defined as IF (interface output) / THEN (user response) statements (Ziefle and Bay, 2005). To exemplify, following the playback of a sentence (IF), the user needs to perform different steps for task completion with the different interfaces. With the matrix interface, word buttons need to be selected (THEN). (IF) all buttons are selected, the next button needs to be selected for the next sentence to be played (THEN). With the slideshow interface, for instance, the last production rule is omitted, as the next sentence is played automatically after the last word is selected. The type interface, in contrast, requires further production rules (e.g. (IF) a keyboard appears, words need to be typed (THEN) and (IF) a word is suggested by the device, it can be selected (THEN)). The overall reduced cognitive complexity of the matrix interface may therefore have led, to a certain extent, to higher preferences of the oHI.

Second, a related explanation could be due to the enhanced error probability with increasing steps, or production rules. In that way, it could explain the higher SRT consistency of the matrix interface for most of the participants and the comparatively lower SRT consistency of the type interface. The probability of success can be modeled with: $p(success) = (1 - p(failure))^{N(steps)}$ (Fisk et al., 2014). It follows, that the increasing number of steps with the type interface may have adversely affected the accuracy of the measurement.

Finally, fine motor skill declines may not be as pronounced in the present oHI group. Consequently, except for one participant, the oHI group may not have experienced that many difficulties in performing the task with the matrix interface, even though buttons and button spaces were comparatively small. Nonetheless, three participants had to be excluded as they were not capable of performing the tasks on a smartphone sufficiently without a pen. However, we have not assessed fine motor skills systematically. For future studies, it would be of interest to capture the exact impact of fine motor skills on UI preferences.

A preference difference between yNH and oHI individuals can be observed with the wheel interface **(RQ2b)**. The wheel preference ratio was higher for oHI than yNH individuals in comparison to the type and matrix interface. For the slide interface there was no pronounced difference between the two groups. Considering that the matrix and type interface both displayed words in smaller font sizes and smaller button sizes than the wheel interface, this suggests a preference trend for larger font and button sizes with the oHI group.

4.4.1.3 Completion time Besides SRT consistency and user preference, it is important for mobile measurements to be as fast as possible, whilst ensuring accurate results. Generally, users prefer faster measurements, and one might risk higher levels of inattentiveness and measurement abortion with longer measurement duration (Möckel et al., 2015). Longer completion times may also be attributed to the higher cognitive complexity of the interfaces. Specifically, if more production rules need to be learned, the completion time can increase, thereby hindering the interface's effectiveness. Additionally, longer completion times may negatively impact the consistency of SRT results. To exemplify, both groups reported forgetting words while inputting results using the wheel and type interface.

For the MST, the matrix and slide interface were the fastest. Regardless, the oHI group took longer to perform the speech tests across all interfaces. This is in line with research indicating that the cognitive decline with age and smartphone unfamiliarity leads to slower response time of elderly (Tsai et al., 2017). The reason for this is, however, that elderly choose their response more carefully to avoid errors. This change in response behavior can yield results as accurate as those obtained by younger users, compensating for any unfamiliarity with the devices (Starns and Ratcliff, 2010). This can be observed with the matrix

interface, where SRT consistency was similar across groups, but the oHI took longer to complete the task.

The most striking group time difference can be observed with the type interface. Here, the time prolongation for completion of the oHI is larger than what would be expected from the observed time increases from the remaining interfaces. It is, thus, clear that the time prolongation goes beyond general potential speed reductions of the elderly, and instead demonstrates a group effect of this interface **(RQ2b)**. An explanation could be that younger individuals are more used to writing text messages with their smartphones, than the elderly. Additionally, the tactile demand was highest in this interface, and therefore, the impairment by possible fine motor skill declines may be most noticeable here. Since the type interface took the longest, provided the least accurate SRT results, and was generally disliked by participants it can clearly be ruled out as a preferred interface.

4.4.1.4 Performance with spatial and fluctuating noise conditions Interestingly, we observed a bias in the ICRA5 condition with the slide interface, which we did not observe with the ICRA5 condition of the simplified matrix test simplified MST. Similarly, a slightly higher bias is present for the slide interface as compared to the matrix interface for the MST with ICRA1 (oHI). Consequently, the bias appears to stem specifically from the slide interface. We infer that higher accuracies would be achieved, if the slide interface would be adapted to address potential error sources resulting in the bias (see Proposal of an adapted interface), or the matrix interface would be used. It also becomes clear that the spatial and fluctuating noise conditions lead to a better SRT separability of the yNH and oHI group, in comparison to the S0N0 condition.

4.4.1.5 Performance with the simplified MST The simplified MST resulted in adequate consistencies, however, was not as consistent as the MST for the yNH group. This can be expected, as it is a reduced MST, both in trials (14 vs. 20) and in sentence length (3 vs. 5). Test-retest variability is also generally higher with simplified MSTs, than with MST. For instance, the test-retest value for the SRT50 with adults for the Italian MST and the simplified Italian MST are 0.6 dB and 1.2 dB, respectively (Puglisi et al., 2021). In our experiment, the average RMSE for yNH and oHI is 1.9 dB. To a certain extent, the RMSE value can therefore be explained by the general test-retest error. As expected, the simplified MST had the shortest absolute

completion time for both yNH and oHI, as it uses fewer trials and shorter sentences compared to the MST. For an elderly target group with potential cognitive declines, the simplified MST can serve as a good alternative to the MST, if (1) none of the other interfaces seem plausible, and (2) the SRT consistency achieved with the simplified MST is considered as being sufficiently high.

4.4.2 Potential crosstalk effect with low-cost consumer electronics

Our study showed that crosstalk between audio presentation channels is an issue with non-professional audio equipment for conditions with a substantial interaural dissimilarity. This crosstalk can be due to "spatial audio processing" of the consumer electronic audio device, electric/electronic channel crosstalk, or due to an acoustic crosstalk path at the headphone that all can be avoided under controlled laboratory conditions, but not for arbitrary consumer electronic devices. In our study, the monaural conditions with the MST (see Fig. 4.4, middle panel) appear to be affected by this effect: If the signal (intended to be played only to one ear) leaks over to the respective other headphone channel, normal hearing listeners can utilize the resulting binaural information which results in lowering the SRT in comparison to the reference value. While this effect can be seen for the monaural S0N90 condition for the yNH group, it is not evident for the oHI group. This could be because the crosstalk may be below the (increased) threshold levels in the oHI group. For the monaural condition users with in-ear headphones could be instructed to remove the earplug for the ear to which no stimulus is presented (ear towards the noise direction). Alternatively, for over- and on-ear headphones, a crosstalk effect could be avoided by only measuring the spatial condition following an indication of hearing impairment via the S0N0 condition.

4.4.3 Proposal of an adapted interface

Not every participant could perform the task with the matrix interface, even though the interface resulted in accurate SRT scores for most participants. The oHI group also noted the small size of the words on the screen with this interface. This may be an issue with future participants and users with larger tactile difficulties than the participants employed in our study. Hence, another interface may be the preferable choice in the future. The slide and wheel interface both achieved adequate SRT consistencies. The slide interface, however, ranked higher in preference and was also faster to complete. In addition, one important

outcome of the content analysis was that participants reported having accidentally pressed the wrong button, without the possibility to correct their mistake. Allowing participants to correct their mistake and to return to earlier parts of the sentence could decrease accidental error sources when performing the MST. As a result, the SRT deviations from the laboratory control session would be reduced. Additionally, the slide interface allows for an even further increase in font sizes. That way, a fast, consistent, and intuitive interface would be provided that avoids the small size when displaying the complete matrix. Figure 4.9 shows the proposed adapted slide interface. We would therefore argue for the adapted slide interface to be implemented for mobile implementations of the MST **(RQ3)**.

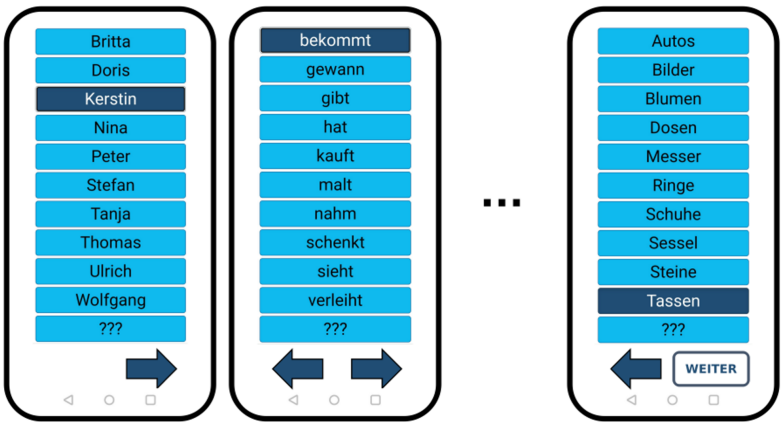

Figure 4.9: Proposed interface for performing the MST via a smartphone. Arrows allow to switch between the different "slides" or columns of the matrix. After a user has selected the understood words, pressing "Next" ("weiter" in German) will start the next sentence.

4.4.4 Limitations and future research

Even though the SRT is rather independent from the absolute presentation level and headphone transfer function, it is possible that some potential variation in SRT occurred due to loosening of the in-ear headphones. To estimate the potential effect for everyday life, investigations into the robustness of different headphone transfer functions (in-ear, on-ear, over-ear) against such mechanical variations and the resulting impact on the measured SRT would be of interest.

Furthermore, the tests were performed with calibrated equipment in a laboratory environment. It would be of interest to investigate how well the system would work in uncontrolled environments, with different uncalibrated setups (smartphones & headphones). The interfaces would not be expected to lead to different results in uncalibrated setups, but household headphones often have different

frequency responses that boost certain frequencies and produce different overall sound levels.

Further, low-cost consumer electronics, as shown in the present study can result in a crosstalk effect, which can impede measuring spatial conditions (see section "Potential crosstalk effect with low-cost consumer electronics"). In addition, additional background noise present in the testing environment may have an influence on measurement results. For that, it could be valuable to integrate a background noise monitoring system, as in Hussein et al. (2016). However, different sound levels are not expected to affect measurement acuity much, as the results of the matrix sentence tests (MST) are based on the SNR rather than on the absolute sound level and spectrum presented. Thus, it is less crucial to perform the measurements at the exact same noise level, and more important that the noise level is at least 20 dB higher than the PTA (Wardenga et al., 2018).

Contrary to speech-in-noise tests that provide higher test-retest variations in their results for a given amount of measurement time and hence cannot clearly detect a training effect, the matrix test exhibits a training effect of about 2 dB (Wagener, 2004). For clinical audiology purposes, training with two test lists of the MST is therefore recommended. In our experiments, the experimenter conducted such a training session with a closed-set test version prior to the control and test session. Participants were, consequently, already familiar with the matrix user interface when they did the measurements with the distinct interfaces. Being presented with the matrix interface may have been less surprising than being presented with the newly-proposed interfaces. This could have increased the preference for the matrix interface among the participants, but at the same time it highlights the good performance of the slide interface in comparison. Displaying the interfaces to untrained participants was not possible though, as the repeated measurements with the distinct interfaces would then have resulted in a training effect. For later implementation of the MST for an openly available web application where usually no training sessions can be performed, it would be important to control for the ongoing training effect, e.g., by monitoring the stability of the ongoing SRT estimate in real time.

Finally, while we aimed to include 15 individuals in both the yNH and the oHI group, we could only include 14 oHI participants. Three additional oHI individuals were unable to complete the MST on a smartphone without special equipment such as a pen. As a result, the sample size is slightly smaller than initially esti-

mated. Future research should, therefore, aim to include larger sample sizes for further investigations into smartphone-based MSTs. Here, our developed mobile version of the MST, which enables remote testing via smartphones, can facilitate large-scale sample collection.

4.5 Conclusions

Our study proves the general feasibility of measuring the matrix sentence test (MST) and the simplified MST via a smartphone and in-ear headphones with both yNH and oHI individuals. For the four distinct interfaces overall group effects exist, next to group effects specific for distinct interfaces. The SRTs of the matrix interface were most consistent with the laboratory control for most of the participants, but not every participant could perform the tasks with this interface and the small size of the interface was noted. It is therefore less suitable for an elderly target group with potential tactile and visual impairments. The slide interface ranked second in terms of SRT consistency, preference, and completion time. At the same time, the participants provided suggestions for the slide interface (i.e., demanding the possibility of correcting the word selection), which could improve SRT consistency and stability. Therefore, the slide interface might eventually turn out to be a better alternative. The findings of this study should be applicable to the MST in all available languages by translating the interfaces to the respective language.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy – EXC 2177/1 – Project ID 390895286.

Bibliography

Allen, N., Sudlow, C., Downey, P., Peakman, T., Danesh, J., Elliott, P., Gallacher, J., Green, J., Matthews, P., Pell, J., et al. (2012). Uk biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, 1(3):123–126.

Borella, E., Ghisletta, P., and de Ribaupierre, A. (2011). Age differences in text processing: The role of working memory, inhibition, and processing speed. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 66(3):311–320.

Bowden, J. L. and McNulty, P. A. (2013). The magnitude and rate of reduction in strength, dexterity and sensation in the human hand vary with ageing. *Experimental gerontology*, 48(8):756–765.

Brand, T. and Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *The Journal of the Acoustical Society of America*, 111(6):2801–2810.

Bruns, T., Ooster, J., Stennes, M., and Rennies, J. (2022). Automated speech audiometry for integrated voice over internet protocol communication services. *American Journal of Audiology*, 31(3S):980–992.

Buschermöhle, M., Wagener, K., and Kollmeier, B. (2016). Sprachaudiometrische messungen mit dem verkürzten oldenburger satztest olkisa bei erwachsenen (speech audiometric measurements with the simplified oldenburg sentence test olkisa with adults). *Z. Audiologie*, 55:6–13.

Chun, Y. J. and Patterson, P. E. (2012). A usability gap between older adults and younger adults on interface design of an internet-based telemedicine system. *Work*, 41(Supplement 1):349–352.

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (2001). Icra noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment: Ruidos icra: Señates de ruido artificial con espectro similar al habla y propiedades temporales para pruebas de instrumentos auditivos. *Audiology*, 40(3):148–157.

Farage, M. A., Miller, K. W., Ajayi, F., and Hutchins, D. (2012). Design principles to accommodate older adults. *Global journal of health science*, 4(2):2.

Fisk, A. D., Czaja, S. J., Rogers, W. A., Charness, N., and Sharit, J. (2014). *Designing for older adults: Principles and creative human factors approaches*. CRC press.

Frishammar, J., Essén, A., Bergström, F., and Ekman, T. (2023). Digital health platforms for the elderly? key adoption and usage barriers and ways to address them. *Technological Forecasting and Social Change*, 189:122319.

gGmbH, H. O. (n.d.). Simplified matrix. https://www.hz-ol.de/de/diagnostik-simplified-matrix.html. Accessed: 2024-01-26.

Gomez-Hernandez, M., Ferre, X., Moral, C., and Villalba-Mora, E. (2023). Design guidelines of mobile apps for older adults: systematic review and thematic analysis. *JMIR mHealth and uHealth*, 11:e43186.

Green, P. and MacLeod, C. J. (2016). Simr: An r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4):493–498.

Hsieh, H.-F. and Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, 15(9):1277–1288.

Hussein, Y. S., Wet Swanepoel, D., Biagio de Jager, L., Myburgh, H. C., Eikelboom, R. H., and Hugo, J. (2016). Smartphone hearing screening in mhealth assisted community-based primary care. *Journal of telemedicine and telecare*, 22(7):405–412.

Hwangbo, H., Yoon, S. H., Jin, B. S., Han, Y. S., and Ji, Y. G. (2013). A study of pointing performance of elderly users on smartphones. *International Journal of Human-Computer Interaction*, 29(9):604–618.

Hörzentrum Oldenburg gGmbH (n.d.). Internationale matrixtests. https://www.hz-ol.de/de/diagnostik-matrix.html. Accessed: 2024-01-26.

Kalimullah, K. and Sushmitha, D. (2017). Influence of design elements in mobile applications on user experience of elderly people. *Procedia computer science*, 113:352–359.

Kieras, D. and Polson, P. G. (1985). An approach to the formal analysis of user complexity. *International journal of man-machine studies*, 22(4):365–394.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., and Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International journal of audiology*, 54(sup2):3–16.

Lee, C.-F. and Kuo, C.-C. (2007). Difficulties on small-touch-screens for various ages. In *Universal Acess in Human Computer Interaction. Coping with Diversity: 4th International Conference on Universal Access in Human-Computer Interaction, UAHCI 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007, Proceedings, Part I 4*, pages 968–974. Springer.

Lewis, J. R. (2014). Usability: lessons learned... and yet to be learned. *International Journal of Human-Computer Interaction*, 30(9):663–684.

Möckel, T., Beste, C., and Wascher, E. (2015). The effects of time on task in response selection-an erp study of mental fatigue. *Scientific reports*, 5(1):10113.

Murman, D. L. (2015). The impact of age on cognition. In *Seminars in hearing*, volume 36, pages 111–121. Thieme Medical Publishers.

Ooster, J., Krueger, M., Bach, J.-H., Wagener, K. C., Kollmeier, B., and Meyer, B. T. (2020). Speech audiometry at home: automated listening tests via smart speakers with normal-hearing and hearing-impaired listeners. *Trends in Hearing*, 24:2331216520970011.

Park, D. and Schwarz, N. (2000). *Cognitive aging: A primer*. Psychology Press.

Pastusiak, A., Niemiec, D., KOCIŃSKI, J., and Warzybok, A. (2019). The benefit of binaural hearing among listeners with sensorineural hearing loss. *Archives of Acoustics*, 44(4):709–717.

Puglisi, G. E., di Berardino, F., Montuschi, C., Sellami, F., Albera, A., Zanetti, D., Albera, R., Astolfi, A., Kollmeier, B., and Warzybok, A. (2021). Evaluation of italian simplified matrix test for speech-recognition measurements in noise. *Audiology research*, 11(1):73–88.

Salman, H. M., Wan Ahmad, W. F., and Sulaiman, S. (2023). A design framework of a smartphone user interface for elderly users. *Universal Access in the Information Society*, 22(2):489–509.

Saunders, G. H., Grush, L., Vachhani, J., Echt, K. V., Griest, S., and Lewis, M. S. (2021). Heterogeneity in vision, hand function, cognition, and health literacy among older veterans: Impacts, outcomes, and clinical recommendations for first-time hearing aid users. *Journal of the American Academy of Audiology*, 32(06):355–365.

Schädler, M. R. (2021). Measurement-prediction-framework: Thoughts on the potential to compensate a hearing loss in noise. https://doi.org/10.5281/zenodo.4500810.

Seidler, R. D., Bernard, J. A., Burutolu, T. B., Fling, B. W., Gordon, M. T., Gwin, J. T., Kwak, Y., and Lipps, D. B. (2010). Motor control and aging: links to age-related brain structural, functional, and biochemical effects. *Neuroscience & Biobehavioral Reviews*, 34(5):721–733.

Smits, C., Kapteyn, T. S., and Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *International journal of audiology*, 43(1):15–28.

Starns, J. J. and Ratcliff, R. (2010). The effects of aging on the speed–accuracy compromise: Boundary optimality in the diffusion model. *Psychology and aging*, 25(2):377.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779.

Swanepoel, D. W., Myburgh, H. C., Howe, D. M., Mahomed, F., and Eikelboom, R. H. (2014). Smartphone hearing screening with integrated quality control and data management. *International journal of audiology*, 53(12):841–849.

Tsai, T.-H., Tseng, K. C., and Chang, Y.-S. (2017). Testing the usability of smartphone surface gestures on different sizes of smartphones by different age groups of users. *Computers in Human Behavior*, 75:103–116.

Wagener, K. (2004). Factors influencing sentence intelligibility in noise. *BIS Verlag*.

Wagener K, K. B. (2005). Evaluation des oldenburger satztests mit kindern und oldenburger kinder-satztest. *Z Audiol*, 44:134–143.

Wardenga, N., Zokoll, M., Kollmeier, B., and Maier, H. (2018). Influence of background noise level on speech reception in normal und hearing impaired persons. *Laryngo-Rhino-Otologie*, 97(S 02):10496.

Wong, C., Thwaites, H., and Khong, C. (2010). Mobile user interface for seniors. *Design Principles and Practices*, 4:231–250.

World Health Organization (2021). World report on hearing.

Ziefle, M. and Bay, S. (2005). How older adults meet complexity: aging effects on the usability of different mobile phones. *Behaviour & information technology*, 24(5):375–389.

Zokoll, M. A., Wagener, K. C., Brand, T., Buschermöhle, M., and Kollmeier, B. (2012). Internationally comparable screening tests for listening in noise in several european languages: The german digit triplet test as an optimization prototype. *International Journal of Audiology*, 51(9):697–707.

S.2 Supplementary Materials

Table S.1: The concept names and explanations are shown, next to examples from provided comments (translated from German to English via DeepL for an unbiased and standardized translation). Underlined words highlight what caused a comment to be counted in the respective concept in longer comments (with multiple concepts).

Concept	**Concept explanation**	**Example comment parts**
Annoying	The interface had annoying properties	- Typing was super annoying - As I said, somewhat cumbersome and sometimes annoying, because you did not know in which direction the searched word was and so forgot the other words. An overview of the words available for selection would have been nice.
clear	The layout was clear	- overview was positive - Clearly designed and intuitive to use
Comprehensible	The interface was comprehensible/ understandable	- Understandable - It was comprehensible, but it took some time to find and select the right word, which often caused the sentence sound to be lost and the words could not be remembered.
correction_ possible	It was possible to correct one's input	- a bit more confusing than B, but no problem if you get clicked wrong
Cumbersome	It was cumbersome to perform the task or the task was too complex	- As I said, somewhat cumbersome and sometimes annoying, because you did not know in which direction the searched word was and so forgot the other words. An overview of the words available for selection would have been nice.

Concept	**Concept explanation**	**Example comment parts**
diff_handle	The interface was difficult to handle	- it was difficult to write down the answer freely; unsure whether to capitalize the words or not - Entering words into the text fields is quite difficult with this smartphone
easy_handle	The interface was easy to handle	- although small, but well selectable
easy	Performing the tasks with this interface was easy or simple	- Super easy to use. Self-explanatory
Fast	The interface was fast (timewise)	- Nice and simple, short and clear, . . . , fast, ...
free_order	One could freely choose the order of providing the answers	- Overview of all possibilities, any order of input possible
good_size	The size of the layout on the smartphone was good	- Font size big enough, scrolling no problem
Intuitive	Performing the task with this layout was intuitive	- Clearly designed and intuitive to use - Super easy to use. Self-explanatory
Old_answers_ marked	Negative perception that old answers were still marked.	- Again, clicked fields remain highlighted, which can be a bit confusing
Mistype	Mistyping occurred / wrong selection of a word	- similar, bad: quickly move on if mistyped; if last word tapped: immediately move on to next sentence - I couldn't correct if I had clicked the wrong way.
no_correction	It was not possible to correct one's input	- In case of mistyping, is it possible to come back? - I couldn't correct if I had clicked the wrong way.

Concept	**Concept explanation**	**Example comment parts**
no_free_order	One could not freely choose the order of providing the answer	- No free order
Slow	It took time to complete the task with this interface, it was not fast	- time-intensive - It took quite a bit of time to write it out, so I wasn't sure about the last few words.
small_size	The size of the layout on the smartphone was rather small	- Although small, but well selectable
Unclear	The layout was not clear	- ...,no overview of words,...
unintuitive	Performing the task with this layout was not intuitive	- unclear whether order of words is important; unclear why not also possible in one line
words_forgotten	Sentence sound lost / Words forgotten/ You had to remember the words for too long	- As I said, somewhat cumbersome and sometimes annoying, because you did not know in which direction the searched word was and so forgot the other words. An overview of the words available for selection would have been nice.

5 General Discussion

This thesis made significant contributions to the field of audiology by introducing an auditory profile generation pipeline for large-scale data analysis. The pipeline utilizes a purely data-driven approach to generate auditory profiles and has a special emphasis on enabling remote testing. Remote testing allows for the collection of larger datasets without restrictions on measurement location, which is highly relevant for covering the entire population.
Auditory profiles can be generated from datasets that are collected remotely, in the lab, or in clinics and then be used in several ways. First, the profiles can be generated from the datasets to understand the underlying data structure. Second, the profiles can be merged from multiple datasets to work towards big data in audiology and establish a global auditory profile. Finally, auditory profiles can be estimated for individual patients using the provided classification models.
Additionally, the thesis also advanced user interface research in audiology for remote testing on smartphones. It specifically focused on adapting the graphical user interface of the matrix sentence test, which is a highly relevant audiological measure for remote testing, to make it compatible with smartphones. This adaptation is particularly beneficial for the often elderly group of hearing loss patients who may have problems with the small screen size of smartphones.

5.1 Contribution of this thesis

Chapter 2 introduced the concept of auditory profiles and describes a proof-of-concept for the subdivision of patients into these profiles using a data-driven approach. The proposed profile-generation approach aligns with existing profiling methods, but is more sensitive to subtle differences in measurement ranges that may have been overlooked previously, due to its data-driven nature, which operates without prior assumptions on the profile distributions. The auditory profiles, thus, summarize information from datasets in smaller patient groups (auditory profiles), each with unique characteristics that distinguish them from other auditory profiles. A profile generation pipeline was developed to generate auditory profiles using model-based clustering. For this, a dataset with 595 patients was used. To obtain a robust estimate of the model parameters (underlying profile number and covariance structure) for the dataset, a bootstrapping approach was implemented. This approach allowed for the estimation of these parameters by repeatedly sampling from the dataset. Finally, different configurations of random forest classification models were compared to identify the best model for classifying patients into the auditory profiles using the available audiological

measures. These classification models can then be used to classify new patients into the profiles in both research and clinical settings. Overall, this proof-of-concept demonstrates the feasibility of subdividing patients into auditory profiles in a purely data-driven manner, providing a useful tool for understanding and classifying patients based on their audiological measures. The profile generation pipeline is flexible and can be applied to various datasets. Although the profile generation approach is independent of specific measures, the resulting profiles are contingent on the underlying data, including the number of profiles, included measures, and covered ranges of patients. It is, therefore, essential to integrate further datasets with varying measures and diverse types of hearing deficits.

In Chapter 3 the profile generation pipeline described in Chapter 2 is extended by a federated learning approach. Additionally, classification models are built for varying feature combinations, such that the profiles can be used in a variety of settings. The 13 profiles described in Chapter 2, were merged with 31 profiles generated from a second dataset (N=1272). The combination of these profiles resulted in a combined set of 13 auditory profiles. To merge the profiles, a profile similarity index using the overlapping density (Pastore and Calcagnì, 2019) was calculated for the common measures. The two datasets contain profiles that exist in both datasets, as well as unique profiles that are only present in one of the datasets. This integration of additional datasets demonstrated the feasibility of expanding the auditory profile set, which should also be generalizable to new datasets. To enable the combination of profiles, datasets may vary in terms of included measures, but a certain number of overlapping measures is needed. The additional information from the remaining measures can be used to estimate plausible ranges for these measures, although with higher uncertainty, as compared to the measures used to merge the profiles. The federated learning approach enables the generation of profiles on sensitive data that is subject to privacy restrictions. For this, the profiles can be generated at the sensitive data location and only the fully anonymized profile distributions are shared to integrate the profiles into the combined global set of auditory profiles.

Chapter 4 contributes to remote testing with the matrix sentence test. The goal is to both develop auditory profiles from large-scale datasets obtained from mobile devices, and classify users of a remote testing tool into one of the auditory profiles using the provided classification models in Chapters 2&3. For this purpose a speech test is needed that can be measured on a mobile device. To achieve this, the matrix sentence test was implemented in a browser-based

online application, allowing for remote measurements via smartphones. A cross-over study was conducted to compare the mobile measurements (using a calibrated setup: smartphone with household headphones) to laboratory control measurements. In addition to a younger normal-hearing group, an older hearing-impaired participant group was included to evaluate the specific impact of mobile measurements for the target group. The first purpose was to assess the general feasibility of conducting the matrix sentence test on a smartphone. The second purpose was to provide an appropriate user interface for the matrix sentence test. The traditional matrix interface requires the display of all 5x10 words on the screen of the smartphone. However, given that the primary target group is predominantly elderly, who may experience challenges with vision and dexterity, three alternative interfaces were evaluated for usability (slide, wheel, type). Against expectations, the traditional matrix interface performed well in terms of accuracy, time efficiency, and user preference. However, not all participants were able to perform the task with this interface and the small size of the interface was noted as a limiting factor. The slide interface ranked second. Every participant was able to perform the task with this interface, and suggestions for improvements were provided to enhance its usability. This finding suggests that the slide interface would be a more suitable choice for a smartphone-based remote testing tool, and highlights the importance of involving the target group in the development of such applications. In future studies, the older hearing-impaired participant group could be expanded to include individuals with more severe dexterity challenges. In this case, it is expected that the slide interface would perform even more effectively than in the current participant group, as the smaller button sizes of the matrix interface may pose an even greater challenge. Overall, this study successfully implements the matrix sentence test in a format that is also suitable for the elderly population, marking an important step towards developing a remote testing tool that can facilitate both remote data collection and diagnostics.

5.2 Auditory profiles for patient characterization

The auditory profiles generated in this thesis are capable of characterizing datasets using a combination of audiogram, ACALOS, and GOESA measures. Although additional features are present in the two analyzed datasets, these three measures were deemed crucial in the classification models and are available for both datasets. This is plausible, as they collectively cover different aspects of hearing loss. While the audiogram covers threshold information, both ACALOS

and GOESA cover different aspects of suprathreshold information. Since only these three main measures were common to both datasets, further reducing the number of measures required for patient characterization is not feasible with the current data. Nonetheless, this highlights the importance of including these measures alongside the audiogram, as relying solely on the audiogram for classifying patients into profiles did not yield adequate performance (see Chapter 3).

To achieve a potential reduction in required audiological measures for patient characterization, a larger set of features would need to be included, initially. This could include, for instance, varying speech tests, to determine the most valuable in terms of patient group separability, but more importantly the addition of new measures. In both the test battery of Van Esch et al. (2013) and the BEAR test battery (Sanchez-Lopez et al., 2021) spectro-temporal resolution was included. The former used the F&T test (Larsby and Arlinger, 1998) and the latter a tone-in-noise test. The BEAR test battery further included a spectro-temporal modulation test. Spectro-temporal modulation tests provide additional information beyond the audiogram in predicting aided SRTs in noise (Bernstein et al., 2016; Zaar et al., 2024), which is highly relevant for hearing aid fitting and could therefore also prove beneficial if included in the auditory profiles.

The current set of auditory profiles mainly cover hearing aid candidates next to few normal hearing individuals due to the source of the two datasets (Research dataset and diagnostic dataset of the Hörzentrum Oldenburg gGmbH). In the future, it would be of great interest to expand the auditory profiles such that cochlear implant patients are also included. Here, datasets mostly include aided performance with hearing aids to assess the speech perception, as the residual hearing needed for unaided speech testing can be too low. In the current set of auditory profiles, however, aided performance was not yet integrated, as it was conceived as an outcome criterion. For cochlear implant patients, however, it can be an indication criterion for the implantation of a cochlear aid. An integration of aided measures into the auditory profiles could, therefore, be plausible in the future.

A benefit of the auditory profiles is that they are derived purely data-driven. While this enables big data analyses, it does not use expert knowledge in the generation of the profiles. Including expert knowledge as labels in the auditory profiles in the form of, for instance, treatment recommendation, audiological

findings, or first fits for hearing aids, is however, important. Including expert knowledge can facilitate the usage of the auditory profiles also in clinical practice and enrich the profiles with important practical information. If profiles are directly related to treatment recommendations, they could also work as part of clinical-decision support system.

One option to connect the auditory profiles to expert knowledge would be by letting audiological experts define audiological findings and treatment recommendations for the different profiles. Such a procedure was used in Buhl et al. (2019, 2020), where audiological experts coded the Common Audiological Functional Parameters (CAFPAs) based on patients test results. The CAFPAs are designed to represent the functional aspects of the human auditory system in an abstract, measurement-independent manner. They function as an interpretable intermediate representation layer in a clinical-decision support and are connected to both audiological findings and treatment recommendations. Including expert knowledge in the derivation of the CAFPAs ensures the plausibility of the CAFPAs. Likewise expert-based ratings could validate auditory profiles and highlight the most valuable profiles. Next to letting experts rate the profiles, it would also be possible to connect the CAFPAs directly to the profiles and in that way include expert knowledge in the profiles. For the patients included in dataset A (see Chapter 2), CAFPA labels are either already available or could be predicted using the classification models provided in Saak et al. (2020). For the patients in dataset B (see Chapter 3), the classification models would need to be adjusted such that the CAFPAs can be predicted based on the available measures in dataset B.

5.3 Auditory profiles in the context of big data analytics

5.3.1 Federated learning

In Chapter 3 the auditory profiles are adapted to function within a federated learning approach. That means profiles can be generated from a larger pool of available data, as they can be computed locally at the sensitive data location, thereby maintaining data privacy (Pfitzner, 2021). This approach allows for the aggregation of large-scale data from various sources, including remotely collected data via smartphones, research datasets, and clinical datasets. Each of these data sources offer unique contributions to the characterization of the audiological patient population. For instance, smartphone-based data collection offers the advantage of accessibility and scalability, potentially capturing information from

a wide demographic due to its easy access and the absence of geographical constraints. Research datasets often provide a larger variety of included features and can thus aid in detecting specific predictors for profile patterns and aid in adapting the current audiological measures used in clinical practice to new insights. Clinical datasets, in turn, provide data for more severe hearing loss patterns, and models derived from this data pool can be directly implemented in clinical practice.

Across each data source "local APs" can be computed that describe the specific data source, while the combination of local APs results in the "global APs" (see Figure 5.1 for a visualization of the federated learning principle with the auditory profiles). Next to profile merging, it also enables the comparison of profile characteristics across data sources. For instance, differences in hearing loss patterns between clinical datasets and research datasets can be assessed by comparing the profile distributions of the "local APs". These in turn, can be compared to the global APs to assess the contribution of individual data sources to the global AP set.

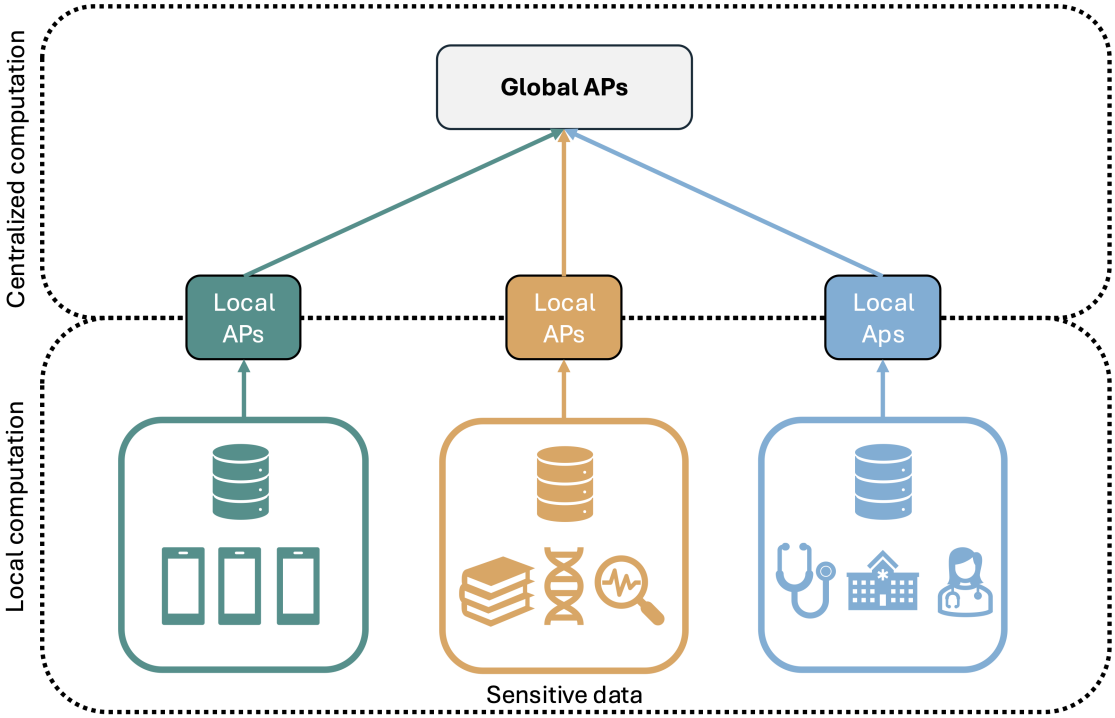

Figure 5.1: Visualization of the federated learning principle of the profile generation pipeline. Sensitive datasets from mobile measurements, research institutes, and clinics can be used to generate auditory profiles (APs) locally. Anonymized profile information from local APs is used to merge the local APs into the global APs. This computation step can occur centralized at a research institute location.

5.3.2 Global auditory profile set

To establish comprehensive global auditory profiles, it is essential to integrate diverse datasets that may vary in terms of the included measures. This can encompass different audiological test batteries or incorporate distinct measures such as genetic information. The profile generation pipeline can generally be applied to these varying types of datasets, as the profile generation process is independent of the specific measures, provided they contain continuous features to meet the requirements of the clustering algorithm.

For profile integration through merging, however, a certain number of overlapping features is required to produce plausible and informative integrated profiles. To exemplify, in line with existing research (e.g., Jepsen and Dau, 2011; Oetting et al., 2016), the profiles show, that profiles with similar audiogram ranges can exhibit distinct differences in other features. Consequently, relying solely on the audiogram for merging profiles would result in a loss of information regarding other features. As a consequence, the merging procedure is currently limited to datasets that share common features.

This raises the question of how to integrate datasets with different measures or completely new data sources. To integrate genetic information, for instance, datasets with both genetic and audiological data are needed initially, such as those collected in the PRESAGE project, which focuses on a comprehensive assessment of early-onset age-related hearing loss (Hochmuth et al., 2024). In this case, genetic sub-profiles could be derived in the first step. In the second step the classification models provided in chapters 2 and 3, could be used to classify patients into one of the available profiles, thus linking profiles to the genetic sub-profiles.

One consequence of merging profiles across datasets is that not all patient within a profile may have information on all features generally included in that profile. For example, dataset A & B share common features, but also contain additional features. When profiles are merged based on these common measures, the additional measures are retained within the profiles. In such cases, the partially available feature information for the additional features can be used to estimate probable ranges for these features, although with higher uncertainty.

5.3.3 Integration of future datasets until profile convergence

The future integration of further local APs into the global AP set from additional datasets, generates the question at which point the global AP set can be considered to actually cover the complete audiological population. To assess this, we can make use of the common principle of iteration until convergence in machine learning (Hastie et al., 2009). Iterating an algorithm until a specific parameter converges is used for model-based clustering, and training classification models, among others. For the auditory profiles, the parameter of interest would refer to the number of profiles. One can assume that with the addition of further datasets the number of profiles will likely increase to some extent. At some point, however, when sufficient hearing loss patterns are included, the integration of additional datasets will not lead to new auditory profiles, indicating that the global APs converged. To avoid local convergence, it's essential to be mindful of the datasets being integrated. If the same patient distributions are repeatedly combined, the profiles may converge prematurely, only to be expanded later when new patient patterns are integrated. An example of this would be the late integration of cochlear implant (CI) patients into the dataset, which would presumably add new profiles to the existing profile set. Figure 5.2 visualizes the concept of the iteration until convergence with the auditory profiles. Convergence is reached, when no new profiles are added to the global AP set and the number of total profiles remains stable.

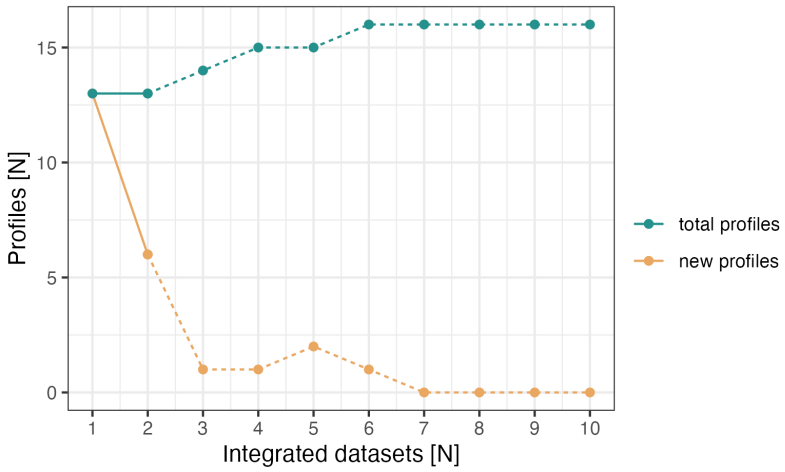

Figure 5.2: Concept visualization of dataset integration until the profile numbers converge. Solid line shows the data for the two integrated datasets. Dashed lines show extrapolated data for the integration of further datasets.

5.4 Auditory profiles in the context of remote testing

Next to the application of the auditory profiles as a facilitator for big data analytics in the field of audiology (see 5.3), they also have the potential to be used in remote testing. Users of a remote testing tool available on smartphones could perform several audiological measures to receive an estimate of the individual hearing deficits. Here, the three factors, namely thresholds (audiogram), loudness perception (ACALOS), and speech perception (GOESA, MST), could be assessed, as these factors were determined as relevant factors for the current auditory profiles. In the future, if additional measures are integrated into the auditory profiles, the remote testing tool could be expanded to incorporate these measures. For measuring the audiogram, however, calibrated devices would be needed to obtain useful results, which is hindered due to the non-trivial task of self-calibrating devices by the non-professional target group of users.

5.4.1 Profile-based missing feature estimation

If certain desired measures, such as the audiogram, cannot be measured on the remote testing tool, the profiles could be used for a statistical estimation of the missing audiological measure. That means the provided classification models, using only the information from the remaining measures (e.g., ACALOS, GOESA, and the age), could be used to classify individuals into an auditory profile and estimate the missing data from the measurement ranges of the predicted profile. For this, several methods could be worthwhile, which are explained with the audiogram as an example for the missing measure. First, a simple approach would be to use the mean, or median audiogram of the respective profile. Second, the audiogram of the individual within the profile that is most similar to the individual of interest could be used. Finally, classification models could be trained for the specific task of estimating audiograms from the provided audiogram ranges of the profile and the measured data. Via these three approaches one could estimate individual audiograms, without needing to measure the audiogram. Generally, it would be of interest to compare the estimated audiograms to both calibrated and uncalibrated controls. The calibrated audiograms would indicate the error of the estimated audiogram. The value of measuring uncalibrated audiograms could be assessed with the uncalibrated audiograms. These could serve to correct the shape of the audiogram estimated via one of the three approaches.

5.4.2 Smartphone-based remote testing tool

The auditory profiles can also serve as a statistical classification system that works in the background of a smartphone-based remote testing tool. Providing individuals with a tool to monitor their hearing performance could motivate individuals to seek help from a hearing care professional earlier. Currently, 8.9 years pass on average between hearing aid candidacy to the actual provision of a hearing aid (Simpson et al., 2019). Given the adverse consequences of hearing loss (poorer quality of life, social isolation, mental health, education, and employment, among others) (Arlinger, 2003; World Health Organization, 2021), it is highly relevant to motivate individuals to seek hearing healthcare early on.

For the auditory profiles, this means they could be connected to specific treatment recommendations, such as profile-based first fits for hearing aids. In the remote testing tool, this could be reflected through a simulation of a personalized hearing aid, which could motivate individuals if they can experience the benefit of a hearing aid. Here, it is likely that profiles that have similar audiogram ranges, but varying ranges for ACALOS, would benefit from different hearing aid settings (Dreschler et al., 2008; Launer et al., 2016), which could be reflected in the different first fits for profiles. When remotely fitting, the profile-based first fits could then be adjusted using self-adjustment to result in an optimal performance (Gößwein et al., 2023). To demonstrate the benefits of the remotely fitted hearing aid, the open Master Hearing Aid (openMHA, Kayser et al. (2022)) could be connected to the remote testing application, such that real-time audio-processing would be feasible. Another beneficial factor of remote testing could be that it could foster remote diagnostics, especially for individuals who are immobile and do not have easy access to hearing care professionals otherwise. Here, the results of the audiological tests and the respective auditory profile could be sent to a hearing care professional for further evaluation.

5.5 Summary

This thesis contributed to advancing big data analytics and remote testing in the field of audiology. An auditory profile generation pipeline was developed that can characterize individuals into auditory profiles. Audiological measures that contribute to the discriminability of the profiles and describe complementary aspects of hearing patterns are based on threshold, loudness perception, and speech understanding. The pipeline can be applied to varying data sources, due

to its federated learning approach that can result in fully anonymized profiles. With the developed merging approach, profiles can be combined across datasets, which paves the way for big data analytics in the field of audiology, as data can be aggregated in the form of auditory profiles without needing to share sensitive data for the computation. For the profiles to be used in practice, in a remote testing setting, varying classification models, which differ with respect to the required features, are provided so that users can be classified into a profile. As the matrix sentence test is an ideal candidate for mobile speech testing, due to its accuracy and repeatability, a mobile implementation was developed, and an appropriate user interface was designed. In that way, the mainly elderly target group can easily perform the tasks on a smartphone, and remote testing in audiology is facilitated.

Bibliography

Arlinger, S. (2003). Negative consequences of uncorrected hearing loss-a review. *International journal of audiology*, 42:2S17–2S20.

Bernstein, J. G., Danielsson, H., Hällgren, M., Stenfelt, S., Rönnberg, J., and Lunner, T. (2016). Spectrotemporal modulation sensitivity as a predictor of speech-reception performance in noise with hearing aids. *Trends in hearing*, 20:2331216516670387.

Buhl, M., Warzybok, A., Schädler, M. R., Lenarz, T., Majdani, O., and Kollmeier, B. (2019). Common audiological functional parameters (cafpas): statistical and compact representation of rehabilitative audiological classification based on expert knowledge. *International journal of audiology*, 58(4):231–245.

Buhl, M., Warzybok, A., Schädler, M. R., Majdani, O., and Kollmeier, B. (2020). Common audiological functional parameters (cafpas) for single patient cases: Deriving statistical models from an expert-labelled data set. *International Journal of Audiology*, 59(7):534–547.

Dreschler, W. A., Esch Van, T. E., Larsby, B., Hallgren, M., Lutman, M. E., Lyzenga, J., Vormann, M., and Kollmeier, B. (2008). Charactering the individual ear by the" auditory profile". *Journal of the Acoustical Society of America*, 123(5):3714.

Gößwein, J. A., Rennies, J., Huber, R., Bruns, T., Hildebrandt, A., and Kollmeier, B. (2023). Evaluation of a semi-supervised self-adjustment fine-

tuning procedure for hearing aids. *International journal of audiology*, 62(2):159–171.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hochmuth, S., Koifman, S., Warzybok-Oetjen, A., Avan, P., Kollmeier, B., and Radeloff, A. (2024). The precision audiology for age-related hearing loss project (presage): Improving the diagnosis of untimely age-related hearing loss. *Laryngo-Rhino-Otologie*, 103(S 02).

Jepsen, M. L. and Dau, T. (2011). Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss. *The Journal of the Acoustical Society of America*, 129(1):262–281.

Kayser, H., Herzke, T., Maanen, P., Zimmermann, M., Grimm, G., and Hohmann, V. (2022). Open community platform for hearing aid algorithm research: open master hearing aid (openmha). *SoftwareX*, 17:100953.

Larsby, B. and Arlinger, S. (1998). A method for evaluating temporal, spectral and combined temporal-spectral resolution of hearing. *Scandinavian audiology*, 27(1):3–12.

Launer, S., Zakis, J. A., and Moore, B. C. (2016). Hearing aid signal processing. *Hearing aids*, pages 93–130.

Oetting, D., Hohmann, V., Appell, J.-E., Kollmeier, B., and Ewert, S. D. (2016). Spectral and binaural loudness summation for hearing-impaired listeners. *Hearing Research*, 335:179–192.

Pastore, M. and Calcagnì, A. (2019). Measuring distribution similarities between samples: a distribution-free overlapping index. *Frontiers in psychology*, 10:1089.

Pfitzner, B, S. N. A. B. (2021). Federated learning in a medical context: a systematic literature review. *ACM Transactions on Internet Technology (TOIT)*, 21(2):1–32.

Saak, S. K., Hildebrandt, A., Kollmeier, B., and Buhl, M. (2020). Predicting common audiological functional parameters (cafpas) as interpretable intermediate representation in a clinical decision-support system for audiology. *Frontiers in Digital Health*, 2:596433.

Sanchez-Lopez, R., Nielsen, S. G., El-Haj-Ali, M., Bianchi, F., Fereczkowski, M., Cañete, O. M., Wu, M., Neher, T., Dau, T., and Santurette, S. (2021). Auditory tests for characterizing hearing deficits in listeners with various hearing abilities: The bear test battery. *Frontiers in neuroscience*, 15:724007.

Simpson, A. N., Matthews, L. J., Cassarly, C., and Dubno, J. R. (2019). Time from hearing aid candidacy to hearing aid adoption: A longitudinal cohort study. *Ear and hearing*, 40(3):468–476.

Van Esch, T. E., Kollmeier, B., Vormann, M., Lyzenga, J., Houtgast, T., Hällgren, M., Larsby, B., Athalye, S. P., Lutman, M. E., and Dreschler, W. A. (2013). Evaluation of the preliminary auditory profile test battery in an international multi-centre study. *International journal of audiology*, 52(5):305–321.

World Health Organization (2021). World report on hearing.

Zaar, J., Simonsen, L. B., and Laugesen, S. (2024). A spectro-temporal modulation test for predicting speech reception in hearing-impaired listeners with hearing aids. *Hearing Research*, 443:108949.

Acknowledgements

I would like to express my sincere gratitude to Birger Kollmeier for making my PhD project possible and for your guidance over the years - from my Master's thesis to the end of my PhD. Thank you for your constant support in helping me grow professionally, while also ensuring the social aspect was never overlooked with the numerous gatherings, such as the Sommerfest, Weihnachtsfest, Medi-Workshop, and the writing workshop on Hvar. It has been a pleasure to be a part of the Medi, of Hearing4all, and the Oldenburg hearing ecosystem in general.

Thank you Mareike, for introducing me to hearing research in the first place and your continuous support in the last years. I have tremendously enjoyed our discussions and thought experiments, both in Oldenburg and during my visit to CERIAH in Paris. They always sparked new ideas and strengthened my scientific curiosity.

Thank you Andrea, for sparking my interest in statistics and machine learning and for always fostering a warm and welcoming working environment.

I would also like to thank all members of the SPHEAR group for the opportunity to discuss both scientific and administrative topics during our weekly meetings. Your support throughout my PhD journey has been invaluable.

Finally, I would like to thank my friends and family. Thank you Florian, for always making me smile and look at the bright side of things. Thank you Jana, for all of your emotional support and for being my rehearsal audience in our time together in Oldenburg. Thank you Sarah, Sandra, and Saskia for our wonderful sister bond, your warmth, patience, and guidance in both personal and professional aspects. At last, I would like to thank my parents. Thank you for always having an open ear and your continuous support in all matters.

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Dissertation hat weder in Teilen noch in ihrer Gesamtheit einer anderen wissenschaftlichen Hochschule zur Begutachtung in einem Promotionsverfahren vorgelegen. Teile der Dissertation wurden bereits veröffentlicht, wie an den entsprechenden Stellen angegeben. Außerdem versichere ich, dass ich die allgemeinen Prinzipien wissenschaftlicher Arbeit und Veröffentlichung, wie sie in den Leitlinien guter wissenschaftlicher Praxis der Carl von Ossietzky Universität Oldenburg festgelegt sind, befolgt und keine kommerziellen Vermittlungs- oder Beratungsdienste in Anspruch genommen habe.

Hamburg, 20.09.24

Samira Saak