



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Multivariate Analysis of Audiological Data from WHISPER and Virtual Hearing Clinic Platforms: A Machine Learning Approach

TESI DI LAUREA MAGISTRALE IN
BIOMEDICAL ENGINEERING - INGEGNERIA BIOMEDICA

Author: **Ilaria Staiano**

Student ID: 994711

Advisor: Prof. Alessia Paglialonga

Co-advisors: Marta Lenatti, Ania Warzybok-Oetjen

Academic Year: 2023-2024

Abstract

Hearing degeneration cases are increasingly frequent nowadays among the adult population; however, this condition often remains under-diagnosed, leading to long-term consequences such as declining auditory capabilities and cognitive issues like anxiety, depression, and attention deficits. This lack of adequate diagnosis frequently exacerbates overall health deterioration. An effective screening process can early identify risk factors and allow timely interventions, thereby mitigating the severity of negative outcomes. Screening tools such as those offered by the WHISPER (Widespread Hearing Impairment Screening and Prevention of Risk) platform developed by CNR-IEIIT and Politecnico di Milano, and the Virtual Hearing Clinic developed by Carl von Ossietzky University of Oldenburg, play a crucial role in this context.

WHISPER provides a straightforward, effective, and clinician-independent screening test, which includes a Speech-in-noise test, a risk factor questionnaire, and a Digit Span Test (DST) to assess cognitive abilities. The collaboration among CNR-IEIIT, Politecnico di Milano, and the University of Oldenburg has facilitated the combined use of both platforms. The main objective of the project is to develop automated methods for multivariate analysis of audiological data using machine learning techniques, aimed at identifying and classifying auditory and cognitive issues in population screening contexts.

The analyses conducted in this thesis provide valuable insights into the impact of various features on auditory and cognitive performances, and the effectiveness of different clustering and classification methods. Data analysis confirmed the expected age-related hearing loss trend, with key variables such as age, Pure Tone Average (PTA), number of stimuli, and total test time playing a crucial role in identifying homogeneous groups through clustering techniques. Despite the challenges posed by the small dataset size, models like Random Forest and SVM demonstrated notable robustness.

In collaboration with the University of Oldenburg, a new dataset has been created to enable cross-validation of speech-in-noise tests. Future research should focus on expanding the dataset and adopting advanced techniques to enhance the reliability of clustering and classification results, thereby deepening our understanding of the complex interactions

between hearing, aging, and cognitive performance.

This thesis lays the groundwork for further research in the field, emphasizing the importance of comprehensive data collection and methodological rigor in advancing our understanding of the dynamics between hearing, aging, and cognitive performance.

Keywords: hearing, Speech in Noise test, cognitive decline, classification, profiling

Abstract in lingua italiana

I casi di degenerazione uditiva sono sempre più frequenti nella popolazione adulta, tuttavia questa patologia rimane spesso sotto-diagnosticata, con gravi conseguenze a lungo termine tra cui il declino delle capacità uditive e problemi cognitivi come ansia, depressione e deficit dell'attenzione. Questa mancanza di diagnosi adeguata spesso porta al peggioramento della salute complessiva dell'individuo. Un processo di screening efficace può individuare precocemente i fattori di rischio e permettere interventi tempestivi, mitigando così l'entità delle conseguenze negative. Gli strumenti di screening come quelli proposti dalle piattaforme WHISPER (Widespread Hearing Impairment Screening and PrEvention of Risk) sviluppata da CNR-IEIIT e Politecnico di Milano e Virtual Hearing Clinic sviluppata dalla Carl von Ossietzky University of Oldenburg giocano un ruolo cruciale in questo contesto.

WHISPER offre un test di screening semplice, efficace e indipendente dal personale clinico, che include uno Speech-in-noise test, un questionario sui fattori di rischio e un Digit Span Test (DST) per valutare le abilità cognitive. La collaborazione tra CNR-IEIIT, Politecnico di Milano e l'Università di Oldenburg ha permesso l'utilizzo congiunto di entrambe le piattaforme. L'obiettivo principale del progetto è sviluppare metodi automatici per l'analisi multivariata dei dati audiologici tramite tecniche di machine learning, al fine di identificare e classificare problemi uditivi e cognitivi in contesti di screening di popolazione.

Le analisi condotte in questa tesi forniscono preziose informazioni sull'impatto delle diverse caratteristiche sulle prestazioni uditive e cognitive, e sull'efficacia di vari metodi di clustering e classificazione. L'analisi dei dati ha confermato il trend atteso di perdita uditiva correlata all'età, con variabili chiave come età, Pure Tone Average (PTA), numero di stimoli e tempo totale del test che hanno giocato un ruolo fondamentale nell'identificare gruppi omogenei attraverso tecniche di clustering. Nonostante le sfide rappresentate dalle dimensioni ridotte del dataset, modelli come Random Forest e SVM hanno dimostrato una robustezza notevole.

In collaborazione con l'Università di Oldenburg, è stato creato un nuovo dataset che

permette la cross-validazione dei test di speech in noise. Le future ricerche dovrebbero concentrarsi sull'espansione del dataset e sull'adozione di tecniche avanzate per migliorare l'affidabilità dei risultati di clustering e classificazione, approfondendo così la nostra comprensione delle interazioni complesse tra udito, invecchiamento e performance cognitive.

Questa tesi getta le basi per ulteriori ricerche nel campo, enfatizzando l'importanza di una raccolta dati esaustiva e di un rigore metodologico nel progredire nella comprensione delle dinamiche tra udito, invecchiamento e performance cognitive.

Parole chiave: udito, test di intelligibilità del parlato, declino cognitivo, classificazione, profilazione

Contents

Abstract	i
Abstract in lingua italiana	iii
Contents	v
1 Introduction	1
1.1 The auditory system	1
1.2 Hearing loss	2
1.3 Risk factors	3
1.4 Hearing tests	3
1.5 Speech in noise tests	7
1.6 Working memory tests	9
1.7 Thesis objectives	10
2 Materials and methods	13
2.1 Test battery	13
2.1.1 Speech in noise tests	13
2.1.2 Working memory test	18
2.1.3 Risk factors questionnaire	19
2.1.4 Summary of extracted features	20
2.2 Acquisition protocol	23
2.2.1 Phase 1	23
2.2.2 Phase 2	25
2.3 Data analysis	26
2.3.1 Dataset characterization	27
2.3.2 Data preprocessing	28
2.3.3 Clustering	32
2.3.4 Classification	38

3	Results	47
3.1	Summary of dataset - phase 1:	47
3.2	Summary of dataset - phase 2:	47
3.3	Analysis of the whole dataset:	48
3.3.1	Characterization of test variables:	48
3.3.2	Clustering using only Whisper features:	58
3.3.3	Classification using only Whisper features:	69
3.3.4	Clustering adding DST and risk factors:	78
3.3.5	Classification adding DST and risk factors:	94
3.4	Comparison among Whisper, DTT and OLSA:	102
4	Discussions	117
4.1	Dataset characterization	118
4.2	Clustering of subjects' profiles	119
4.3	Classification of clustered subjects	122
4.4	Comparative analysis between SIN tests.	126
5	Conclusions	129
	Bibliography	131
	List of Figures	137
	List of Tables	141
	List of Abbreviations	143
	Acknowledgements	145

1 | Introduction

1.1. The auditory system

The auditory system assumes the critical responsibility of processing auditory stimuli, coming from outside the human body and transmitted through the area of the brain that is dedicated to the processing of such signal, that is, the temporal cortex. This system orchestrates the intricate conversion of sound waves, initially perceived as pressure waves in the air, into electrical signals that are subsequently relayed to the central nervous system. Comprising three distinct anatomical components - namely, the outer, middle, and inner ear - the auditory apparatus functions as a sophisticated conduit for the transmission of auditory information.

The outer ear is the most external component, the initial point of contact with the incoming sound waves, and together with the middle ear, they collaborate in transmitting sound to the inner ear, where the actual conversion into electrical signals happens. The outer ear comprises the pinna and the ear canal. The pinna, a skin-covered flap situated on the side of the head, gathers sound waves and directs them inward to the ear canal. These components work together to enhance the transmission of information, aid in sound localization, and perform mechanical filtering, preventing the passage of air pressure vibrations at extremely low and high frequencies. The middle ear, occupying an intermediary position in the auditory pathway, transmits the vibrations of the eardrum, also known as the tympanic membrane, in response to air waves to the ossicles (malleus, incus, and stapes). These ossicles move in synchrony at the same frequency, generating wave-like movements of the fluid in the inner ear. The primary function of the inner ear is to convert sound into neural action potentials. It consists of the cochlea, the main sensory organ for hearing, which resembles a snail-like spiral with three internal channels. Within the cochlea are the basilar membrane and the Organ of Corti. The basilar membrane encodes sound based on its characteristic frequency, with each segment of the membrane vibrating at a specific frequency, similar to a spectrum analyzer. The Organ of Corti acts as a transducer, converting vibrations into nerve impulses through specialized ciliated cells. These cells, arranged in a row on the membrane, serve various functions,

with the inner hair cells functioning as the primary transducers, whose bending triggers the opening of ionic channels in the cell membrane. [1]

The auditory system can be damaged by various factors, such as unhealthy habits or traumatic events, which can lead to a decline in hearing ability, both in early stages and age-related.

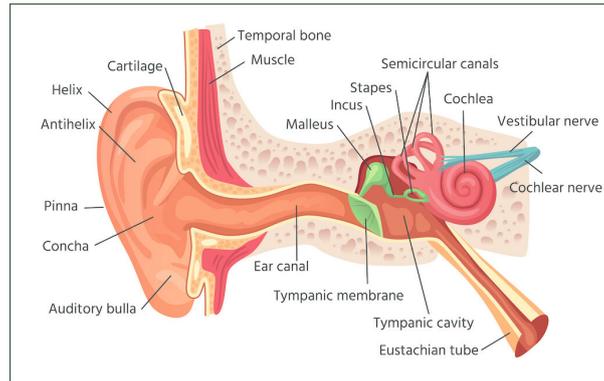


Figure 1.1: The main elements of the auditory system.

1.2. Hearing loss

Cases of hearing degeneration are becoming increasingly common nowadays, especially among the adult population. According to the World Health Organization (WHO), hearing loss is the first among the 20 leading causes of moderate-to-severe disability in adults [2]. Around 466 million people worldwide are affected by disabling hearing loss and this number is estimated to increase to over 700 million in 2050 [2].

Despite the high incidence of this condition, it often goes undiagnosed, leading to long-term consequences. In addition to a decline in hearing abilities, it is frequently associated with cognitive degeneration effects such as anxiety, depression, and attention deficits. Unfortunately, this situation is often overlooked, resulting in a worsening of the individual's health condition. The majority of people that suffer from hearing loss tend to live for years without looking for help, and it leads to an increase in the progression of disabling [3].

In this scenario, it is fundamental to raise awareness about the importance of early detection of hearing loss and to promote widespread screening initiatives, targeting all ages but especially young adults. In this way, it is possible to identify risk factors and treat this condition in the early stages, reducing the severity of its consequences. For this purpose, screening tools that are accessible to everyone play a crucial role, allowing

a greater number of people to be reached compared to lengthy and complex clinical tests.

1.3. Risk factors

The causes behind hearing loss may be different, and they can be divided into two main categories: congenital and acquired causes. Congenital causes are genetic factors that can be hereditary and non-hereditary, and complication during pregnancy or birth. Acquired causes for hearing loss also include individual's habits, and so they are not related to age. They include hearing infections and certain kinds of diseases that can lead to age-related hearing loss in some individuals (like stroke, meningitis, depression, obesity, diabetes, etc.).

Notably, the incidence of cardiovascular disease in patients with hearing loss has also been studied, showing a positive correlation [4]. One of the main factors is the constant exposure to noise, for example in case of workers that use noisy machines every day. Also individuals that use personal audio devices without paying attention to the volume used or that regularly participate in recreational activities where high-volume music or sounds are played (for example discos or concerts) are at major risk. Additionally, cardiovascular diseases, whether caused by smoking or not, are a significant cause for hearing loss and consumption of alcohol can be considered a risk factor too. Awareness-raising initiatives are crucial to make a difference in the detection of hearing loss and also in preventing them.

1.4. Hearing tests

During the years, different types of diagnostic hearing tests have been developed in order to evaluate people's hearing abilities from different perspectives. These tests are chosen based on the context of application, the age of the participants and the part of the ear that wants to be tested. Some of the most known tests are: Pure Tone Threshold Audiometry, Tympanometry, Auditory Brainstem Response (ABR), Otoacoustic Emissions (OAE), Speech Testing, Speech in Noise Tests (SIN).

Pure Tone Threshold Audiometry: Pure Tone Threshold Audiometry is generally considered the screening test choice for adults. The reason why it is usually not performed on infants or non-collaborative subjects is because it requires the active participation of the patient during the whole measurement. This test helps identify hearing thresholds, which are used to describe the severity and frequency affected from the hearing loss. The procedure involves the administration of pure sounds to the subject through the use of

an audiometer, with different frequency and amplitude. The test is conducted for both ears, using frequencies ranging from 125 Hz to 8000 Hz (125-500-1000-2000-4000-6000-8000Hz). A graph is then produced, called audiogram, that shows the minimum value (in decibel) that has been detected by the participant for each frequency and each ear. [5]

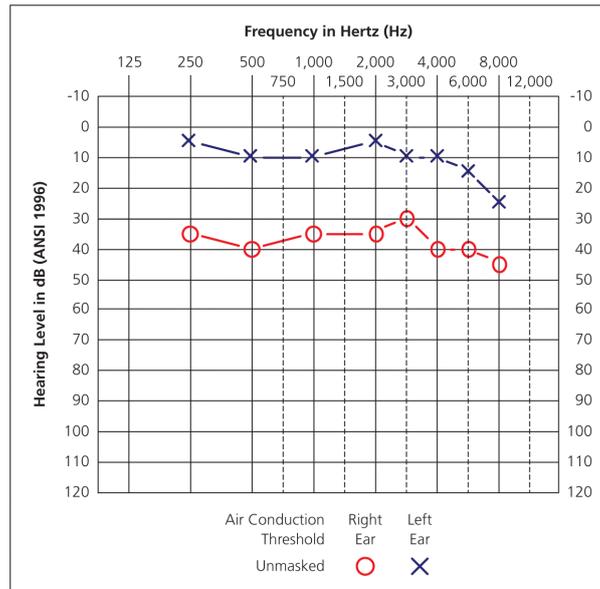


Figure 1.2: Audiogram of a flat conductive hearing loss from [6]. The horizontal axis represents sound frequency, ranging from low to high pitch. The vertical axis represents sound intensity. Thresholds for the right ear are marked with a red circle, while thresholds for the left ear are marked with a blue X.

The hearing threshold for each ear is typically computed by averaging the four central frequencies (500-1000-2000-4000 Hz), obtaining the Pure Tone Average (PTA). If the threshold is higher than 20dB hearing loss is detected and it can have different severity. The threshold reported is not the only one possible, because there are different criteria and standards adopted in literature, but here the one by the WHO is used, based on the following table.

Grade	Hearing threshold^f in better hearing ear in decibels (dB)	Hearing experience in a quiet environment for most adults	Hearing experience in a noisy environment for most adults
Normal hearing	Less than 20 dB	No problem hearing sounds	No or minimal problem hearing sounds
Mild hearing loss	20 to < 35 dB	Does not have problems hearing conversational speech	May have difficulty hearing conversational speech
Moderate hearing loss	35 to < 50 dB	May have difficulty hearing conversational speech	Difficulty hearing and taking part in conversation
Moderately severe hearing loss	50 to < 65 dB	Difficulty hearing conversational speech; can hear raised voices without difficulty	Difficulty hearing most speech and taking part in conversation
Severe hearing loss	65 to < 80 dB	Does not hear most conversational speech; may have difficulty hearing and understanding raised voices	Extreme difficulty hearing speech and taking part in conversation
Profound hearing loss	80 to < 95 dB	Extreme difficulty hearing raised voices	Conversational speech cannot be heard
Complete or total hearing loss/deafness	95 dB or greater	Cannot hear speech and most environmental sounds	Cannot hear speech and most environmental sounds
Unilateral	< 20 dB in the better ear, 35 dB or greater in the worse ear	May not have problem unless sound is near the poorer hearing ear. May have difficulty in locating sounds	May have difficulty hearing speech and taking part in conversation, and in locating sounds

Figure 1.3: Classification of Hearing Thresholds by WHO. Grades of hearing loss and related hearing experience in quiet and noisy environments.

Pure Tone Threshold Audiometry does not assess the ability to recognize voices or to determine the direction from which sounds are coming. This means that while it is effective for identifying the presence and extent of hearing loss, it does not evaluate how well a person can understand speech, especially in noisy environments, nor does it measure the ability to localize sounds, which is crucial for understanding where sounds originate in space. These aspects of auditory perception are important for effective communication and situational awareness and require additional tests.

Tympanometry: Tympanometry is performed in order to assess middle ear effusion or Eustachian tube dysfunction, providing information about the mobility or compliance of the tympanic membrane, the volume of the external ear canal and the pressure within the middle ear. There are different kinds of tympanograms: type A (normal), type B

(that indicates the presence of fluid behind the tympanic membrane) and type C (that means Eustachian tube dysfunction). This test is really useful for middle ear disease detection and also follow-up. The tympanogram results are interpreted by a healthcare professional, such as an audiologist or otolaryngologist. They consider the tympanogram pattern along with other clinical findings to diagnose middle ear conditions and determine appropriate treatment.[7]

Auditory Brainstem Response: It is a reliable method to evaluate cochlear function, being used also intraoperatively and to predict postoperative hearing impairment. It is a neurophysiological test used to assess the function of the auditory nerve and brainstem pathways involved in hearing. During an ABR test, electrodes are placed on the patient's scalp, typically at specific locations such as the vertex (top of the head) and behind each ear. These electrodes pick up the electrical activity generated by the auditory nerve and brainstem in response to the auditory stimuli. The patient is typically presented with a series of brief clicks or tone bursts through headphones or ear inserts. These stimuli are presented at different intensity levels to assess the hearing threshold. The responses are very small and are amplified and filtered to improve their detectability. The electrical responses are typically averaged over multiple presentations of the stimulus to improve the signal-to-noise ratio. This helps to distinguish the neural responses from background electrical activity. The averaged electrical responses are analyzed to identify characteristic waveforms, known as waves I, II, III, IV, and V. These waves represent the sequential activation of different neural structures along the auditory pathway, from the cochlea to the brainstem. The presence, latency (timing), and morphology (shape) of these waves are assessed by an audiologist or healthcare provider. Any abnormalities in the ABR waveform may indicate dysfunction or pathology along the auditory pathway, such as hearing loss, auditory nerve disorders, or brainstem abnormalities.[8]

Otoacoustic Emissions: Otoacoustic emissions are sounds that the cochlea produces in response to sounds presented to the ear. During an otoacoustic emissions test, the patient is typically presented with various types of sounds, such as clicks or tones, through headphones or ear inserts. These sounds can be presented at different frequencies and intensities. After presenting the sound stimuli, a sensitive microphone is placed in the patient's ear canal to pick up the otoacoustic emissions produced by the cochlea in response to the sounds. The microphone detects the faint sounds emitted by the inner ear and records them. The recorded otoacoustic emissions are analyzed by specialized equipment. The analysis includes assessing the presence, strength, frequency, and other characteristics of the emissions. The presence and characteristics of otoacoustic emissions provide valuable information about the function of the cochlea. Normally functioning cochlea

produce otoacoustic emissions, while cochlea with certain types of hearing loss or abnormalities may not produce emissions or may produce weaker emissions. OAEs are often used as part of newborn hearing screening programs to assess the hearing of newborn infants quickly and noninvasively. They can help diagnose hearing loss, particularly in cases where traditional audiometry may be challenging, such as with infants or individuals with developmental disabilities. They can also be used to monitor the effects of ototoxic medications, noise exposure, or other factors that may affect hearing over time. Overall, otoacoustic emissions testing is a valuable tool in assessing hearing function, particularly in populations where traditional audiometric testing may be difficult or impractical.[9]

1.5. Speech in noise tests

Speech-in-noise (SIN) tests are diagnostic assessments used to evaluate a person's ability to understand speech in the presence of background noise. During a speech-in-noise test, the individual is typically presented with recorded speech stimuli, such as sentences or words, played through headphones or speakers. These speech stimuli are presented against a background of varying levels of noise, typically white noise, or speech-shaped noise. The intensity of the background noise is adjusted to create different signal-to-noise ratios (SNRs). SNR is the difference in decibels (dB) between the level of the speech signal and the level of the background noise. Hence, lower SNRs indicate more challenging listening conditions. During test execution the participant is asked to repeat or identify the speech stimuli presented in the presence of noise. The test may involve repeating words, sentences, or discriminating between different speech sounds. The individual's responses are recorded and scored. The performance on the SIN test is typically quantified in terms of the percentage of correctly identified or repeated speech stimuli at each SNR level. The results are used to assess the individual's ability to understand speech in noisy environments. Higher scores indicate better speech-in-noise perception, while lower scores suggest difficulty hearing in noisy situations. SIN tests are used in various clinical settings for different purposes: they help identify individuals with auditory processing disorders, sensorineural hearing loss, or other hearing-related difficulties, particularly in challenging listening environments. They assess the effectiveness of hearing aids or other assistive listening devices in improving speech perception in noise. They guide the development of auditory training programs or rehabilitation strategies to improve speech-in-noise perception. SIN tests are also used in research studies to investigate factors affecting speech perception, such as aging, cognitive function, language proficiency, and hearing disorders. In particular, SIN tests don't necessarily need a soundproof room with calibrated equipment and trained test personnel, making them more flexible compared to the Pure Tone

Threshold Audiometry, less expensive and feasible for large-scale screening. Over the years, these tests have started to be performed through the Internet, making it possible to perform them independently for the user. [10] The main outcome of the SIN test is the speech reception threshold (SRT) that is a measure used in audiology to determine the level of speech intensity (in decibels Hearing Level, dB HL) at which an individual can just barely recognize sounds. The SRT serves as an important clinical measure because it reflects the softest level of speech that an individual can detect and understand, providing valuable information about their ability to perceive speech in everyday listening situations. There are a variety of SIN tests, that differ from each other based on age, procedure, noise, time, speech material, test setting etc. Some examples of SIN tests are: QuickSIN, Words in Noise (WIN), Listening in Spatialized Noise–Sentences (LiSN-S), and Coordinate Response Measure (CRM) [11].

QuickSIN: This test assesses speech perception in noise and may also indicate cognitive processes. It uses a progressive protocol composed by IEEE sentences pronounced by a female voice, using Auditec four-talker babble as background noise. QuickSIN is employed for pre and post hearing aid fitting.

Words in Noise: WIN evaluates speech perception with minimal cognitive influence. The background noise is Causey six-talker babble, with a progressive protocol and NU-6 words with female voice as target stimuli. It is employed for pre-hearing aid fitting.

Listening in Spatialized Noise–Sentences: This test evaluates spatial processing disorder and is used for auditory processing disorder (APD) evaluation. It comprises sentences pronounced by a female voice with children’s stories as background noise. LiSN-S employs an adaptive procedure for assessment.

Coordinate Response Measure: CRM analyzes spatial hearing abilities. It can be performed with adaptive or progressive protocol, using multi-talker babble as noise. The test utilizes sentences as stimuli, often with spatially separated speech or competing talkers. CRM is commonly used in research and clinical settings to evaluate spatial hearing and speech perception in complex listening environments.

The effectiveness of SIN tests can vary based on several factors related to the speech stimuli, including the linguistic complexity (such as phonemes, words, or sentences), which can influence test outcomes. Additionally, factors like the speaker’s gender (male or female), the characteristics of the background noise (e.g., broadband, narrow-band, speech-like noise), and the testing environment (e.g., headphones or sound field) can all impact speech intelligibility in noisy conditions. These variables must be considered when interpreting the results, as they can affect the accuracy and reliability of the assessment.

1.6. Working memory tests

Hearing loss can have a significant impact on various cognitive functions, including working memory. Working memory refers to the cognitive system responsible for the temporary storage and manipulation of information needed for tasks such as comprehension, learning, and reasoning [12]. When hearing loss is present, especially in the higher frequencies, it can be difficult to perceive and process speech, and it can impair verbal working memory. Individuals also expend more cognitive resources on deciphering speech, leading to an increased cognitive load that can make tasks that require simultaneous processing and storage of information heavier. Cognitive challenges increase as the hearing loss gets worse, becoming a relevant aspect to take into consideration when dealing with people that suffer from hearing impairments. Individuals with hearing loss may employ compensatory strategies, for example lip-reading or relying more on visual cues, but it may not fully mitigate the cognitive challenges associated with hearing impairment, particularly in situations with high cognitive demands. Being aware of the impact of cognitive decline derived from hearing loss, several ways to quantify it have been developed. Some of the most widely used tests include:

Digit Span Test (DST): In this test, participants are required to repeat sequences of digits in the same order (forward digit span) or in reverse order (backward digit span). The length of the longest sequence successfully recalled serves as a measure of working memory capacity. [13]

Letter-Number Sequencing: This test assesses working memory by requiring participants to mentally manipulate and reorder sequences of letters and numbers according to specific rules. [14]

N-back Task: In this task, participants are presented with a sequence of stimuli (e.g., letters, numbers, or shapes) and must indicate whether the current stimulus matches the one presented "n" items back in the sequence. The task difficulty can be adjusted by varying the value of "n," with higher values requiring more demanding working memory processes. [15]

Spatial Span Test: Similar to the Digit Span Test, the Spatial Span Test assesses visuospatial working memory by requiring participants to repeat sequences of spatial locations in the same or reverse order. [16]

Reading Span Test: Participants are given sentences to read either on paper or a computer screen. After each sentence, they make a processing judgment (such as whether the sentence is true or false). Following a series of sentences, they are asked to recall items

from memory. The difficulty of the task is adjusted by changing the number of sentences in each recall block. The test score, typically the number of items correctly recalled, serves as a measure of working memory capacity. [17]

Operation Span Task: This task requires participants to solve simple math problems while simultaneously remembering a series of unrelated words. The number of correctly recalled words provides a measure of working memory capacity. [18]

These tests vary in terms of their cognitive demands, sensory modalities (verbal vs. visuospatial), and complexity. Clinicians often use a battery of tests to comprehensively assess working memory function and identify potential declines. Additionally, computerized versions of these tasks are increasingly used in research and clinical settings for their standardized administration and scoring procedures.

1.7. Thesis objectives

Within the scope of this thesis, the primary aim is to develop automated methods for conducting multivariate analyses, utilizing both supervised and unsupervised machine learning techniques, on audiological data collected from two distinct platforms. The first platform is the WHISPER platform (Widespread Hearing Impairment Screening and PrEvention of Risk), jointly developed by CNR-IEIIT and Politecnico di Milano. This platform facilitates the administration of a straightforward, effective, and clinician-independent screening test, which includes a Speech-in-noise assessment, a risk factor questionnaire, and a Digit Span Test to evaluate the cognitive abilities of the subjects under examination [19].

Additionally, the Virtual Hearing Clinic platform, developed by the Carl von Ossietzky University of Oldenburg, serves as another instrumental tool in this endeavor [20]. The utilization of both platforms is made feasible through collaborative efforts between the two institutions.

The overarching goal of the thesis project is to facilitate the identification of potential hearing impairments and cognitive challenges within population screening contexts. To achieve this, the developed methods aim to characterize individual profiles, employing methodologies such as clustering techniques, and to effectively identify issues, particularly those related to auditory function, in accordance with established standard criteria, such as the classifications of hearing loss delineated by the WHO. Furthermore, an additional specific objective of this thesis, in conjunction with the Oldenburg team, entails the generation of a novel dataset encompassing subjects who have undergone testing utilizing

both platforms.

This dataset, collected during a 2-month visiting period in Oldenburg, will serve as a foundation for the cross-validation of profiling and classification methodologies, leveraging the unique attributes of each dataset. The new dataset involved screening with a combination of Whisper test, DST, as well as the Oldenburg Sentence Test (OLSA) and Digit Triplet Test (DTT). By incorporating these tests , that perform speech in noise assessments and cognitive evaluations, the project aims to ensure robustness and reliability in the derived insights and conclusions.

2 | Materials and methods

2.1. Test battery

In this section, the tests employed during the data collection phase are presented. Specifically, three SIN tests (Whisper test, Oldenburg sentence test, and Digit Triplet test), a cognitive test (Digit Span test), and a risk factors questionnaire were utilized. The inclusion of a cognitive test and a risk factors questionnaire was motivated by the potential correlation between hearing loss and cognitive decline, an intriguing aspect that merits thorough examination.

These tests have been used during this thesis, in order (i) to acquire new data to be used as an enrichment of an already existing dataset (Phase 1, Italy) and (ii) to create a new dataset meant to compare different SIN tests (Phase 2, Oldenburg). The acquisition protocols used in Phase 1 and Phase 2 were slightly different and will be described in Sections 2.2.1 and 2.2.2, respectively.

2.1.1. Speech in noise tests

Whisper test

The Whisper speech-in-noise test was developed as part of the Widespread Hearing Impairment Screening and PrEvention of Risk (WHISPER) project, a collaboration between CNR-IEIIT and Politecnico di Milano. The main goal of the project is to promote remote screening for hearing loss and cognitive decline in adults. The project outcome is a test platform that includes a language-independent speech-in-noise test, a risk factor questionnaire, and a cognitive test [19]. From now on, the Whisper speech-in-noise test will be simply called Whisper test, or simply Whisper.

The platform presents a first page where generic personal information are inserted, such as age, gender, language, Pure Tone Threshold Audiometry values (Figure 2.1). The type and brand of device used for the measurement (earphones or headphones) are entered in a subsequent page (Figure 2.2), where participant is then given the opportunity to set a

comfortable volume level (after hearing examples of stimuli they can set the volume in order to be able to listen clearly). The Whisper test is performed separately for each ear, and before starting, the participant can select which ear to measure.

Figure 2.1: Whisper first page: age, gender, language and thresholds of Pure Tone Threshold Audiometry are inserted.

Figure 2.2: Whisper second technical information page, to enter the measurement device, select ear and set comfortable volume level.

The Whisper test is minimally dependent on the listeners' native language because the stimuli used are vowel-consonant-vowel in stationary speech shaped noise (e.g., aba, ada, aga) with an optimized adaptive procedure. The VCVs were pronounced by a professional language native English male speaker who pronounced the VCVs with no prosody, with the stress on the second vowel and with constant pitch [19].

The adaptive staircase method is frequently used in this kind of procedures, and it consists of adjusting the presentation level of the stimuli based on the subject's responses. The questions are proposed in closed form, with 3 alternatives (Figure 2.3).

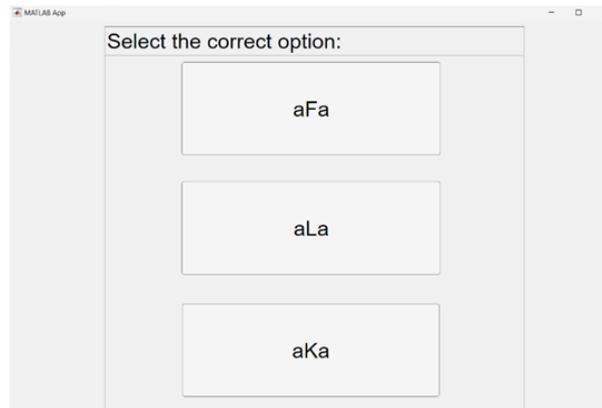


Figure 2.3: Interface for Whisper test, where 3 alternatives closed-form questions are presented. English male speaker with background noise pronouncing VCVs.

The test gets easier when the participant gives incorrect answers and gets more difficult when the responses are correct, following a one-up/three-down (1U3D) logic (after one incorrect response, the test becomes easier, and after three correct responses, it becomes more challenging). Staircase tracking algorithms are guided by non-parametric statistical models to reach the desired target point on the psychometric curve, without relying on specific assumptions about the underlying perception model [21]. Typically, a staircase procedure will stabilize at a target point where the probability of decreasing the presentation level matches the probability of increasing it. For the 1U3D method, the target point corresponds to 79.4% intelligibility. To analyze these inherent differences, the psychometric curves of the 12 VCVs was estimated using the Short-Time Objective Intelligibility (STOI) measure [22].

STOI is a quantitative measure used to assess how well speech can be understood in the presence of background noise. It operates by analyzing short-time segments of speech signals and comparing them with corresponding noisy segments. STOI computes a score that represents the similarity between the clean speech and the degraded speech caused by noise, providing a numerical estimate of intelligibility. STOI values were computed for each VCV across a range of SNRs from -50 to +20 dB, in 2 dB increments. At each SNR, intelligibility was estimated by averaging 100 simulated instances of VCV plus noise. The psychometric curve for each VCV was determined by fitting the STOI values (in 0.25 dB increments) to a cumulative normal model (sigmoid function)[23].

Using a k-means clustering approach [24], the 12 psychometric curves were grouped into four clusters (CLS1: asa; CLS2: afa, aga, aka, ata; CLS3: aba, ada, ala, ana, apa, ara; CLS4: ama). The average psychometric curves for these clusters are displayed in Figure 2.4.

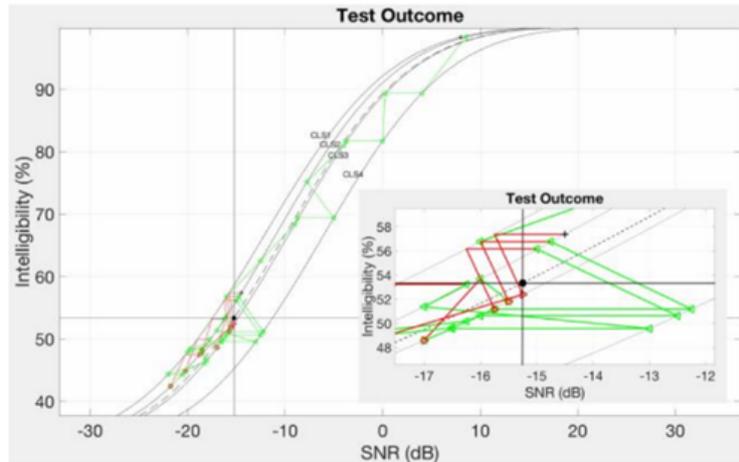


Figure 2.4: Average psychometric curves for the four clusters of VCVs.

In the proposed staircase method, the 1U3D rule was applied using intelligibility steps across the psychometric curves of the four clusters, as shown in Fig. 2.4. Specifically, the adaptive procedure adjusts to lower intelligibility levels (green lines) after three correct responses and to higher intelligibility levels (red lines) after one incorrect response. To prevent abrupt changes in SNR during testing, the 1U and 3D SNR steps were restricted to +5.4 dB and -4 dB, respectively, maintaining an optimal ratio between the steps as recommended by [25]. This approach allows the intelligibility staircase to be terminated after 12 reversals, with the intelligibility threshold estimated as the average of the mid-points of the last four ascending runs, following the guidelines in [25]. The SRT of the SNTCLS was then calculated along the average psychometric curve of the 12 VCVs as the SNR value at the intelligibility threshold (indicated by the black mark on the dashed line in Fig. 2.4) [26].

One of the main outcomes of the Whisper test is the SRT value, that is stored and computed as the average SNR of the last 4 ascending runs (ascending runs= trials between one incorrect answer and 3 correct answers in a row). Other variables about the number stimuli, percentage of correct responses etc, are stored and extracted from the test. They are described more in details in Section 2.1.4 ('Summary of extracted features').

Matrix test: Oldenburg sentence test

The OLSA test is the German version of the Matrix Test [27], first developed by Hagerman (1982) for the Swedish language, and available in different languages today.

The speech material consists of a base matrix of 50 words (10 names, 10 verbs, 10 numerals, 10 adjectives, 10 objects) from which grammatically correct but semantically

unpredictable five-word sentences are created with a random combination of one word from each group (see Figure 2.5 for an example). This is done in order to maintain the language-specific sentence structure for the given language, to assure that the listeners are familiar with the grammar of the sentences and avoid biasing due to listener's syntactic competence. The test can be performed using an open-set or a closed-set response format.

In the context of this thesis, the closed-set response format was used, to enable the test to be self-administered by the participant, by pressing the appropriate response on a keyboard, touchscreen or any other response device, avoiding the necessity of repeating the words by the subject and scoring them by the test administrator, breaking the language barrier. 100 sentences are generated in a way that accounted for all possible combinations of word transitions. For the measurements, the participant is presented with 20 sentences in succession, each of which consists of five words in German and always has the same structure: Name, Verb, Number, Adjective, Object. After the performance of a sentence, a table appears on the screen, in which the subject can select understood words and then press OK to continue with the next sentence. It is possible to guess or to leave the empty space. The speaker voice is a female voice, and as background noise it is used an ICRA1 [27]. The SRT is retrieved as the SNR yielding 50% speech intelligibility.

At the end of the test, a final page shows the curve of intelligibility [%] with respect to the stimuli, together with other technical information (e.g. the noise level, the SNR, the % deviation of intelligibility from the reference value, etc), and a series of features are stored in a .XML file (see Section 2.9 for more details about the most important features extracted).

Andrea	cerca	due	bottiglie	azzurre
Chiara	compra	quattro	macchine	belle
Luca	dipinge	cinque	matite	bianche
Marco	manda	sette	pelle	grandi
Maria	possiede	otto	pietra	nere
Matteo	prende	nove	porte	normali
Sara	regala	dieci	scatole	nuove
Simone	trascina	venti	sedie	piccole
Sofia	vole	poche	tavole	rosse
	vole	molte	tazze	utili
Anna	compra	quattro	matite	grandi

OK

Figure 2.5: Example of the Matrix Test in Italian. The OLSA test is the German version of this Matrix Test, where sentences are presented in this way but in German language.

Digit Triplet Test

The DTT [28] consists of spoken numbers in background noise, precisely, it comprise 27 sequences of 3 numbers each, that are presented to the participant who has to reproduce the sequence using a keypad (Figure 2.6). The main feature that is estimated is the SRT, retrieved as the SNR yielding 50% speech intelligibility. The speech material for the German and Italian versions is composed by digits between 0 and 9, composed as triplets. The speaker, a female singer and speech therapist, maintained a consistent speaking effort and speech rate, averaging about 120 syllables per minute. To ensure uniformity and eliminate any long-term trends in speaking effort, the level of all recordings was adjusted to achieve the same average root mean square (RMS) level as the announcements in all recorded triplets.

Subsequently, the recorded triplets were manually segmented into individual digits using Syntrillium Cool Edit 2000, and the two best realizations for each digit at each position in a triplet were selected. Silent intervals of 200 ms were inserted between digits within a triplet.

The quasi-stationary masking noise was generated by superimposing the speech material (the digits from versions 1 and 2) thirty times, with variable time intervals between elements ranging from 5 ms up to 2000 ms. This method ensures that the long-term spectrum of the noise aligns with the spectrum of the speech material, thereby optimizing masking and creating a sharp discrimination function [29].

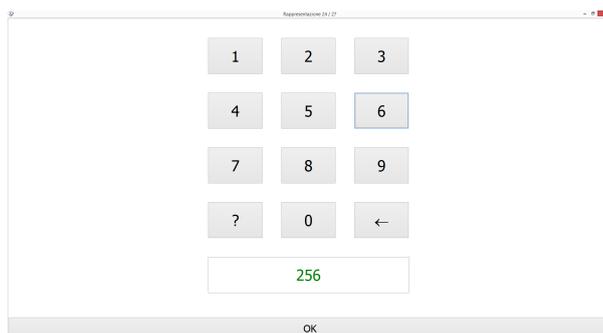


Figure 2.6: Interface of the DTT, where digits are inserted.

2.1.2. Working memory test

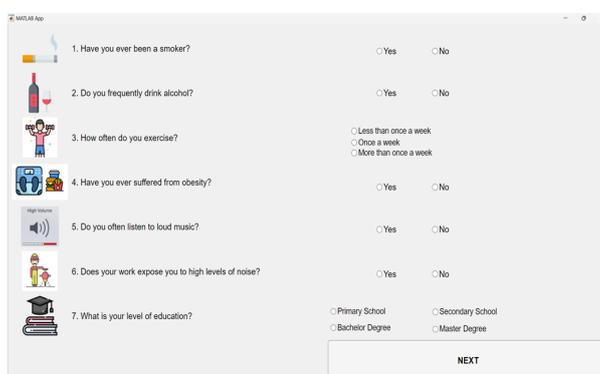
Digit Span Test

The DST [13] is used to measure cognitive ability and short-term memory capacity. It involves the reproduction of a series of numbers, either in the order they are presented

(forward digit span) or in reverse order (backward digit span), immediately after seeing them on the screen. In the context of this thesis, the forward version was used. The platform shows a sequence of digits (from 0 to 9) one at a time, starting from a sequence of 3 numbers, and after the presentation of the sequence, the participants have to reproduce the sequence in the right order using a keyboard (the maximum time to complete the task is 20 seconds). If the sequence is completely correct, the next sequence is one digit longer than the previous one, otherwise it keeps the same length for another trial. The maximum length of the sequence is 9 digits, while the test stops when two consecutive incorrect answers are given. Each digit is shown for 0.7 seconds, while between one sequence and the next one there is 1 second of pause. At the end of the test, the score is shown, that represents the maximum length of the sequence that was correctly guessed by the user (Digit Span Score, DSS). The sequences are randomly generated, assuring that the same sequence is not shown more than once during the same test.

2.1.3. Risk factors questionnaire

The risk factors questionnaire comprises a series of 18 questions regarding the subject's medical history, habits, and lifestyle. The questions are presented in closed form (Figure 2.7) and gather information on the following risk factors related with hearing loss and cognitive decline: diabetes, cardiovascular diseases, cholesterol, depression, COVID-19, family history of hearing loss, tinnitus, ear infections, head trauma, stroke, meningitis, smoking, alcohol consumption, physical activity, obesity, listening to loud music, exposure to noisy work environments, and education level. In this protocol, the questionnaire is administered prior to the beginning of the Whisper test, and the collected data is saved in an Excel file through a MATLAB platform.



The screenshot shows a MATLAB App window titled "MATLAB App" containing a questionnaire. The questions are as follows:

1. Have you ever been a smoker? (Yes/No)
2. Do you frequently drink alcohol? (Yes/No)
3. How often do you exercise? (Less than once a week, Once a week, More than once a week)
4. Have you ever suffered from obesity? (Yes/No)
5. Do you often listen to loud music? (Yes/No)
6. Does your work expose you to high levels of noise? (Yes/No)
7. What is your level of education? (Primary School, Secondary School, Bachelor Degree, Master Degree)

A "NEXT" button is located at the bottom right of the questionnaire area.

Figure 2.7: Example of questions from the risk factors questionnaire. There are in total 18 questions.

2.1.4. Summary of extracted features

Features extracted for Whisper test + Risk factor questionnaire + Digit Span Test			
ID - Subject ID	dx1, dx2, dx3, dx4, dx5, dx6, dx7 sx1, sx2, sx3, sx4, sx5, sx6, sx7 - Air conduction pure tone audiometry for right and left ear (250, 500, 1000, 2000, 4000, 6000, 8000 Hz)	Screening3classes - 0 if PTA <= 20dB HL, 1 if 20<PTA<=40 dB HL, 2 if PTA>40dB HL	#trials - number of proposed stimuli (SIN test)
Age - Subject Age	PTA - Pure Tone Average (average of the pure tone thresholds at central frequencies(500-4000Hz)	total_time - total SIN test time (sec)	srt - speech recognition threshold in dB SNR (SIN test)
Ear - Ear tested with SIN	Screening20 - Binary PTA (0 if PTA<= 20dB HL, 1 otherwise)	% correct - percentage of correct responses (SIN test)	avg_timecount - average response time in seconds (SIN test)
Gender - Gender at birth	Screening40 - Binary PTA (0 if PTA<= 40dB HL, 1 otherwise)	#correct - number of correct responses (SIN test)	volumeLevel - volume amplification coefficient
Native_language - first language	Risk factor YES/NO variables - cardiovascular disease (cardio), cholesterol, covid, depression, diabetes, drink, ear infections, family history hearing loss (family_history_HL), head trauma, meningitis, obesity, smoke, stroke, high_volume_exposure, work_exposure, tinnitus	education - level of education (primary school, secondary school, bachelor, master)	exercise level - how often (less than one a week, once a week, more than once a week)
Place_testers - testers group	avgSingleDigitTimes - average time to insert i-th digit of the sequence in DST	binaryResults - binary result for each proposed sequence (1=correct, 0=incorrect) in DST	cancelPress - stores the number of times the subject presses the cancel key for each single trial (1 value for each single trial)
digitSpanScore - max sequence length guessed by the user in the DST	correctPercentages - percentage of the sequence guessed by the user in the DST ((1 value for each single trial)	responseTimes - total response time (sec) for each proposed sequence in DST	touchScreen - use of a touch screen

Figure 2.8: Summary of the main features extracted from Whisper, risk factors questionnaire, and DST.

Most significant features extracted for Matrix Test + Digit Triplet Test					
ID - Subject ID	MEASUREMENTDATE - date of the measurement (year/month/day/msec/sec/min/h)	SpeechLevel - (65dB)	SNR - signal to noise ratio's value		
MEASUREMENTSTATUS - «complete» if the measurement was finished.	MEASUREMENTID - «Digit3» for DTT, «Olisa» for OLSA	Intelligibility - value of intelligibility in %	TrialsDone - number of trials (20 sentences for OLSA, 27 sequences for DTT)		
Noise - «id 0lnoise» for OLSA, «id digit3noise» for DTT	DisplayedWords - words that the subject actually sees on the screen	TargetWords - words that are presented before shuffling (if shuffle is performed in closed test). It is equal to DisplayedWords in the context of this thesis because no shuffle is performed.	SelectedWords - it contains the symbol "." in correspondence of a wrong word guessed by the subject and the right word is shown when it is correct		
ClosedSelection - it contains the actual words that are selected by the client, also the wrong ones	ClientLanguage - selected language of the test	TransducerName - model of the transducer (HDA200)	MeasurementStart - time when the acquisition begins		
MeasurementEnd - time when the acquisition ends	ReadableResult - Final results of the whole test. It contains values of Intelligibility: [%]; Type of transducer (always Headphones); Channel(left or right); Model of transducer (HDA200), Speech level(65,0dB) Noise level; Noise Type.	TransducerType - type of transducer used (headphone)			

Figure 2.9: Summary of the main features extracted from the OLSA and the DTT tests.

Elaborated features for the analysis		
<p>avgSingleDigitTimes_mean – average time to insert i-th digit of the sequence in DST over all the trials e.g, the average of avgsingleDigitTimes.</p>	<p>response_first_numbers - total time required to enter the first sequence of the DST (of 3 digits).</p>	<p>incorrect_avg - Average single digit typing time for incorrect trials in DST.</p>
<p>cancelPressTot – it stores the total number of times the cancel key is pressed by each user during the entire DST.</p>	<p>correctPercentagesMean - it stores the average Percentage of the sequence guessed by each user during the entire DST.</p>	<p>correct_avg - Average single digit typing time for correct trials in DST.</p>

Figure 2.10: Summary of the main features that have been subsequently elaborated for the analysis, all the tests.

This section reports a summary of the features extracted in each component of the test battery, namely, Whisper, risk factors questionnaire and DST (Figure 2.8), OLSA and DTT (Figure 2.9). For the various analysis, different subsets of variables have been utilized and others have been excluded, as reported more in details in the Results chapter. The following variables for Whisper and DST have never been employed: Screening3classes, Screening20, Screening40, volumeLevel, Place_Testers, touchScreen. Some of the variables were excluded because they have been computed in different ways during time, volumeLevel for example was not included because of the introduction of calibration (explained in the Acquisition protocol section). Risk factors related to covid, exercise, and tinnitus were not considered as they were added in the platform recently, hence they were collected just for a few subjects. Finally, some additional features have been further elaborated for the analysis (Figure 2.10).

2.2. Acquisition protocol

This section describes the two acquisition protocols used in this thesis. The first acquisition phase was conducted in Italy, at Politecnico di Milano, with the aim of acquiring data to expand an existing dataset. The experimental protocol was approved by the Politecnico di Milano Research Ethical Committee (Opinion No. 13/2022, April 13, 2022). Special attention was given to recruiting volunteers aged between 30 and 40 years, as the number of data points in the existing dataset for this age group was lower compared to other age groups. Nonetheless, younger volunteers were also tested since the acquisitions took place at the university and colleagues were available to be tested. In total, 33 new subjects were added to the dataset. The second acquisition phase was conducted at the University of Oldenburg, involving data collection from normal-hearing subjects under 40 years of age, with the objective of comparing SIN tests (Whisper, OLSA, and DTT). 17 subjects were acquired, forming a second dataset that was subsequently used for the comparative analysis. Overall, during the development of this thesis, 50 new subjects were acquired in total. The two acquisition protocols are detailed below.

2.2.1. Phase 1

Participants

This phase includes the measurements performed at Politecnico di Milano, that have been used to enlarge an already existing dataset, composed of data acquired during laboratory testing and opportunistic hearing screening initiatives. The dataset comprises tests for

single and both ears. During the measurement that have been performed for this thesis, a priority has been given to looking for subject around 30-40 years old, but volunteers of all ages were admitted too.

Procedure

During the initial phase of the protocol, participants engaged in a comprehensive series of assessments comprising Pure Tone Threshold Audiometry, the DST, and the Whisper test. Before starting, participants were presented with and required to sign an informed consent form, expressing their voluntary participation in the study, that guaranteed anonymity.

The initial segment of the session focused on conducting Pure Tone Threshold Audiometry measurements utilizing a designated auditory assessment device, a clinical audiometer (Amplaid 177+, Amplifon with TDH49 headphones). Both the right and left ears of participants underwent evaluation across a spectrum of frequencies including 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz, and 6000 Hz. This systematic examination facilitated the determination of auditory thresholds across a range of frequencies, providing valuable insights into participants' hearing acuity.

Following the Pure Tone Threshold Audiometry assessments, participants proceeded to undertake the DST. The DST was always kept right after the Pure Tone Threshold Audiometry measurement in order to avoid the risk of biasing the measurement due to subject's fatigue. Before the measurement, participants received a comprehensive explanation of the test procedures, ensuring standardized administration and complete participant understanding. The DST was conducted under controlled conditions to mitigate potential confounding variables, ensuring the reliability and validity of test outcomes.

Subsequently, participants underwent the Whisper test. At first, it consists of a series of questions about general information (age, gender at birth, first and second spoken language). These questions are presented in closed form, using an easy-to-use and intuitive platform developed with MATLAB. The test involves the use of earphones to deliver auditory stimuli to one ear at a time. This sequential approach allowed for precise evaluation of auditory perception in each ear independently. The entire duration of the measurement session was slightly more than 30 minutes.

By adhering to this protocol, the study aimed to comprehensively evaluate auditory function, short-term memory capacity, and auditory processing abilities in participants, contributing to a deeper understanding of auditory cognition and its implications.

2.2.2. Phase 2

Participants

The second phase of the data acquisition was performed at the university Carl von Ossietzky of Oldenburg, using as target German native listeners younger than 40 years old, all normal hearing.

Procedure

The procedural framework for the measurement session was collaboratively established and agreed upon by both the research teams at Politecnico di Milano and Oldenburg University. The complete duration of the measurement session encompassed approximately one hour, meticulously structured to ensure methodological consistency and adherence to ethical guidelines.

Each participant was provided with an informed consent form, explaining the purpose and procedures of the study, and expressing their voluntary participation. It's noteworthy that participant anonymity was strictly maintained throughout the study.

The measurement session commenced with the administration of Pure Tone Threshold Audiometry using a designated device, an Interacoustics AC40 (calibrated by the Interacoustics service having licence for this). Notably, if participants had undergone Pure Tone Threshold Audiometry measurements within the recent past (3 months), the previously obtained results were utilized to minimize redundancy and optimize efficiency.

Following the Pure Tone Threshold assessment, participants progressed to undertake the DST, a standardized measure of short-term memory capacity and cognitive function. Subsequently, participants underwent the Whisper, OLSA, and DTT tests in a randomized order. This randomized sequencing was adopted to mitigate the potential influence of fatigue on performance and to maintain the independence of performance from the order of administration.

By organizing the measurement session in this manner, the research teams aimed to ensure the integrity and reliability of the data collected, while also safeguarding participant well-being and adherence to ethical standards throughout the study duration.

Calibration

For the protocol used in Phase 2, it was decided to calibrate the equipment used for the DST and Whisper tests, as the OLSA and DTT were already being performed with

calibrated equipment. Up to that point, to conduct Whisper, the computer volume was set at the midpoint of the volume controller, allowing the listener to adjust it to their preferred level (the level they found comfortable); this adjustment was stored as a gain in the software. The calibration procedure involved using an artificial ear, specifically the Brüel & Kjær (B&K) 4153 artificial ear (Figure 2.11).



Figure 2.11: Artificial Ear (Brüel & Kjær (B&K) 4153)

It was decided to consistently use the same equipment throughout the study (laptop MSI Thin GF63 12VF-291IT and headphones HDA200), setting the volume value corresponding to an output level of 65 dB during calibration. This volume value was then applied to all participants.

For calibration, a calibration signal needed to be defined. It was decided to utilize the noise employed during the measurements as a masker, namely a Gaussian noise filtered with the spectrum of the international masking signal. Consequently, the speech material in the software was calibrated using a stationary speech-shaped noise, generated from the speech material itself to match its long-term spectrum.

Subsequently, the final noise, filtered and normalized based on the VCVs, was extracted and played through the artificial ear. The amplifier gain was adjusted to calibrate and achieve an output level of 65 dB from the headphones.

2.3. Data analysis

The data analysis was conducted entirely using Python 3. This section describes the theory behind the techniques used, specifically regarding dataset characterization, data pre-processing, clustering, and classification.

2.3.1. Dataset characterization

The dataset already existing and the newly acquired datasets (i.e., data collected during Phase 1 and Phase 2), went through a process of dataset characterization. This analysis was done in order to understand the data acquired, the most important features and their distribution, some missing values or abnormalities in the data, and to prepare them to be further processed and then make clustering and classification more efficiently.

Histograms, scatterplots, feature reduction, standardization and also some feature analysis and modification were performed (for more details, see the Results chapter). In addition, correlation measures for some variables were performed, to measure the degree to which two variables relate to each other. If two variables are correlated, it means that a change in one variable is associated with a change in the same direction (positive correlation) or in the opposite direction (negative correlation) in the other variable. Two types of correlation were exploited: Spearman and Pearson correlation.

Pearson correlation measures the strength and direction of a linear relationship between two continuous variables, assuming normally distributed data. The coefficient of correlation ranges from -1 (perfect negative linear relationship) to 1 (perfect positive linear relationship), with 0 indicating no linear relationship. Spearman correlation measures the strength and direction of a monotonic relationship using ranked data. It does not assume normal distribution and can capture both linear and non-linear relationships. The coefficient also ranges from -1 to 1, with similar interpretations as Pearson's. Pearson is best for linear relationships with normally distributed data, while Spearman is more versatile and can handle ordinal data and non-linear relationships.

Analyzing data before performing clustering or classification is crucial for ensuring the quality and reliability of the results. It helps identify and handle missing values, outliers, and inconsistencies, and allows for understanding data distributions and relationships, which are essential for selecting appropriate algorithms and preprocessing steps.

In Figure 2.12 the pipeline for data analysis that has been adopted in the thesis is summarized. More details about data preprocessing, clustering, and classification steps are reported in the following.

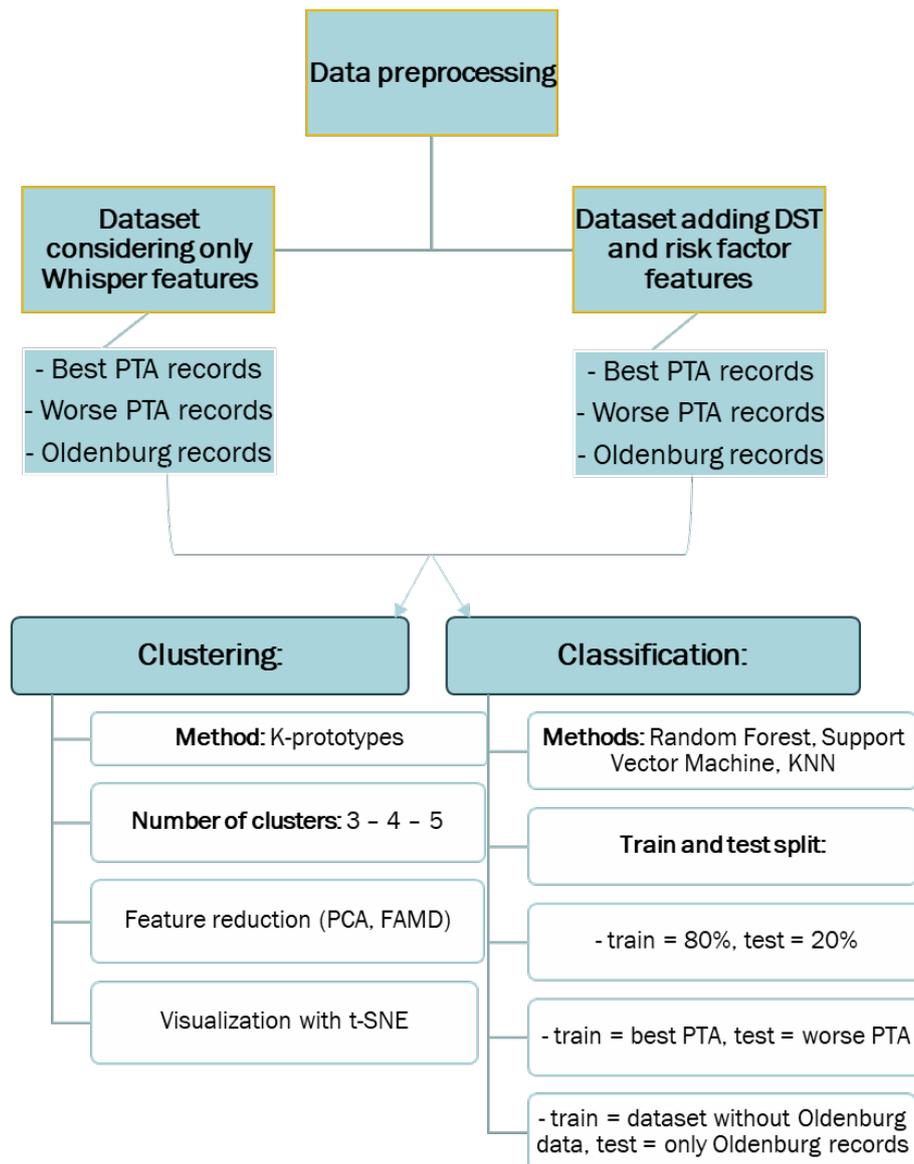


Figure 2.12: Pipeline of data preprocessing, clustering and classification analysis.

2.3.2. Data preprocessing

Data preprocessing is a crucial step in the data analysis and Machine Learning (ML) pipeline. It involves transforming raw data into a clean and usable format, which is essential for improving the quality and performance of models. Preprocessing steps like data cleaning, normalization and standardization help in correcting errors, inconsistencies, missing values and noise, that are often present in raw data. Models trained on preprocessed data generally perform better because the data is more consistent and representative of the underlying patterns. Preprocessing helps in highlighting relevant features and reducing the impact of irrelevant or misleading information. Preprocessing

techniques such as dimensionality reduction and feature selection help in minimizing the amount of data that needs to be processed, thereby reducing computational costs and improving the efficiency of the learning algorithms. Common data preprocessing techniques are:

Data cleaning: there are several techniques to deal with missing values, including deletion (removing rows or columns with missing values), imputation (replacing missing values with mean, median, mode, or predictions), and interpolation. For outlier detection and removal, instead, statistical methods (e.g., Z-scores, IQR) or machine learning techniques are frequently used.

Data transformation: it includes different techniques, for example normalization, that can be performed in different ways.

Normalization: Scaling features to a common range, typically [0, 1] or [-1, 1], or to have zero mean and standard deviation (std) equal to one (Standardization), to ensure that no single feature dominates the learning process.

To standardize data means to transform them according to the following formula:

$$z = \frac{x - \mu}{\sigma}$$

where: μ is the mean of the features in the training set, x is the original value, σ is the standard deviation of the feature in the training set, z is the standardized value.

Encoding Categorical Data: Encoding categorical data is essential because most ML algorithms require numerical input. By converting categorical variables into numerical values, algorithms to identify patterns and make accurate predictions can be effectively used. Two ways of performing this task are One-Hot Encoding and Label Encoding.

- *One-Hot Encoding:* Converts categorical variables into a series of binary variables (0 or 1) to ensure they can be handled by ML algorithms.
- *Label Encoding:* Assigns a unique integer label to each category. This is suitable for ordinal data where the categories have a meaningful order.

Dimensionality Reduction: it involves reducing the number of variables under consideration, in order to enhance model performance and generalization, and also to improve computational efficiency and have a simplified data visualization and interpretation. Some of the most frequently applied techniques are Principal component analysis, Factor Analysis of Mixed Data and t-Distributed stochastic Neighbor Embedding.

Principal component analysis (PCA).

PCA reduces the number of features while retaining most of the data's variability by transforming features into a new set of orthogonal components. To determine the number of components to retain, the explained variance method is used. This method involves examining the cumulative explained variance ratio and selecting the number of components that account for a desired proportion of the total variance, often 90% or 95%. By doing so, PCA ensures that the reduced feature set maintains the essential information present in the original dataset. Additionally, it's important to note that PCA operates exclusively on numerical variables. This means that categorical or non-numeric data must be preprocessed or excluded from the analysis before applying PCA. By focusing solely on numeric variables, PCA identifies the directions (principal components) in the feature space that capture the most significant variance, making it particularly suitable for datasets with numerical attributes [30].

Factor Analysis of Mixed Data (FAMD): this method is utilized for feature reduction too. Unlike PCA, which operates solely on numerical variables, FAMD can handle both categorical and numerical variables, providing a comprehensive view of the dataset. FAMD allows for the exploration of underlying data structures and the identification of latent patterns. Similar to PCA, FAMD transforms variables into a new set of orthogonal components. As in the PCA, to determine the number of components to retain, the explained variance method was employed. By leveraging FAMD, essential information present in the original dataset is preserved in the reduced feature set. The inclusion of categorical variables makes FAMD suitable for datasets with mixed data types, allowing for a more nuanced understanding of the data's structure and relationships. Additionally, FAMD offers the advantage of assessing the contribution of each feature to the newly created components. This feature contribution analysis allows for a deeper understanding of the role and importance of individual variables in shaping the data's structure. By observing the percentage of variance explained by each feature within the components, researchers can identify the most influential variables driving the underlying patterns. This capability enhances the interpretability of the results and aids in identifying the key factors driving the observed relationships in the dataset [31].

t-Distributed stochastic Neighbor Embedding (t-SNE): It is a technique utilized for dimensionality reduction, particularly for visualization purposes, while maintaining the intrinsic structure and relationships within the data. t-SNE achieves this by transforming high-dimensional data into a lower-dimensional space where similar instances are modeled by nearby points, while dissimilar instances are modeled by distant points. Unlike PCA, which focuses on preserving global structures, t-SNE prioritizes the preservation of local

structures, making it particularly effective for revealing clusters and patterns in complex datasets. Additionally, t-SNE is often used as an exploratory tool to gain insights into the underlying structure of the data. However, it's important to note that t-SNE is computationally intensive and sensitive to hyperparameters, requiring careful tuning for optimal performance.

The main hyperparameters are the *perplexity* and *learning rate*. The *perplexity* parameter influences the selection of nearest neighbors in high-dimensional spaces. It represents the number of neighbors t-SNE considers during dimensionality reduction. A higher *perplexity* implies considering more neighbors for each point, while a lower *perplexity* means considering fewer neighbors. Typically, *perplexity* values range from 5 to 50. It's essential to experiment with different *perplexity* values to achieve the best visualization of the data. Additionally, the *learning rate* controls the speed at which points move in the new dimensional space during optimization. Higher *learning rates* lead to faster point movements but may result in instability or suboptimal outcomes. Conversely, lower *learning rates* can slow down optimization and require more iterations to reach a stable solution. The choice of *learning rate* depends on the dataset's characteristics and often requires experimentation to find the optimal value for stable and efficient convergence [32].

In this thesis, the different configurations of hyperparameters that have been tried are the following (in the Results chapter, only the best configuration is reported):

```
configurations = [  
'perplexity': 5, 'learning_rate': 50,  
'perplexity': 30, 'learning_rate': 200,  
'perplexity': 50, 'learning_rate': 500,  
'perplexity': 30, 'learning_rate': 'auto',  
'perplexity': 30, 'learning_rate': 500,  
'perplexity': 30, 'learning_rate': 50  
]
```

Data preprocessing is a fundamental step that significantly impacts the quality and performance of ML models. By improving data quality, enhancing model performance, reducing computational complexity, and ensuring compatibility with various algorithms, preprocessing transforms raw data into a format that is more suitable for analysis and

obtaining more accurate and efficient ML outcomes.

2.3.3. Clustering

Clustering is a class of unsupervised ML models that aims to subdivide the records of a dataset into homogeneous groups of observations, called clusters, so that observations belonging to one group are similar to one another and dissimilar from observations included in other groups. Clustering models serve diverse purposes. In certain scenarios, the clusters produced can offer insightful interpretations of the phenomena under study. Moreover, the division into clusters frequently acts as the first stage in a data mining effort, laying the groundwork for applying further methodologies customized for each cluster. Additionally, clustering facilitates exploratory data analysis by spotlighting outliers and pinpointing observations that could potentially constitute distinct clusters on their own, thereby aiding in dataset dimensionality reduction.

Clustering can be useful in the context of hearing data for several reasons. Hearing datasets often contain complex auditory signals with various patterns related to speech, noise, and other sound sources and clustering can help identify and categorize these patterns, making it easier to analyze and interpret the data. These algorithms can also extract relevant features from the auditory data, such as frequency components, temporal characteristics, and spectral features. These extracted features can then be used for various tasks such as speech recognition, sound classification, and acoustic scene analysis. Clustering can aid in segmenting long audio recordings into meaningful segments based on similarities in acoustic features. This segmentation can be useful for tasks such as speaker diarization, where different speakers need to be identified and separated in the audio stream. It can also be useful to differentiate between signal and noise components in the auditory data. By clustering similar noise patterns, it becomes possible to develop noise reduction techniques that selectively attenuate unwanted noise while preserving the desired signal components. Clustering techniques can be used to analyze individual hearing profiles and preferences. By clustering similar hearing profiles, personalized hearing aid settings can be developed to optimize sound amplification and improve the listening experience for individuals with hearing impairments. Overall, clustering in a hearing dataset can facilitate various tasks, ultimately advancing the understanding of auditory perception and improving hearing-related technologies and treatments.

The process of clustering consists in applying algorithms to unlabeled data and evaluate the best number of clusters to divide the dataset in. Here follows a brief explanation of the theory behind the most widely used algorithms for clustering: K-means, K- modes

and K-Prototypes .

Clustering techniques

K-Means: K-Means partitions unlabeled data into distinct groups by identifying similarities in features and recurrent patterns across the dataset. The k-Means clustering process is an iterative one, which entails dividing a set of n data points into K different clusters based on their similarity and average distance from the centroid of each formed subgroup. The workflow of the K-Means algorithm unfolds as follows:

1. Determine the value of K to specify the number of clusters ($n_clusters$) to be established.
2. Randomly select K points to serve as cluster centroids ($cluster_centers$).
3. Assign each data point to the nearest centroid, forming the predefined clusters based on their proximity.
4. Update the centroid position for each cluster.
5. Iterate over the process of reassigning each data point to the new closest centroid for each cluster.
6. If any reassignments occur, return to step 4; otherwise, proceed to step 7.
7. Conclude the process.

The primary objective is to partition the data in a manner that ensures points within the same cluster exhibit greater similarity to each other than to points in other clusters.

K-Means clustering offers several advantages. Firstly, it boasts simplicity in both understanding and implementation, making it computationally efficient and suitable for handling large datasets. Additionally, it demonstrates scalability, accommodating substantial volumes of data effectively.

However, K-Means does come with its drawbacks. Notably, its sensitivity to initialization can lead to varying outcomes based on the initial placement of cluster centroids. Determining the optimal number of clusters poses a challenge, often requiring domain expertise or trial-and-error approaches. The assumption of spherical and similar-sized clusters may not always reflect the true data structure, impacting the algorithm's performance. Outliers can significantly influence cluster assignments and centroid positions. Moreover, K-Means may struggle with non-convex clusters, often producing spherical results, and may not perform well when clusters vary significantly in size. Despite its

drawbacks, K-Means remains widely utilized and effective, particularly when the data structure aligns with the algorithm's assumptions.

K-Modes: K-Modes is a clustering algorithm used to group similar data points into clusters based on their categorical attributes. Unlike traditional clustering algorithms that use distance metrics, K-Modes works by identifying the modes or most frequent values within each cluster to determine its centroid. This method is ideal for clustering categorical data such as customer demographics, market segments, or survey responses. While K-Means uses mathematical measures (distance) to cluster continuous data (the lesser the distance, the more similar the data points are), K-Modes uses the dissimilarities (total mismatches) between the data points to cluster categorical data (the lesser the dissimilarities the more similar data points are). Here there are the principal steps of this algorithm [33]:

1. Pick K observations at random and use them as leaders/clusters.
2. Calculate the dissimilarities and assign each observation to its closest cluster.
3. Define new modes for the clusters.
4. Repeat 2–3 steps until there is no re-assignment required.

K-Modes presents several advantages: the resulting clusters are easily interpretable, comprising distinct categories rather than continuous values, which enhances understandability and actionability, particularly in domains where categorical attributes hold significance. K-Modes also exhibits robustness to outliers compared to numerical-focused algorithms like K-Means, as categorical data is less influenced by outliers, with each category representing a unique value. Moreover, it demonstrates scalability, efficiently handling large datasets with numerous categorical variables. The main limitations are: it is restricted to categorical data, rendering it unsuitable for datasets containing numerical attributes. Additionally, like K-Means, its performance can be sensitive to the initial placement of cluster centroids, impacting clustering outcomes. High cardinality categorical variables may pose challenges, increasing computational complexity and memory usage, potentially affecting performance and scalability.

K-Prototypes: As previously described, K-Means algorithm is not ideal for datasets with categorical variables due to its reliance on the Euclidean distance, which is suitable only for numerical data. Conversely, K-Modes offers the opportunity to deal with categorical variables, but presents limitations when dealing with high-cardinality categorical variables and mixed data types. In response, the K-Prototypes algorithm emerged. This method calculates distance between numerical features using Euclidean distance akin to

K-Means, while also considering the distance between categorical features based on the number of matching categories. First introduced by Huang in 1998 [34], K-Prototypes offers a solution tailored for datasets encompassing both numerical and categorical attributes. Here are the steps involved in the K-Prototypes algorithm:

1. Initialization: Initialize the number of clusters (k) and randomly select k data points as the initial cluster centroids. For numerical attributes, calculate the mean of each attribute for each cluster. For categorical attributes, calculate the mode (most frequent category) for each attribute for each cluster.
2. Assignment Step: Assign each data point to the nearest cluster centroid based on a distance metric that combines both numerical and categorical attributes. One commonly used distance metric for numerical attributes is the Euclidean distance, while for categorical attributes, the Hamming distance or another appropriate dissimilarity measure is used. The overall distance between a data point and a cluster centroid is the sum of the distances for numerical attributes and categorical attributes.
3. Update Step: Recalculate the cluster centroids based on the newly assigned data points. For numerical attributes, update the centroids by calculating the mean of each attribute for the data points in each cluster. For categorical attributes, update the centroids by calculating the mode for each attribute for the data points in each cluster.
4. Convergence Check: Repeat the assignment and update steps until convergence criteria are met. Convergence may be defined by a maximum number of iterations or when the cluster assignments and centroids no longer change significantly between iterations.
5. Finalization: Once convergence is achieved, the algorithm stops, and the final cluster assignments and centroids are obtained. Each data point is assigned to the cluster with the nearest centroid based on the combined distance metric.

The K-Prototypes algorithm offers several advantages. It effectively combines numerical and categorical attributes to form clusters that best represent the underlying structure of the dataset. By considering both types of attributes, it can handle diverse datasets with mixed data types more accurately than traditional clustering algorithms that only handle numerical data, making it versatile for real-world datasets with diverse attribute types.

Resulting clusters are easily interpretable, aiding in decision-making based on cluster

characteristics. Additionally, K-Prototypes demonstrates robustness to outliers, similar to K-Modes, as categorical attributes are less influenced by outliers, resulting in more stable cluster assignments. Moreover, it provides flexibility in data representation, allowing for a comprehensive analysis of the dataset and potentially revealing more nuanced insights.

However, K-Prototypes has some disadvantages. Similar to K-Means and K-Modes, its performance can be sensitive to initialization, impacting clustering outcomes. Handling both numerical and categorical attributes increases computational complexity, leading to longer runtimes and higher memory usage, particularly for larger datasets. Choosing the optimal number of clusters (K) can be challenging, requiring careful consideration and potentially involving trial and error or domain expertise. Furthermore, interpreting clusters with mixed data types may pose challenges compared to clusters with a single data type, necessitating additional analysis and visualization techniques to understand relationships between attributes within each cluster.

Given the presence of categorical and numerical data in the datasets that have been used in this thesis, K-Prototypes is the algorithm that has been utilized for the clustering step.

Clustering evaluation

The choice of the number of clusters is made using the so-called elbow method. It involves running a clustering algorithm, for a range of different cluster numbers (k) and calculating the corresponding sum of squared distances (SSD) from each point to its assigned cluster center. This SSD is also known as the within-cluster sum of squares (WCSS). The idea is to plot the WCSS against the number of clusters and observe the resulting curve. Initially, the WCSS decreases sharply with an increasing number of clusters, but after a certain point, the rate of decrease slows down and forms an "elbow" shape. The point where this elbow occurs is considered the optimal number of clusters [35]. An important issue in clustering analysis is evaluating the quality of the clusters formed, and it is crucial to ensure that the clustering algorithm effectively captures the underlying structure of the data. Three commonly used indices for this purpose are the Silhouette score, Davies-Bouldin index, and Calinski-Harabasz index.

Silhouette Score: The Silhouette score [36] measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The Silhouette coefficient for each sample is calculated using the formula:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2.1)$$

where: $a(i)$ is the average distance between the sample i and all other points in the same cluster. $b(i)$ is the minimum average distance from the sample i to all points in the nearest cluster that i is not a part of. The Silhouette score ranges from -1 to 1: A score close to 1 indicates that the sample is well-matched to its own cluster and poorly matched to neighboring clusters. A score close to 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters. A score close to -1 indicates that the sample has been assigned to the wrong cluster. A high average Silhouette score indicates well-defined and distinct clusters, suggesting that the clustering algorithm has performed effectively.

Davies-Bouldin Index: The Davies-Bouldin index [37] evaluates the average similarity ratio of each cluster with its most similar cluster. It is calculated using the formula:

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{s_i + s_j}{d_{ij}} \right) \quad (2.2)$$

where: k is the number of clusters. s_i is the average distance between each point in cluster i and the centroid of i . d_{ij} is the distance between the centroids of clusters i and j . The Davies-Bouldin index ranges from 0 to ∞ , with lower values indicating better clustering quality. A low Davies-Bouldin index means that clusters are compact and well-separated from each other.

Calinski-Harabasz Index: The Calinski-Harabasz index [38], also known as the Variance Ratio Criterion, measures the ratio of the sum of between-cluster dispersion and within-cluster dispersion. It is defined as:

$$CH = \frac{\text{Tr}(B_k)/(k-1)}{\text{Tr}(W_k)/(n-k)} \quad (2.3)$$

where: $\text{Tr}(B_k)$ is the trace of the between-cluster dispersion matrix. $\text{Tr}(W_k)$ is the trace of the within-cluster dispersion matrix. k is the number of clusters. n is the total number of samples. The Calinski-Harabasz index is higher for better-defined clusters. A high Calinski-Harabasz index indicates that the clusters are dense and well-separated from each other, suggesting an effective clustering.

Centroids and *medoids* are key points in the context of clustering, used to represent clusters and facilitate the interpretation of results. Centroids are the average points of a cluster in feature space, calculated as the arithmetic mean of the coordinates of all points in the cluster. They represent the 'heart' or geometric center of the cluster. On the other hand, medoids are the points in the cluster that minimize the sum of distances

from other points in the cluster. They are the most representative points of the cluster in terms of similarity to other points. In summary, while centroids provide an average representation of the cluster, medoids are actual points in the dataset that best represent the cluster's structure. Both are useful for understanding the distribution and shape of clusters and can be used to interpret and compare clustering results.

2.3.4. Classification

After clustering, it is also very common to apply classification algorithms to categorize data points into the clusters that have been selected. This approach enhances significantly the analytical outcome and refines the understanding by assigning labels or categories to these clusters.

Classification is a supervised machine learning technique used to categorize data points into predefined classes or labels based on their features. It involves training a model to categorize data points into predefined classes, enabling the prediction of outcomes or the identification of patterns within datasets. In the context of hearing data, classification can be immensely useful for tasks such as identifying different types of auditory signals, classifying hearing impairments, or predicting patient outcomes based on clinical data. A critical aspect of building a reliable classification model is the division of data into a training set and a test set.

The training set is used to train the model, enabling it to learn the relationships and patterns in the data. The test set, on the other hand, is used to evaluate the model's performance on unseen data. This division is crucial because it allows for an unbiased assessment of the model's ability to generalize to new, unseen data, which is a key indicator of its real-world applicability. Typically, the data is randomly split into approximately 70-80% for training and 20-30% for testing. This ensures that the model has sufficient data to learn from while also having a separate set of data to validate its predictions. Other ways of splitting the dataset are used based on the context.

By assessing the model on the test set, it can be identified any issues such as overfitting, where the model becomes overly attuned to the training data's idiosyncrasies rather than capturing generalizable patterns. In the realm of hearing data analysis, overfitting could manifest as the model mistakenly learning from irrelevant noise or outlier data points, compromising its ability to accurately classify new, unseen data.

To mitigate overfitting, techniques such as cross-validation, regularization, and feature selection are commonly employed. Cross-validation involves splitting the training data into

multiple subsets, training the model on some subsets, and validating it on the remaining subsets, thus ensuring the model's robustness across different data samples. Regularization techniques add a penalty for more complex models to prevent them from fitting the noise in the training data. Feature selection involves choosing the most relevant features for the model, reducing the dimensionality of the data, and minimizing the risk of overfitting. These practices ensure that the model learns meaningful patterns without being overly influenced by noise or irrelevant features, thereby improving its generalization performance and making it more reliable for practical applications in hearing data analysis.

In the following pages, the theory behind the three main algorithms that have been exploited during the work of the present thesis are presented.

Random Forest

The Random Forest algorithm is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) of the individual trees. It is known for its robustness and ability to handle large datasets with high dimensionality. Random Forest combines multiple decision trees to improve classification accuracy. Each tree is constructed from a different bootstrap sample of the training data, and features are randomly selected at each split. Each tree in the forest is trained on a different subset of the training data, created by randomly sampling with replacement (bootstrap sampling). This helps in reducing variance and preventing overfitting.

At each split in the decision tree, a random subset of features is chosen. This ensures that the trees are diverse and reduces the correlation between individual trees. For classification, each tree in the forest outputs a class prediction. The final prediction is made by taking the majority vote among all the trees. There are several key parameters:

- Number of Trees ($n_estimators$): This parameter specifies the number of decision trees in the forest. Increasing the number of trees generally improves performance but also increases computational cost.
- Maximum Depth of Trees (max_depth): This parameter limits the depth of the trees. Limiting the depth can prevent overfitting, especially in noisy datasets.
- Minimum Samples Split ($min_samples_split$): The minimum number of samples required to split an internal node. Increasing this value can lead to more generalized trees.

- Minimum Samples Leaf (*min_samples_leaf*): The minimum number of samples required to be at a leaf node. Setting this parameter helps in smoothing the model, especially for regression tasks.
- Number of Features (*max_features*): The number of features to consider when looking for the best split. It can be set to a specific number or to "sqrt" (square root of the total number of features) or "log2" (logarithm base 2 of the total number of features).
- Bootstrap: A boolean parameter indicating whether bootstrap samples are used when building trees. If set to False, the whole dataset is used to build each tree.

The parameter grid that has been utilized for all the analyses of the present thesis is the following:

```
param_grid = {
    'n_estimators': [20, 25, 30, 50, 100, 150],
    'max_depth': [None, 5, 10, 15, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
```

Support Vector Machine

The Support Vector Machine (SVM) is a widely applied supervised algorithm, due to its effectiveness in high-dimensional spaces and its ability to handle non-linear data through kernel functions. SVM aims to find the optimal hyperplane that separates data points of different classes with the maximum margin. The margin is defined as the distance between the hyperplane and the closest data points from each class, known as support vectors.

Linear SVM: For linearly separable data, SVM finds a hyperplane $w \cdot x + b = 0$ that maximizes the margin. The optimization problem is formulated as:

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad (2.4)$$

subject to $y_i(w \cdot x_i + b) \geq 1$ for all i , where y_i are the class labels, x_i are the data points, w is the weight vector, and b is the bias.

Non-Linear SVM: For non-linearly separable data, SVM uses kernel functions to transform the data into a higher-dimensional space where a linear separator can be found.

Common kernel functions include:

Linear Kernel: $K(x_i, x_j) = x_i \cdot x_j$

Polynomial Kernel: $K(x_i, x_j) = (x_i \cdot x_j + c)^d$

Radial Basis Function (RBF) Kernel: $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$

Sigmoid Kernel: $K(x_i, x_j) = \tanh(\alpha x_i \cdot x_j + c)$

The choice of kernel function (e.g., linear, polynomial, RBF, sigmoid) determines how the data is transformed. The kernel function can significantly impact the SVM's performance, especially in non-linear cases. The parameter C controls the trade-off between achieving a low training error and a low testing error (generalization). A small C allows more misclassifications (higher bias), while a large C aims to classify all training examples correctly (higher variance). The SVM algorithm is particularly efficient when the number of dimensions exceeds the number of samples, it is versatile because of the different number of kernel functions that can be specified and it is robust to overfitting, particularly when using an appropriate value of the parameter C . Some disadvantages of this approach are the intensive computational effort required, especially for large datasets, and the sensitivity to noisy data because outliers can affect the hyperplane's position.

The parameter grid used for SVM throughout all the analyses is the following:

- **C:** 0.001, 0.01, 0.1, 1, 10, 100
- **kernel:** linear, rbf, poly, sigmoid

k-Nearest Neighbors Classifier

The k-Nearest Neighbors (KNN) algorithm is a simple, yet effective, non-parametric classification algorithm that classifies a data point based on the majority class among its k-nearest neighbors in the feature space. For a given test sample, the algorithm calculates the distance between the test sample and all training samples.

The algorithm identifies the k-nearest neighbors to the test sample based on the calculated distances. The class labels of the k-nearest neighbors are examined, and the most frequent class label among them is assigned to the test sample. The key parameters of the algorithm are:

- *Number of Neighbors (k):* The parameter k defines the number of nearest neighbors to consider when making a classification decision. The choice of k can significantly impact the algorithm's performance: Small k (e.g., $k = 1$) means that the model can

be sensitive to noise and may overfit the training data, instead, a large k signifies that the model becomes more generalized but may underfit the data.

- *Distance Metric*: The distance metric determines how the distance between data points is computed. Commonly used metrics include:
 - Euclidean Distance: $d(p, q) = \sum_{i=1}^n (p_i - q_i)^2$
 - Manhattan Distance: $d(p, q) = \sum_{i=1}^n |p_i - q_i|$
 - Minkowski Distance: A generalized form that includes both Euclidean and Manhattan distances.
- *Weighting*: The KNN algorithm can assign weights to the neighbors based on their distance from the test sample. Common weighting schemes include:
 - Uniform Weights: All neighbors are equally weighted.
 - Distance Weights: Closer neighbors are given higher weights, which can be defined as $1/d(p, q)$ where $d(p, q)$ is the distance.

The main advantages of this algorithm are its simplicity and flexibility, because it can be applied to various types of data without the need for parameter estimation, and its non-parametric nature, so the no need of assumptions about the underlying data distribution.

The parameter grid used in the thesis for KNN is the following:

- **n_neighbors**: 3, 5, 7, 9

For each algorithm, in the Results chapter are reported the best combination of parameters obtained from the grids reported here. Tuning of the parameters has been performed through grid search and cross-validation.

Cross validation and other performance measures

Cross-validation is a resampling technique used to assess the performance of machine learning models and to estimate how well they will generalize to unseen data. It is commonly used to evaluate the robustness and reliability of predictive models, especially when the dataset is limited or prone to overfitting.

The cross-validation technique that has been exploited in this thesis is the *k-fold cross-validation*. In k-fold cross-validation, the original dataset is randomly partitioned into k equal-sized folds. The model is trained k times, each time using $k-1$ folds as training data and the remaining fold as the validation set. This process is repeated k times, with

each fold used exactly once as the validation set. The performance metric (e.g., accuracy, precision, recall) is then averaged over all k iterations to obtain a single estimate of model performance. Moreover, the best combination of parameters found through cross-validation can then be tested and evaluated on an external test set composed of unseen data.

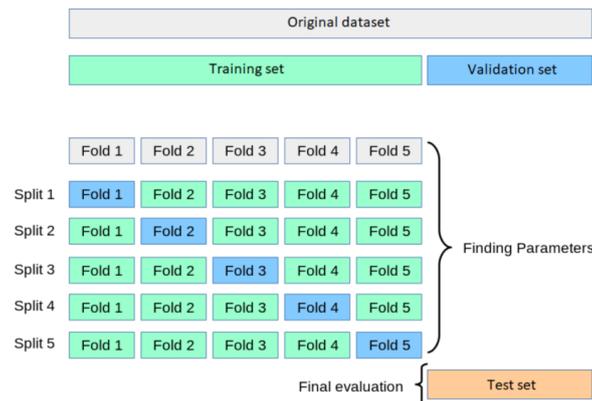


Figure 2.13: Illustration of k -fold cross validation with extension for hold outs (scikit-learn, 2021)[39]

Sometimes it is also useful to execute a *nested CV*, for example when the dataset has a small size. This may lead to high variability in model's performance, as it was the case in some of the analyses of the present thesis. For K (=number of folds) times the entire dataset is randomly split into external training and external test sets (for example using train and test split: 80-20%). Each external training fold is divided into internal training and internal validation sets. Internal cross-validation is performed in order to find the best parameters. The model with the best parameters is then trained on the external training set (80% of the data) and evaluated on the external test set (20% of the data). This approach provides a more robust estimate of the model's performance by reducing the variability due to random train-test splits and ensures that the chosen hyperparameters are well-optimized for the entire dataset. This method has been applied in one of the splitting methods for classification of the data (random splitting with training = 80%, test = 20%).

Another important issue about ML models is their performance and the metrics used to evaluate it. Here is a brief description of the performance metrics that have been considered during the development of the present thesis. Each one of these metrics can be computed as micro-average or macro-average. Macro-averaging gives equal weight to each class, while micro-averaging gives equal weight to each instance.

Precision: Precision measures the fraction of positive instances correctly identified by the model out of all instances that the model classified as positive. It is calculated as the ratio of true positives (correctly classified positive instances) to the sum of true positives and false positives (negative instances incorrectly classified as positive). In other words, precision indicates how precise the model is in classifying positive instances.

Recall (or Sensitivity): Recall measures the fraction of positive instances that were correctly identified by the model compared to the total positive instances present in the data. It is calculated as the ratio of true positives to the sum of true positives and false negatives (positive instances incorrectly classified as negative). In other words, recall indicates how well the model can find all positive instances.

F1-score: The F1-score is the harmonic mean of precision and recall. It is useful when wanting to consider both precision and recall in a single measure. The F1-score is particularly useful when classes are imbalanced, i.e., when there are many more instances of one class than the other. It is calculated as $2 \times (\textit{precision} \times \textit{recall}) / (\textit{precision} + \textit{recall})$.

Support: Support is the number of samples for each class present in the data. It is useful for understanding the distribution of classes in the dataset and if there are any class imbalance issues.

Accuracy: Accuracy measures the overall correctness of the model's predictions by calculating the fraction of correctly classified instances out of all instances in the dataset. It is calculated as the ratio of the number of true positives and true negatives (correctly classified instances) to the total number of instances in the dataset. In other words, accuracy indicates the proportion of instances that the model classified correctly, regardless of their class.

Macro-averaging computes metrics (e.g. precision, recall) for each class and then averages these values across all classes to derive a final score. This method treats every class equally, disregarding the class size. In contrast, micro-averaging aggregates true positives, false positives, and false negatives across all classes to compute metrics based on the total counts, giving equal importance to each instance. The choice between macro- and micro-averaging depends on the problem at hand and the significance of each class or instance. Macro-averaging is beneficial when all classes hold equal importance and an overall performance assessment across them is desired. It also helps in handling imbalanced datasets by ensuring each class contributes equally. However, macro-averaging can sometimes obscure true performance. For instance, poor performance in a small, less significant class can still impact the overall score equally, potentially masking critical issues. Conversely, excellent performance in majority classes might overshadow poor

performance in minority classes. Micro-averaging, on the other hand, is suitable when wanting to emphasize the total number of misclassifications across the dataset, akin to accuracy. Yet, it can exaggerate the performance of the dominant class, leading to inflated scores and potentially overlooking poor performance in smaller classes. The choice of metric should consider factors like class distribution, their importance, and specific evaluation goals to provide a comprehensive assessment. [40]

With this chapter, it has been provided a comprehensive overview of the test battery, the clustering and classification techniques employed in the analysis of hearing data. By detailing the methodologies utilized, including K-Means, K-Prototypes, and other relevant algorithms, this chapter lays the groundwork for the subsequent analysis and interpretation of the results. Through rigorous application of these methods to hearing data, the aim has been to uncover valuable insights that contribute to the understanding of auditory perception and pave the way for advancements in the field of hearing science.

3 | Results

This chapter presents the principal findings of the study, detailing the analysis of the collected data and the interpretation of the results in relation to the research questions and hypotheses. The following sections will explore the key outcomes, highlighting significant trends and patterns observed throughout the study.

3.1. Summary of dataset - phase 1:

The dataset acquired during the first phase of data acquisition (section 2.2.1) includes 33 Italian subjects collected during lab testing and opportunistic hearing screening initiatives. Among the 33 subjects that have been acquired, 16 are women and 17 are men, with age from 22 to 64 years (age median of about 27 years). Of these 33 subjects, 4 subjects have shown sign of slight or moderate hearing loss, according to their PTA value. Only some of the 33 subjects tested performed the test on both ears, for a total of 53 new records. Initially, an exploratory analysis of the existing dataset and this new subset was conducted to better understand the data for more effective clustering and classification.

For each record, the complete set of features from Whisper, DST, and the risk factor questionnaire, listed in Figure 2.8, was extracted and analysed.

3.2. Summary of dataset - phase 2:

The dataset associated with the second phase of the protocol (section 2.2.2) was acquired during the stay at the University of Oldenburg and consists of 34 ears tested from 17 normal-hearing subjects under the age of 40. Specifically, 7 were women and 10 were men. One subject took the tests in Italian, while the others took the tests in German language. Only one subject, German, took the DTT for one ear, while all subjects took all tests for both ears. Comparative analyses of the 3 speech-in-noise tests used were performed on this dataset to observe possible correlations, similarities, or differences. The features extracted are the same for the Whisper test and the DST, with the addition of features from the DTT and OLSA tests in Figure 2.9.

3.3. Analysis of the whole dataset:

In this section, the main results are shown for the dataset containing Whisper, DST and risk factor features. The overall dataset comprises 434 subjects. Overall, 278 subjects performed the test for only one ear, while 156 subjects performed it for both ears. Therefore, the dataset comprises 590 records in total, each one related to a single ear tested.

Specifically, 375 records are related to female subjects, whereas the remaining 215 records are related to male subjects. The dataset comprises subjects aged between 17 and 89 years, and includes individuals speaking over 12 different languages. In terms of audiometric thresholds, 403 records showed $PTA < 20\text{dB}$, while PTA between 20dB and 35dB was found in 113 records. Finally, the records with $PTA > 35\text{dB}$ were 74.

Whisper and PTA measurements are available for each record of the dataset (i.e., since the first subject acquired in 2018). DST measurements and most of the risk factors (15 out of 18) were included in a more recent phase, hence they are present just for 266 records.

3.3.1. Characterization of test variables:

In this section, the distributions of the key features within the dataset will be examined. This process will provide an important overview of the demographic and audiologic characteristics of the subjects included in the study.

The distribution of subjects age in the dataset has been observed in order to explore the range of represented ages, with particular attention to the most represented age groups.

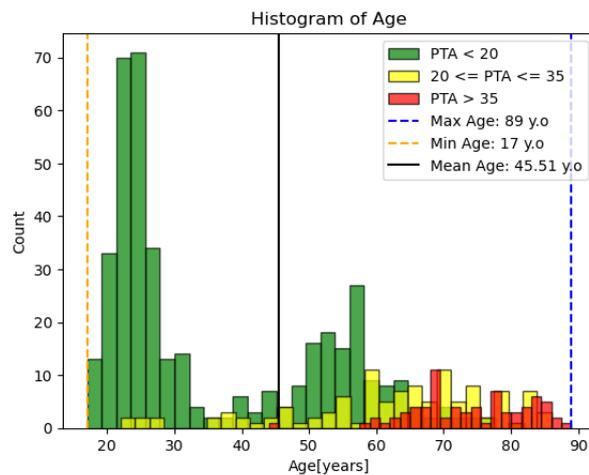


Figure 3.1: Age histogram with highlighted PTA bands; minimum, maximum and mean values are reported.

As highlighted in Figure 3.1, the number of subjects between the ages of 30 and 50 appears to be lower compared to other age groups, especially the younger one. This demonstrates the decision, for the first phase of the protocol, to seek and favor subjects within that age range. In accordance with audiological guidelines and accepted standards by the WHO, hearing levels have been divided into three categories according to the PTA value (obtained as the mean of frequencies ranging from 500 to 4000 Hz) to reflect common practice in diagnosing and classifying hearing disorders. Specifically, tested ears can be characterized by normal hearing (PTA <20dB green), mild hearing loss (PTA between 20 and 35dB), and moderate hearing loss (PTA >35dB). These intervals are widely recognized and adopted in audiology to describe the severity of hearing loss and guide clinical decisions. From the graph, it can already be observed a distribution of PTA values above 35 dB (thus, according to the guidelines, corresponding to a significant hearing loss) for the older groups, as usually expected with increasing age. It can also be noted that there is a good number of subjects between 50 and 60 years old that present values of PTA lower than 20 dB.

The histogram displaying the occurrences of PTA values is shown in Figure 3.2. The mean value is below 20 dB, indicating that, on average, the subjects can be considered to be normal-hearing (403 records show values lower than 20 dB). However, 113 records present PTA values between 20dB and 35dB, indicating slight/mild hearing loss. Moreover, there are several records with high PTA values (> 35dB, precisely 74 records), highlighting that some of the tested population also exhibits relatively high levels of hearing impairment.

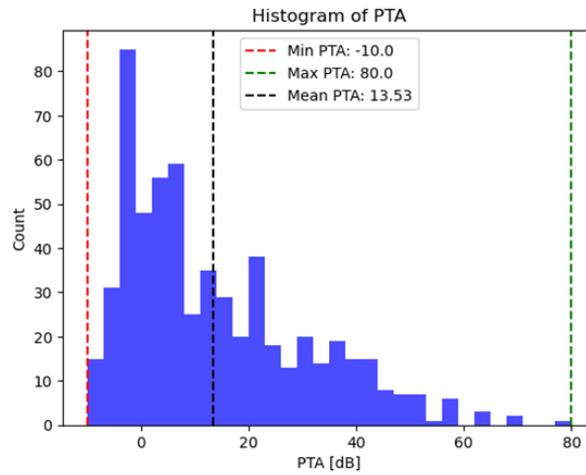


Figure 3.2: Histogram of PTA

Observing the scatterplot between PTA and Age (Figure 3.3), as expected, an increase in PTA values with advancing age is evident, suggesting that aging-related factors influence hearing loss. Nevertheless, a greater variability in PTA values is also observed, especially in the age range starting from 45-50 years, indicating that different individuals of the same age can have significantly different PTA values.

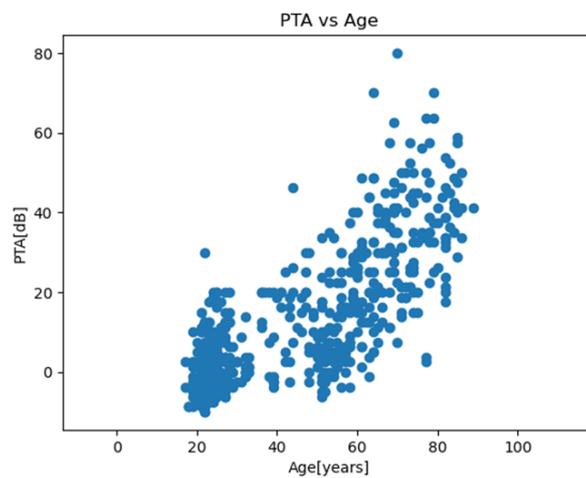


Figure 3.3: Scatterplot of PTA and Age

Subsequently, the distribution of the SRT, i.e., the main output of the Whisper test, was observed (Figure 3.4). The graph highlights the previously explained PTA ranges and reports the minimum, maximum, and mean values for the entire dataset.

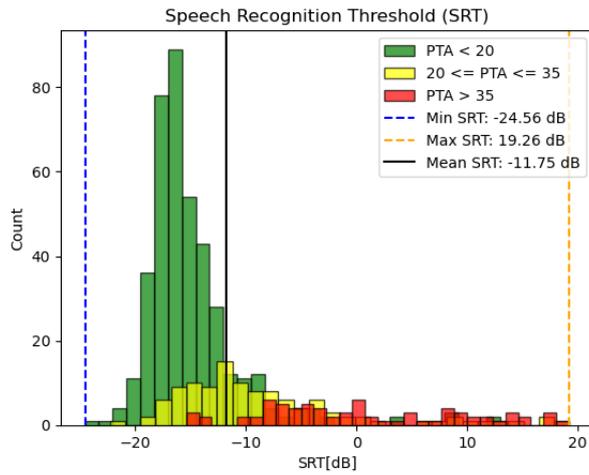


Figure 3.4: Histogram of SRT with highlighted PTA values.

Very negative values of SRT correspond to excellent speech recognition ability in noise, while very high values correspond to difficulties in recognizing speech in noisy conditions. From the graph, it can be observed that the distribution is non-normal (median value = -14.71 dB), with a distribution skewed towards negative values. From the division based on PTA class, it can be observed that high SRT values correspond to high (worse) PTA values, while low SRT values correspond to low (better) PTA values. This trend suggests a relationship between hearing loss and the ability to recognize spoken language in the presence of background noise (SRT). This relationship is confirmed by the value of the correlation coefficient between PTA and SRT (value = 0.679).

Figures 3.5 and 3.6 report the histogram of total time of the Whisper test and the scatterplot of total time and age, respectively.

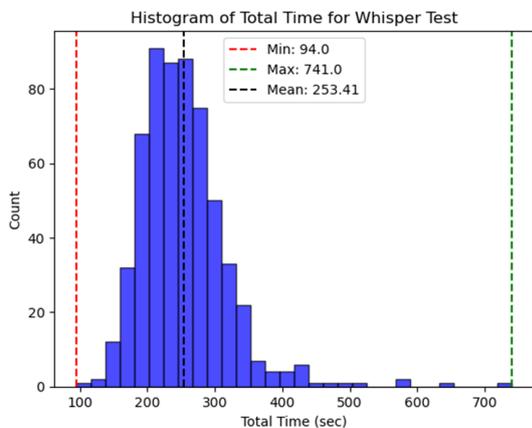


Figure 3.5: Histogram of Total Time for Whisper Test.

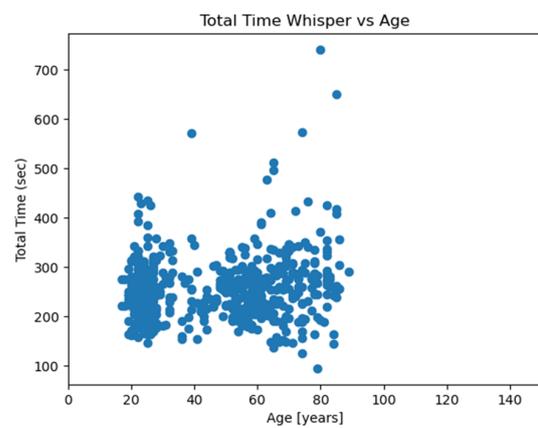


Figure 3.6: Scatterplot of Total Time for Whisper test VS Age.

The histogram in Figure 3.5 shows that most of the data points are between 100 and 400 seconds, with only a few cases exceeding 700 seconds. The scatterplot in Figure 3.6 shows a sparse distribution of data with some dense regions, but it does not indicate a clear trend or correlation between age and the total time for the Whisper test. It is possible that other factors influence the total time for the test beyond age, or that significant individual variability is present.

Moving on to the collected features related to the cognitive test, Figure 3.7 shows the histogram of the Digit Span Score, representing the longest sequence each subject correctly recalled in the DST. The majority of the scores range between 5 and 6, as indicated by the mean value. Value 0 is attributed to subjects that didn't get correctly the first sequence (of length 3) for both the two trials given (and they were only 3).

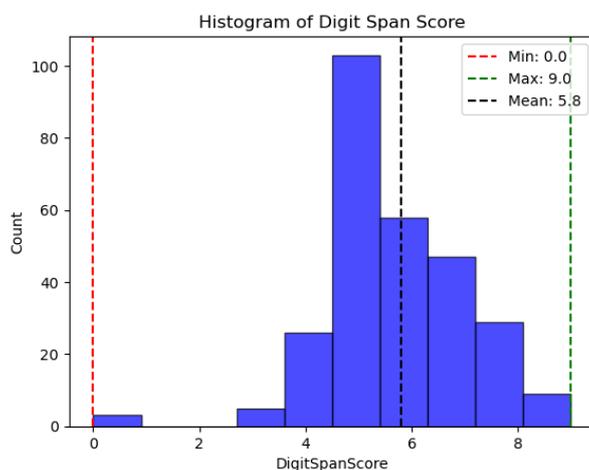


Figure 3.7: Histogram of Digit Span Score.

The following graphs show some DST-related variables that have been reprocessed to obtain additional information and improve data interpretation (more in Figure 2.10 in Summary of extracted features section). Firstly, the variable *avgSingleDigitTimes* was examined. This variable indicates the average time taken to enter each single digit of the current sequence during the DST. The first number in the list indicates the average time taken to type the first digit of the sequence by that user, the second indicates the average time taken to type the second digit, and so on. It has been decided to observe the mean value of the variable for each subject, i.e., the single digit average typing time across all digits. Therefore, the average time for each digit was calculated for each subject, resulting in the new variable *avgSingleDigitTimes_mean*.

For example:

$avgSingleDigitTimes = [0.691, 0.906, 0.837, 0.650, 1.25, 0.631]$, meaning that the mean digit time for the first digit of the sequences is 0.691 sec, while the second digit has a mean digit time of 0.906 sec, and so on. The new variable has value 0.827, computed as the sum of all the elements of the list, divided by the number of the elements (e.g. the mean), and represents the average typing time in the DST for all digits for the considered subject.

The histogram of this new variable is shown in Figure 3.8.

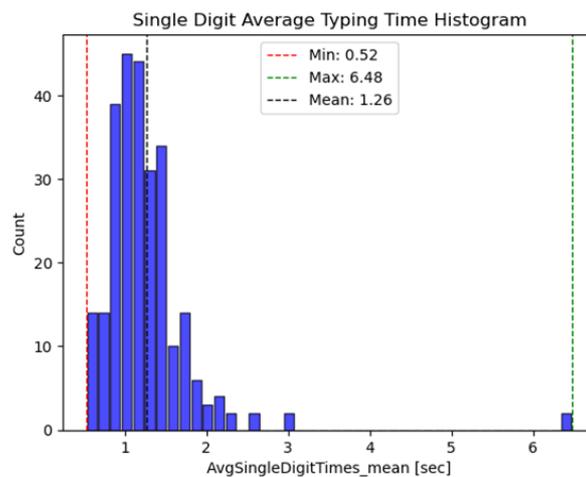


Figure 3.8: Histogram of Single Digit Average Typing Time

The observed values show an average slightly above one second, indicating a relatively high responsiveness. Although there are some slightly higher values, there is just an isolated case exceeding six seconds.

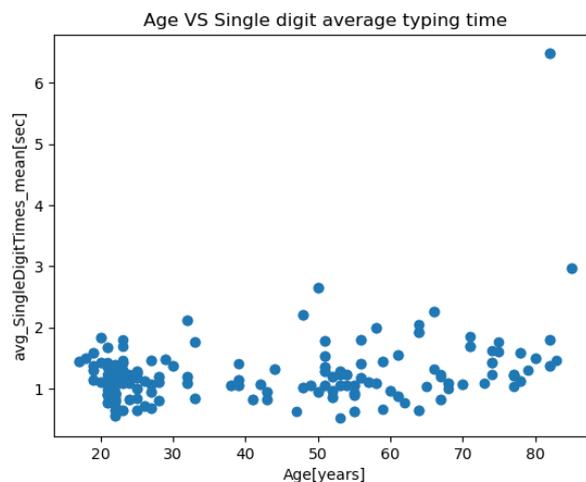


Figure 3.9: Scatterplot of Age and $avgSingleDigitTimes_mean$

To investigate the correlation between single digit average typing time and age of the tested subjects, a scatter plot is shown in Figure 3.9. The displayed graph represents the relationship between age (on the x-axis) and *avgSingleDigitTimes_mean* on the y-axis. As previously mentioned, the observed points are mostly below 2 seconds of average typing time, indicating generally high responsiveness. Younger subjects (between 20 and 30 years old) tend to have more concentrated and faster average typing times (primarily between 0.5 and 1.5 seconds). As age increases (e.g., above 40 years), there is greater variability in typing times, with some subjects exhibiting significantly higher times. Indeed, the scatterplot reveals that the subject with average typing time exceeding six seconds is over 80 years old. The correlation coefficient (of 0.30) is relatively low and correlation between age and *avgSingleDigitTimes_mean* is not really evident, indicating that the increase in average single-digit typing time with age is not strongly pronounced.

It was then observed the total time required to enter the first sequence of the DST (of 3 digits), which is contained in the variable *responseTimes*. This variable contains the total response time for each proposed sequence (in seconds), so by extracting the first value of the variable for each subject, the total time needed to enter the first sequence of the test is obtained (and called 'response_first_numbers'). It was decided to extract the value corresponding to the first sequence only to ensure that there was a value for every subject tested, even for those with a DSS of 0, who finished the test early.

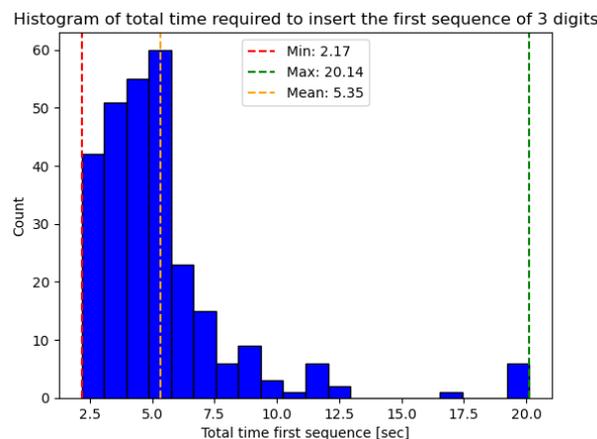


Figure 3.10: Total time required to insert the first sequence of 3 digits.

The histogram in Figure 3.10 shows that the tested subjects needed an average of 5.3 seconds to enter the first DST sequence. Only fewer cases presented high values for this variable, up to a maximum of about 20 seconds (they are 6 cases, and in terms of DSS two of them have 5.0, two have 6.0 and two have 8.0. This could lead to the consideration of some sort of influence between the time needed to insert the first sequence and the

overall performance on the test, but the presence of values 6 and 8 that are in average and above the average does not really support that).

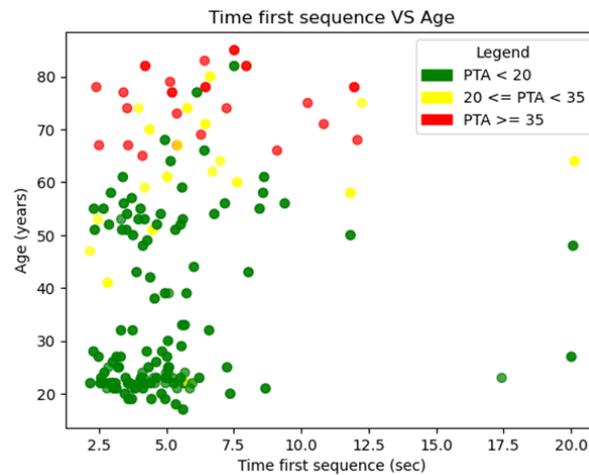


Figure 3.11: Scatterplot of Time for first sequence and Age, highlighted with PTA values.

The scatterplot in Figure 3.11 depicts the typing time of the first sequence against age, with coloration based on the individual subject's PTA value.

It is evident that higher ages are generally associated with higher PTA values, but not necessarily with longer typing times. On the contrary, the scatterplot shows densely populated red points on the left side of the graph, corresponding to rather short typing times. The isolated cases of prolonged times (i.e., higher than 15 seconds) have PTA values below 20dB and age below 70 years, perhaps due to the possible presence of cognitive difficulties not associated with hearing impairments or difficulties in interacting with the testing platform.

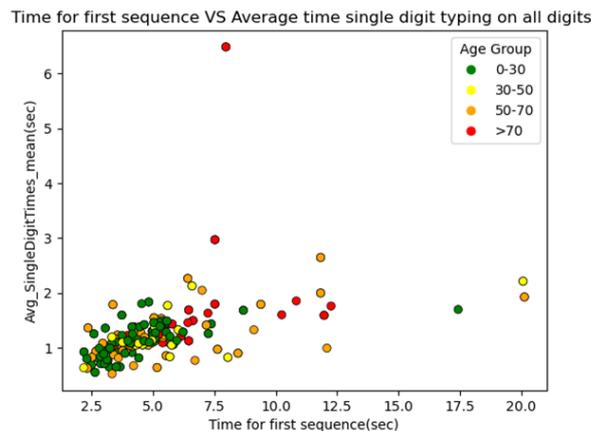
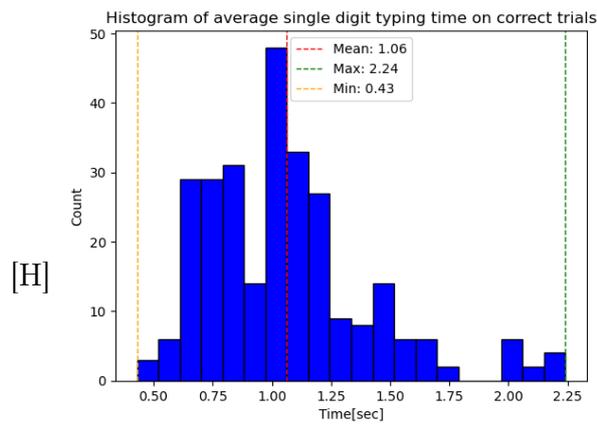


Figure 3.12: Scatterplot of Time for the first sequence of DST and the Average time single digit typing on all digits, highlighted with Age.

Figure 3.12 shows the scatterplot of typing time for the first DST sequence (first element of `avgSingleDigitTimes`) and the average typing time for each digit (`avgSingleDigitTimes_mean`). Specifically, we can observe that subjects exhibiting a high typing time value on the first sequence of 3 digits are in the age groups of 30-50 and 50-70, except for one subject who is younger. In general, the population under 30 years old is clustered at the bottom left of the plot, showing relatively fast typing time values for both the first sequence and individual digits, while both variables increase with age, as expected. Notably, there are several subjects in the older age group who still perform well in terms of timing.

Then, the average typing times for individual digits are observed separately for correct trials and incorrect trials. The variable `binaryResults` contains binary results for each proposed sequence, i.e., 1 if the sequence is guessed correctly in full, 0 if the sequence is partially or entirely incorrect. Therefore, to find the correct trials for each individual subject, simply take the values of 1 from the `binaryResults` variable and mark their positions. Then, corresponding values from the `avgSingleDigitTimes` variable are retrieved, obtaining the variable `correct_avg`. Finally, the average of these values yields the average typing time for the individual digit on correct trials for that subject. The same procedure is followed for incorrect trials, but by searching for the value 0 in `binaryResults` (variable named then `incorrect_avg`).

The two histograms in Figure 3.13 and 3.14 display a rather similar mean value, slightly higher for the incorrect trials, while the maximum value for the incorrect trials exceeds that of the correct trials by more than 1 second, suggesting that in the case of incorrectly entered sequences, typing times are also higher.



[H]

Figure 3.13: Histogram of Average Single Digit Typing Time on Correct trials.

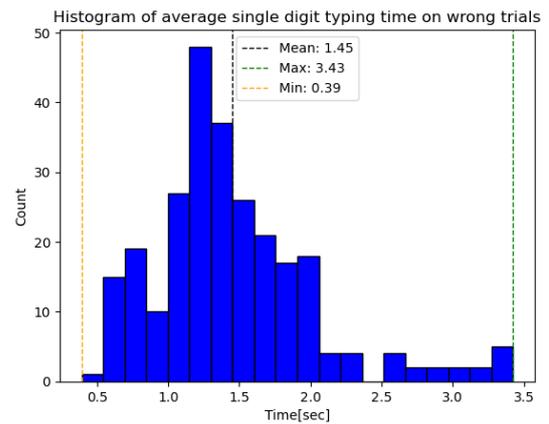


Figure 3.14: Histogram of Average Single Digit Typing Time on Wrong trials.

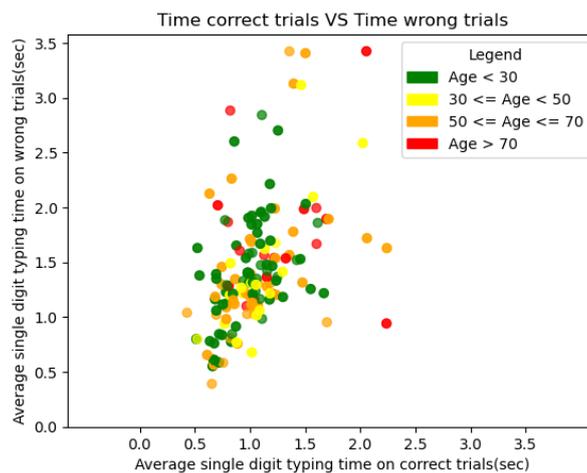


Figure 3.15: Scatterplot of correct trials VS wrong trials average single digit typing time highlighted by Age.

The association between the average single digit typing time in correct and incorrect trials is also investigated based on different age groups. Particularly, Figure 3.15 shows that those who exhibit longer typing times in both correct and incorrect trials (i.e., exceeding 3 seconds) are primarily in the adult age group (over 50 years old), although within the same age group there are subjects with relatively low values too.

Finally, the dataset can be characterized in terms of the answers that were given in the risk factor questionnaire. The following figure (3.16) displays the frequency of responses 'Yes' and 'No' for the risk factors where the response was of binary type. Additionally, other variables related to the risk factor questionnaire but with more than two possible answers have the following numerosity (note: not all the records of the dataset have

same number of risk factor answers because it has been added later in time, but all of the participants that have participated in the study during this thesis underwent the questionnaire):

- **Education:** Master (72), Bachelors (45), High (Secondary school) (131), Middle (Primary school) (32)
- **Tinnitus:** No (122), Yes Both (12), Yes Left (11), Yes Right (4)
- **ExerciseLevel:** Low (39), Medium (42), High (68)

Feature	Yes	No
cardio	34	246
cholesterol	60	220
covid	121	53
depression	36	244
diabetes	12	268
drink	52	228
ear_infections	73	207
family_history_HL	86	194
head_trauma	36	244
meningitis	0	280
obesity	16	264
smoker	105	175
stroke	3	277
high_volume_exposure	118	162
work_exposure	29	251

Figure 3.16: Number of Yes/No responses for each binary risk factor from the questionnaire.

3.3.2. Clustering using only Whisper features:

In this section, the results of clustering performed on the dataset where only the main features related to the Whisper test were selected are presented. Since all the records of the dataset executed the Whisper test, all the 590 records were considered. This first analysis was performed to observe the grouping based on a reduced sample of features strictly related to speech-in-noise perception and pure-tone perception and their contribution to cluster generation. The following 7 Whisper features were considered: 'PTA', 'SRT', 'gender', 'Age', '%correct', '#trials', 'total_time'. The gender variable was binarized (Female: 1, Male: 0).

The results for the following 3 configurations obtained with the K-Prototypes algorithm

are reported: 3, 4, and 5 clusters. The results are shown, at first, for the whole number of records and then also considering, for those subjects who have performed the test for both ears, only the record corresponding to the ear that demonstrated the lower (i.e., better) PTA value (434 records).

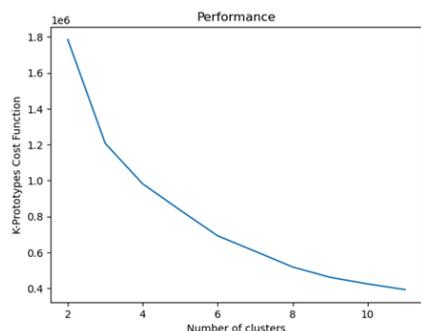


Figure 3.17: Elbow Curve illustrating the reduction in the cost function of K-Prototypes as the number of clusters increases, using exclusively Whisper features.

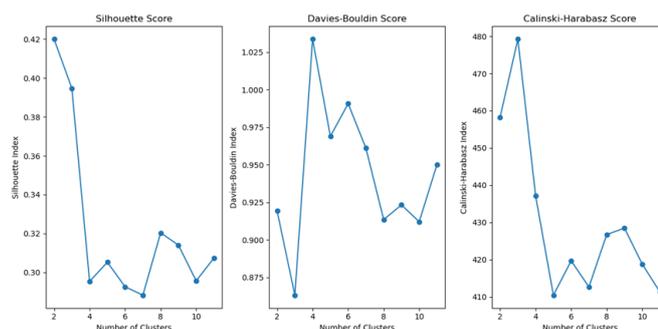


Figure 3.18: Silhouette Score, Davies-Bouldin Score, Calinski-Harabasz Index as the number of clusters increases, using exclusively Whisper features.

The choice of these three clustering configurations is based on the analysis conducted using the elbow curve and the Silhouette, Davies-Bouldin, and Calinski-Harabasz indices. As observed in Figures 3.17, the elbow curve suggests that the optimal number of clusters is 3, as the most significant change in slope occurs at this point. Additionally, the Silhouette score is highest at 3 clusters, the Davies-Bouldin index is low, and the Calinski-Harabasz index is high, exactly as theory would suggest, indicating the potential for achieving denser and better-separated clusters. Nonetheless, it was decided to also attempt clustering with 4 and 5 clusters to perform a comparative analysis and observe if increasing the number of clusters would significantly change the performance or explain something more (including for classification).

Three Clusters:

The results of clustering into 3 clusters using the K-Prototypes method applied to the dataset with Whisper features only (590 records) are reported. The dataset has been grouped in 3 clusters as follows: Cluster 1 contains 305 data points, Cluster 2 contains 263 data points, and Cluster 3 contains 22 data points. Figure 3.19 shows a 3D representation of the obtained clusters, using PTA, SRT, and Age as main features for visualization, with

centroids and medoids highlighted.

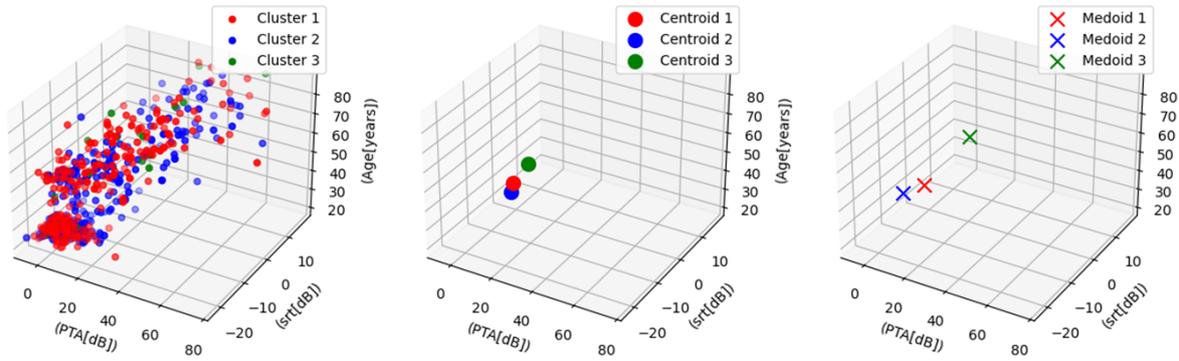


Figure 3.19: 3D representation of 3 clusters, with centroids and medoids highlighted (Whisper features only).

Tables 3.1 and 3.2 shows the complete N-dimensional coordinates of centroids and medoids, respectively, for the 3 clusters that have been derived.

Cluster	PTA[dB]	SRT[dB]	Age	gender	%correct	#trials	total_time[sec]
Cluster 1	14.45	-12.27	48.52	1	89.13	81.04	288.31
Cluster 2	12.42	-11.51	42.19	1	89.07	69.73	208.07
Cluster 3	17.84	-8.76	55.55	1	86.35	96.05	464.75

Table 3.1: Centroid table for 3 clusters (Whisper features only).

Cluster	PTA[dB]	SRT[dB]	Age	gender	%correct	#trials	total_time[sec]
Cluster 1	15	-13.58	49	1	90.48	84	292
Cluster 2	5	-13.83	42	0	90.91	77	215
Cluster 3	33.75	-10.35	76	1	90.72	97	433

Table 3.2: Medoid table for 3 clusters (Whisper features only).

The analysis of the clusters reveals distinct patterns in the auditory and demographic characteristics of the subjects.

Cluster 1 is characterized by subjects that demonstrate relatively good speech recognition in noise, with SRT values of -13.5833 dB (medoids coordinates). The average age in this cluster is around 49 years, and the cluster is predominantly composed by female subjects. The percentage of correct responses is high, approximately 90.48%, with a moderate number of trials and total test time.

Cluster 2 shows subjects with slightly better hearing levels than Cluster 1, with PTA values of 5 dB. Their ability to recognize speech in noise is comparable to Cluster 1, with

SRT values of -13.8333 dB. The average age is slightly lower, around 42 years, with male gender distribution. The performance is similarly high in terms of percentage correct, approximately 90.91%, with a fewer number of trials and shorter total test time compared to Cluster 1.

Cluster 3 comprises older subjects, with an average age of 76 years (even if in centroids coordinates the age reported is lower, around 55.5 years). These individuals exhibit more significant higher PTA values (33.75 dB). Their speech recognition in noise is lower, with SRT values of -10.3542 dB. They still maintain a high percentage of correct responses (90.72%), though they require more trials and longer total test time, reflecting the increased difficulty they face.

It is observed that the most discriminating variables for the three clusters are PTA, the number of trials, and the total time.

Given the presence of several subjects who underwent the Whisper test twice (once for the right ear and once for the left ear), it was decided to repeat the same clustering analysis using, only the record with the better PTA value, i.e., the lowest value, for subjects with double Whisper tests. Therefore, the number of records analyzed was 434.

The dataset with the ears with better PTA records has been grouped in 3 clusters as follows: Cluster 1 contains 14 data points, Cluster 2 contains 202 data points, and Cluster 3 contains 218 data points. Figure 3.20, Table 3.3, and Table 3.4 represent, respectively, the 3D representation of the clusters for the ears with better PTA records only and the corresponding full set of centroids and medoids.

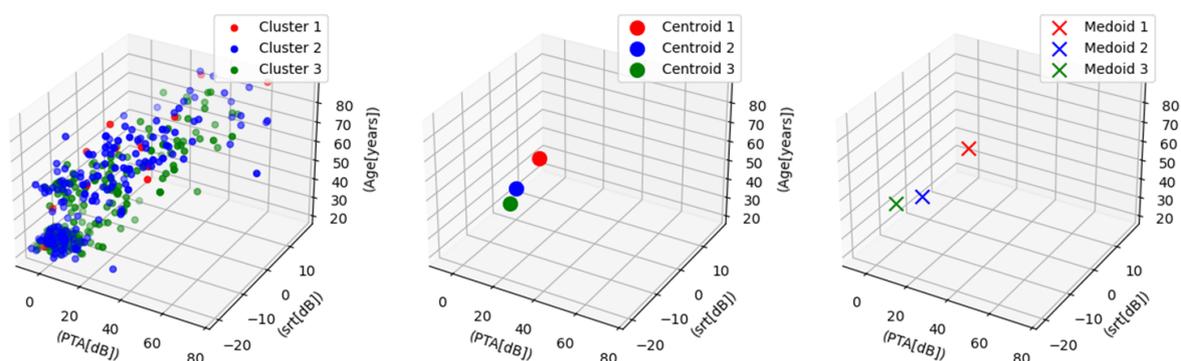


Figure 3.20: 3D representation of 3 clusters only for ears with better PTA records, with centroids and medoids (Whisper features only).

Cluster	PTA[dB]	SRT[dB]	Age	gender	%correct	#trials	total_time[sec]
Cluster 1	21.96	-7.11	63.64	1	85.36	95.36	470.01
Cluster 2	15.85	-11.49	50.84	1	88.97	80.37	285.10
Cluster 3	12.45	-11.34	41.75	1	89.28	69.65	205.92

Table 3.3: Centroids table for 3 clusters - only ears with better PTA records (Whisper features only).

Cluster	PTA[dB]	SRT[dB]	Age	gender	%correct	#trials	total_time[sec]
Cluster 1	33.75	-10.35	76	1	90.72	97	433
Cluster 2	15	-13.58	49	1	90.48	84	292
Cluster 3	2.5	-14.04	42	0	92.31	78	209

Table 3.4: Medoids table for 3 clusters - only ears with better PTA records (Whisper features only).

The clustering analysis, focusing only on the records with ears with better PTA values (i.e., the lowest values) for subjects with double Whisper tests, reveals distinct patterns across the clusters based on medoids values. Cluster 1 comprises older subjects, with an average age around 76 years, predominantly female. They exhibit the highest PTA values (33.75 dB), indicating more significant hearing loss. Their SRT is the poorest at approximately -10.35 dB. Despite the hearing challenges, their percentage of correct responses remains high (90.72%), though they require more trials (97) and longer total test time (433 seconds).

Subjects in the second cluster have PTA values of 15 dB. They show better SRT performance at -13.5833 dB. The average medoid age is 49 years, predominantly female. They maintain a high percentage of correct responses (90.48%) with a moderate number of trials (84) and total test time (292 seconds).

The third cluster includes younger subjects, with an average age of 42 years, mostly male. They exhibit the best hearing levels with PTA values 2.5 dB, and good SRT values of -14.0417 dB. Their performance is the highest in terms of correct responses (92.31%), requiring the fewest trials (78) and shortest total test time (209 seconds).

The inclusion of only ears with better PTA records highlights the significant influence of PTA, number of trials, and total test time in defining the clusters.

Four Clusters:

The dataset with Whisper features only has been grouped in 4 clusters as follows: Cluster 1 contains 223 data points, Cluster 2 contains 144 data points, Cluster 3 contains 18 data points, and Cluster 4 contains 205 data points. Figure 3.21 shows the 3D representation of the obtained clusters, using PTA, SRT, and Age as main features for visualization, with centroids and medoids highlighted. Tables 3.5 and 3.6 shows the complete N-dimensional coordinates of centroids and medoids, respectively, for the 4 clusters that have been derived.

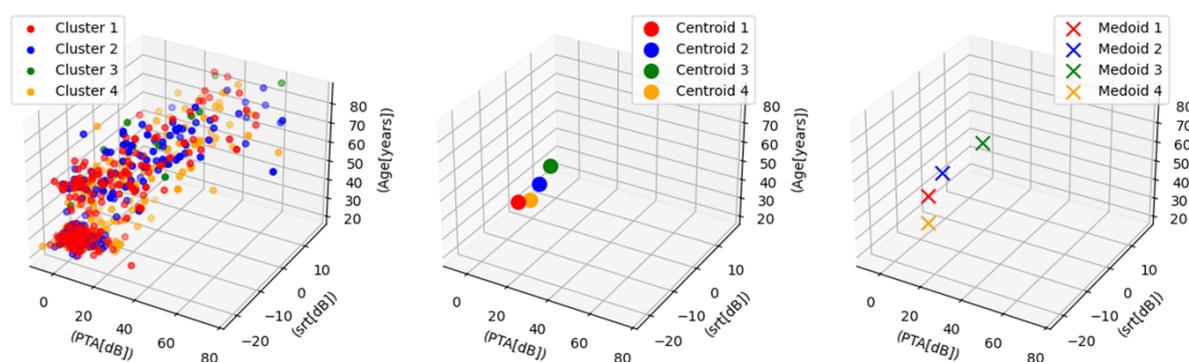


Figure 3.21: 3D representation of 4 clusters with centroids and medoids highlighted (Whisper features only).

Cluster	PTA[dB]	SRT[dB]	Age	gender	%correct	#trials	total_time[sec]
Cluster 1	11.62	-12.62	44.21	1	88.95	77.25	253.19
Cluster 2	12.95	-10.96	42.17	1	89.16	67.44	195.07
Cluster 3	16.97	-12.04	51.24	1	89.27	83.39	314.56
Cluster 4	19.17	-7.57	58.50	1	85.70	96.72	481.75

Table 3.5: Centroid table for 4 clusters (Whisper features only).

Cluster	PTA[dB]	SRT[dB]	Age	gender	%correct	#trials	total_time[sec]
Cluster 1	7.5	-12.36	44	0	87.50	80	253
Cluster 2	12.5	-16.28	36	0	92.65	68	191
Cluster 3	16.25	-13.46	60	0	91.01	89	316
Cluster 4	20	-0.39	63	1	87.83	115	477

Table 3.6: Medoid table for 4 clusters (Whisper features only).

The analysis of medoids of the clusters obtained reveals some distinctive characteristics that can help understand the differences between the identified groups. Particularly, the most significant features appear to be the PTA, the number of trials and the total_time.

In Clusters 1 and 2, for instance, a lower PTA is observed compared to the other clusters, indicating better hearing ability in these two categories. Cluster 2 includes younger subjects than Cluster 1 (36 years).

However, if centroids coordinates are observed, it is evident that Cluster 1 includes individuals with an average age of around 44 years and Cluster 2 has a slightly lower average age, around 42 years. This similarity in terms of age for centroids might suggest a different distribution of other characteristics, such as the number of trials and total_time. These two variables show differences in the two groups, more than the other features, so it is probably based on them that the clustering on centroids has divided the subjects of the first cluster to the subjects of the second one.

On the other hand, in Clusters 3 and 4, an increase in PTA is observed in the medoid table, but still remaining lower (or equal ad in case of Cluster 4) than 20dB, so still in normal-hearing range. Cluster 4 has a higher average age, around 63 years.

Additionally, it's worth noting that in Cluster 4 and 1 show a slightly lower correctness percentage (%correct) compared to the other two clusters. This might indicate that, despite greater hearing loss, other factors such as age or gender or also cognitive capabilities could influence spoken language comprehension abilities.

Similarly, the results of clustering when using 4 clusters but only ears with better PTA records are reported. The dataset with ears with better PTA records has been grouped in 4 clusters as follows: Cluster 1 contains 174 data points, Cluster 2 contains 156 data points, Cluster 3 contains 92 data points, and Cluster 4 contains 12 data points. Figure 3.22, Table 3.7, and Table 3.8 represent, respectively, the 3D representation of the clusters for the ears with better PTA records only and the corresponding full set of centroids and medoids.

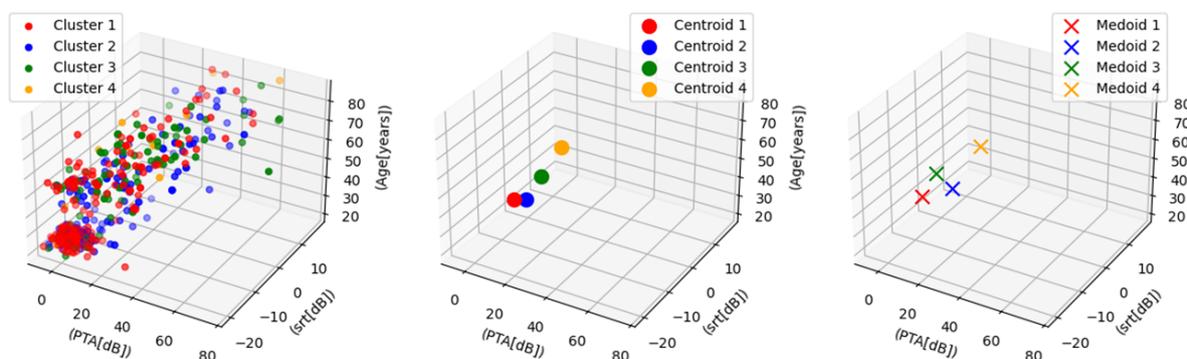


Figure 3.22: 3D representation of 4 clusters only for ears with better PTA records, with centroids and medoids (Whisper features only).

Cluster	PTA[dB]	SRT[dB]	Age	gender	%correct	#trials	total_time[sec]
Cluster 1	10.58	-12.68	43.88	1	89.48	77.98	254.42
Cluster 2	13.74	-10.58	42.25	1	88.91	67.42	194.301
Cluster 3	21.09	-10.53	56.81	1	88.88	81.73	311.77
Cluster 4	24.79	-5.55	67.33	1	84.40	95.67	483.18

Table 3.7: Centroid table for 4 clusters - only ears with better PTA records (Whisper features only).

Cluster	PTA[dB]	SRT[dB]	Age	gender	%correct	#trials	total_time[sec]
Cluster 1	7.5	-12.36	44	0	87.5	80	253
Cluster 2	13.75	-6.06	43	0	88.89	63	191
Cluster 3	16.25	-13.46	60	0	91.01	89	316
Cluster 4	33.75	-10.35	76	1	90.72	97	433

Table 3.8: Medoid table for 4 clusters - only ears with better PTA records (Whisper features only).

Cluster 1 has the lowest PTA while Cluster 4 shows values over 20dB, suggesting some sort of hearing loss, even if with values of SRT lower than Cluster 2 (that has normal-hearing value of PTA). All the clusters have predominance of male subjects except for Cluster 4, that also presents the longest total test time. Number of trials and total time of the Whisper test still seem to be the most distinctive features for clustering, even if in this case also PTA manages to distinct pretty well.

Five Clusters:

The dataset with Whisper features only has been grouped in 5 clusters as follows: Cluster 1 contains 184 data points, Cluster 2 contains 178 data points, Cluster 3 contains 110 data points, and Cluster 4 contains 100 data points, and Cluster 5 contains 18 data points.

Figure 3.23 shows the 3D representation of the obtained clusters, using PTA, SRT, and Age as main features for visualization, with centroids and medoids highlighted. Tables 3.9 and 3.10 shows the complete N-dimensional coordinates of centroids and medoids, respectively, for the 5 clusters that have been derived.

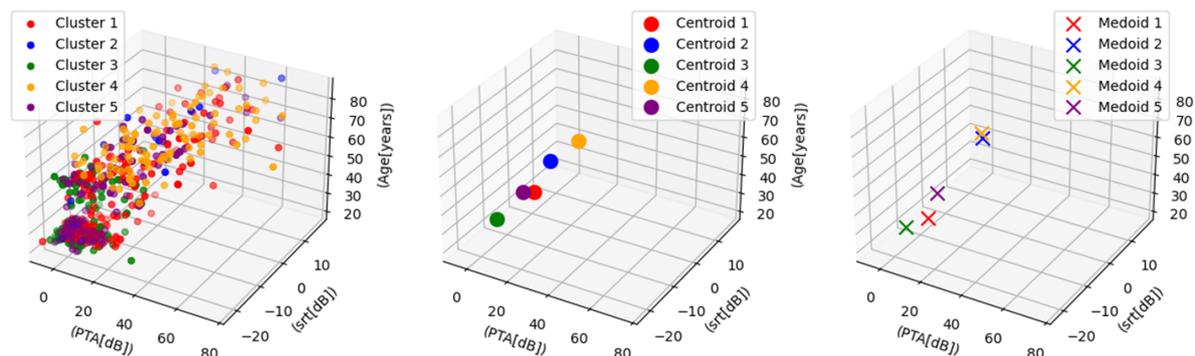


Figure 3.23: 3D representation of 5 clusters with centroids and medoids highlighted (Whisper features only).

Cluster	PTA[dB]	SRT[dB]	Age	gender	%correct	#trials	total_time[sec]
Cluster 1	14.74	-10.49	43.81	1	89.52	66.54	190.96
Cluster 2	19.17	-7.57	58.6	1	85.69	96.72	481.75
Cluster 3	2.69	-15.78	31.95	1	89.32	80.34	245.98
Cluster 4	30.66	-5.59	70.13	1	87.65	68.38	264.39
Cluster 5	13.19	-13.33	46.44	1	89.37	86.25	319.54

Table 3.9: Centroid table for 5 clusters (Whisper features only).

Cluster	PTA[dB]	SRT[dB]	Age	gender	%correct	#trials	total_time[sec]
Cluster 1	12.5	-16.28	36	0	92.65	68	191
Cluster 2	20	-0.39	63	1	87.83	115	477
Cluster 3	1.25	-16.54	28	1	89.48	76	249
Cluster 4	26.25	-5.95	73	1	89.33	75	259
Cluster 5	17.5	-16.47	51	1	90.70	86	329

Table 3.10: Medoid table for 5 clusters (Whisper features only).

This clustering permits to distinguish also a subset of subjects that is older (Cluster 4, 73 years old) but still performs pretty good (lower values of total_time compared to younger groups for example, and also quite good % of correct answers). 5 clusters also include a group for younger people (cluster 3, 28 years old), that have lowest PTA and SRT, and also low number of trials and total time. The inclusion of a fifth cluster makes more clear the distinction based on age compared to the other two configurations of clusters.

Similarly, the results of clustering when using 5 clusters and only the ears with better PTA records are reported. The dataset with the ears with better PTA records has been grouped in 5 clusters as follows: Cluster 1 contains 147 data points, Cluster 2 contains 111 data points, Cluster 3 contains 89 data points, Cluster 4 contains 75 data points, and Cluster 5 contains 12 data points. Figure 3.24, Table 3.11, and Table 3.12 represent, respectively, the 3D representation of the clusters for the ears with better PTA records only and the corresponding full set of centroids and medoids.

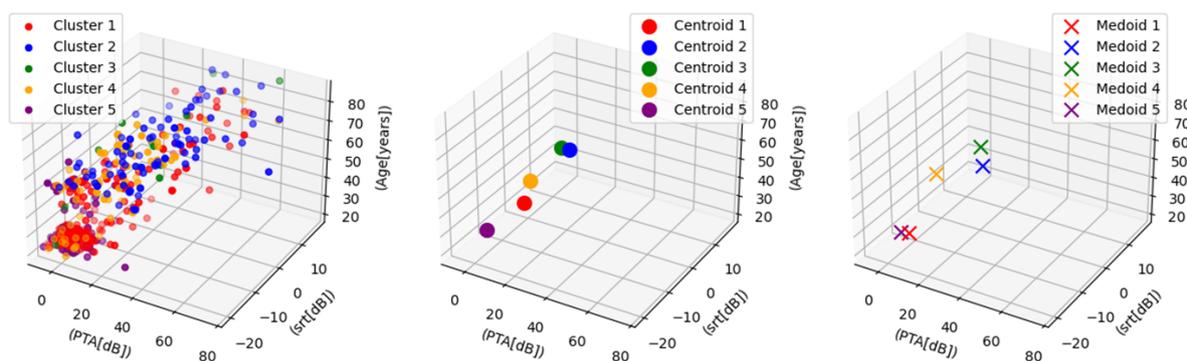


Figure 3.24: 3D representation of 5 clusters only for ears with better PTA records, with centroids and medoids (Whisper features only).

Cluster	PTA[dB]	SRT[dB]	Age	gender	%correct	#trials	total_time[sec]
Cluster 1	13.03	-10.96	41.10	1	88.95	68.05	192.61
Cluster 2	28.82	-5.86	67.64	1	87.92	67.91	252.83
Cluster 3	24.79	-5.55	67.33	1	84.40	95.67	483.18
Cluster 4	17.61	-12.07	55.2	1	89.31	84.13	317.82
Cluster 5	1.13	-16.11	29.29	1	90.24	83.31	257.75

Table 3.11: Centroid table for 5 clusters - only ears with better PTA records (Whisper features only).

Cluster	PTA[dB]	SRT[dB]	Age	gender	%correct	#trials	total_time[sec]
Cluster 1	3.75	-14.83	27	0	89.23	65	195
Cluster 2	32.5	-8.88	64	0	90.28	72	254
Cluster 3	33.75	-10.35	76	1	90.72	97	433
Cluster 4	16.25	-13.46	60	0	91.01	89	316
Cluster 5	-1.25	-14.29	25	1	90.12	81	263

Table 3.12: Medoid table for 5 clusters - only ears with better PTA records (Whisper features only).

Considering only ears with better PTA values with medoids the distinction in age is still more evident. Medoid's gender is more distinctive than in centroids, showing for example Cluster 1 that has younger male and Cluster 5 that has younger female, even if both with close ages.

Also in this case adding the 5 cluster permits to distinguish more classes based on age (especially for medoids coordinates), other than trials and time that seem to be always really differentiating clusters.

t-SNE visualization for 3 - 4 - 5 clusters - Whisper features only:

This subsection deals with the visualization of clusters fitted on the dataset containing the whole amount of data (590 records, using only Whisper features), using the t-SNE visualization tool. With $n_features > 3$ (more specifically 7 features), t-SNE allows clustered data to be projected onto a two-dimensional space using transformed coordinates. T-SNE visualization is reported for 3, 4, and 5 clusters, with perplexity equal to 50 and learning rate equal to 500. These values were chosen after performing several trials with different configurations of perplexity and learning rate (reported in Section 2.3.2), as they proved to provide a clearer visualization of the data points and better values of Silhouette Score, Davies-Bouldin index and Calinski Index.

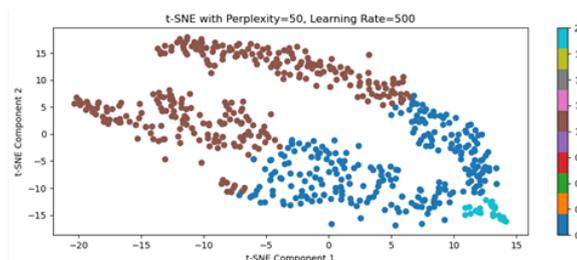


Figure 3.25: t-SNE, perplexity = 50, learning_rate= 500, 3 clusters.

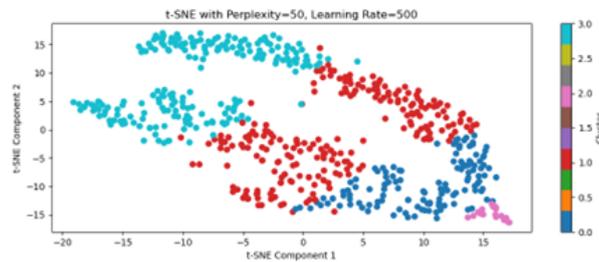


Figure 3.26: t-SNE, perplexity = 50, learning_rate= 500, 4 clusters.

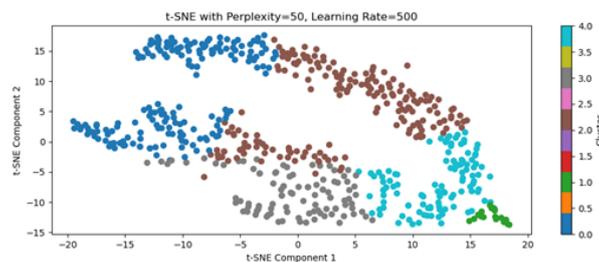


Figure 3.27: t-SNE, perplexity = 50, learning_rate= 500, 5 clusters.

The t-SNE visualizations provided offer insightful perspectives on the clustering behavior of the data under different configurations. In the first plot (Figure 3.25), with three clusters, there is a clear and distinct separation among the clusters, colored brown, blue, and cyan. This indicates that the algorithm effectively distinguishes between three broad groups within the data. As the number of clusters increases to four in the second plot (Figure 3.26), represented by red, blue, pink, and cyan, the separation remains fairly distinct but starts to show some overlap, particularly between the blue and pink clusters. This suggests a more nuanced separation where some clusters may share similarities. The complexity further increases in the third plot with five clusters (Figure 3.27), adding green and grey clusters into the mix. Here, there is noticeable overlap, especially among the brown, grey, and blue clusters, indicating that while the algorithm identifies more detailed groupings, these groups exhibit more commonalities, leading to less distinct boundaries. Therefore, while increasing the number of clusters provides a more detailed view of the data structure, it also introduces greater overlap and ambiguity. This necessitates a balance between the need for detailed analysis and the clarity of cluster separation, tailored to the specific application and nature of the data being studied.

3.3.3. Classification using only Whisper features:

In this section, the results of the classification performed using three methods are presented: Random Forest, SVM, and KNN. Each method was applied to the entire dataset

using only the Whisper features, but the dataset was divided into training and testing sets in three different ways.

Splitting methods:

- train = 80% (472 records), test = 20% (118 records)
- train = ears with better PTA (434 records), test = worse PTA (156 records)
- train = dataset without Oldenburg data (556 records), test = only Oldenburg records (34 records)

In the case of the first type of splitting (80%-20%), there is variability due to different possible choices of parameters and also due to varying compositions of the test and train sets because the random split is formed with different records each time, even if same percentage for train and test, leading to potentially different performance outcomes. This is why nested CV was chosen in this case.

For the second and third types of splitting (ears with better/worst PTA and without/with Oldenburg data), there is a fixed test set, so the variability in performance comes only from different parameter choices. Therefore, simple cross-validation is reported in these cases.

Classification for split: TRAINING = 80%, TEST = 20%

Classification results performed on 3-4-5 clusters for the entire dataset using only Whisper features are presented in this section.

Each classification algorithm for this splitting method underwent a nested CV with 5 folds (as explained in Section 2.3.4).

This process was adopted to ensure robustness and reliability of the results across various data splits. Table 3.13 reports the mean and standard deviation of the performance metrics (computed as macro average) on the 5 random splits in the external training and external test sets.

The optimal parameters obtained for Random Forest were: `max_depth=5` (for 3 clusters), `None` (for 4), `10` (for 5), `min_samples_leaf=1` (for 3 clusters), `1` (for 4), `2`(for 5), `min_samples_split=2` (for 3 clusters), `10` (for 4), `5` (for 5), `n_estimators= 30` (for 3 clusters) , `25` (for 4), `30` (for 5).

The best parameters for SVM were found to be, for all the number of clusters, `C = 100` and linear kernel.

For what concerns KNN, the best parameter has been 3 n_neighbors for 3 clusters and 4 clusters, and it was 7 n_neighbors for 5 clusters.

Table 3.13: Classification report: Mean and standard deviation of performance metrics across 5 random splits into training and external test sets. Splitting method: training = 80%, test = 20%, 3 clusters (Whisper features only).

Metric	Train Mean	Train Std	Test Mean	Test Std
<i>Random Forest</i>				
Accuracy (Macro)	0.997	0.003	0.977	0.010
Precision (Macro)	0.998	0.002	0.983	0.008
Recall (Macro)	0.989	0.013	0.893	0.058
F1-Score (Macro)	0.993	0.007	0.925	0.041
<i>SVM</i>				
Accuracy (Macro)	0.992	0.005	0.979	0.018
Precision (Macro)	0.995	0.003	0.985	0.011
Recall (Macro)	0.994	0.004	0.985	0.014
F1-Score (Macro)	0.994	0.004	0.985	0.013
<i>KNN</i>				
Accuracy (Macro)	0.925	0.019	0.885	0.042
Precision (Macro)	0.948	0.013	0.787	0.145
Recall (Macro)	0.804	0.055	0.692	0.069
F1-Score (Macro)	0.849	0.051	0.713	0.090

Random Forest achieves high accuracy on both training (0.997) and test (0.977) sets, with slightly higher variability in recall and F1-score on test. SVM shows consistently strong performance with accuracy of 0.992 (training) and 0.979 (test), maintaining stable precision, recall and F1-score across both sets. KNN performs less consistently with accuracies of 0.925 (training) and 0.885 (test), indicating more variability and lower generalization compared to the other models. Random Forest and SVM demonstrate superior performance over KNN, particularly in maintaining high accuracy and stability across training and test sets.

Table 3.14 shows the performance of nested CV when considering 4 clusters.

Table 3.14: Classification report: Mean and standard deviation of performance metrics across 5 random splits into training and external test sets. Splitting method: training = 80%, test = 20%, 4 clusters (Whisper features only).

Metric	Train Mean	Train Std	Test Mean	Test Std
<i>Random Forest</i>				
Accuracy (Macro)	0.998	0.003	0.896	0.043
Precision (Macro)	0.958	0.083	0.849	0.110
Recall (Macro)	0.959	0.080	0.824	0.102
F1-Score (Macro)	0.958	0.081	0.827	0.110
<i>SVM</i>				
Accuracy (Macro)	0.996	0.003	0.972	0.014
Precision (Macro)	0.998	0.002	0.959	0.028
Recall (Macro)	0.997	0.002	0.980	0.007
F1-Score (Macro)	0.997	0.002	0.967	0.017
<i>KNN</i>				
Accuracy (Macro)	0.884	0.031	0.769	0.012
Precision (Macro)	0.909	0.028	0.782	0.084
Recall (Macro)	0.836	0.073	0.714	0.083
F1-Score (Macro)	0.859	0.063	0.734	0.082

Including an additional cluster leads to slight changes in model performance: Random Forest has 0.998 accuracy on training data but shows decreased accuracy (0.896) and increased variability in test. SVM continues to exhibit strong performance with accuracy of 0.996 (training) and 0.972 (test), showing consistent precision and recall across both sets (slightly lower in the test set). It suggests robust generalization. KNN demonstrates lower overall performance compared to Random Forest and SVM, with accuracy of 0.884 (training) and 0.769 (test). Precision and recall metrics also decrease, indicating challenges in handling the additional cluster complexity. While Random Forest and SVM maintain competitive performance with the addition of a fourth cluster, KNN struggles more with lower accuracy and less stable metrics.

Table 3.15 shows the performance of nested CV when considering 5 clusters.

Table 3.15: Classification report: Mean and standard deviation of performance metrics across 5 random splits into training and external test sets. Splitting method: training = 80%, test = 20%, 5 clusters (Whisper features only).

Metric	Train Mean	Train Std	Test Mean	Test Std
<i>Random Forest</i>				
Accuracy (Macro)	0.999	0.001	0.909	0.031
Precision (Macro)	0.999	0.001	0.840	0.094
Recall (Macro)	0.994	0.010	0.795	0.058
F1-Score (Macro)	0.997	0.006	0.798	0.071
<i>SVM</i>				
Accuracy (Macro)	0.993	0.006	0.949	0.014
Precision (Macro)	0.994	0.005	0.888	0.066
Recall (Macro)	0.994	0.005	0.919	0.073
F1-Score (Macro)	0.994	0.005	0.900	0.067
<i>KNN</i>				
Accuracy (Macro)	0.869	0.032	0.742	0.027
Precision (Macro)	0.895	0.026	0.715	0.081
Recall (Macro)	0.842	0.043	0.699	0.068
F1-Score (Macro)	0.861	0.038	0.694	0.070

Adding a fifth cluster introduces further complexity to the classification task. Random Forest achieves accuracy of 0.999 on the training data but experiences a slight drop in test accuracy 0.909. Precision metric also decrease, indicating challenges in generalizing to the test set with the additional cluster. SVM maintains high accuracy on both training (0.993) and test (0.949) data, with stable precision and recall metrics, even if higher drop in precision test is experienced. This suggests robust performance in handling the increased cluster complexity. KNN exhibits lower overall performance compared to Random Forest and SVM, with a decrease in the classification metrics. While SVM continues to demonstrate strong and stable performance with the addition of a fifth cluster, and also Random Forest has good performances, KNN struggles with lower accuracy and less stable metrics.

Classification for split: TRAINING = ears with better PTA, TEST = worse PTA

This subsection deals with classification results on the dataset considering Whisper features only, as in the previous case, but considering a different splitting method. Specifically, the training set comprises records where the ears with better PTA was selected (434 records), while the remaining records form the test set (156).

For each classification method, the accuracy on the training and test set obtained with 5-fold-CV on all the combinations of parameters is reported. Moreover, the performance of the model with the best parameters has been applied to the training set and evaluated on the test set and is commented.

The best parameters obtained using 5-fold-CV are:

Random Forest: max_depth=5, min_sample_leaf=1, min_sample_split=10, n_estimators=20. Mean train accuracy (CV=5 over all parameters) = 0.992 ± 0.009 . Mean test accuracy (CV=5 over all parameters) = 0.957 ± 0.028 .

SVM: C = 1, kernel = 'linear'. Mean train accuracy (CV=5 over all parameters) = 0.817 ± 0.194 . Mean test accuracy (CV=5 over all parameters) = 0.799 ± 0.186 .

KNN: n_neighbors = 3.

Mean train accuracy (CV=5 over all parameters) = 0.936 ± 0.016 . Mean test accuracy (CV=5 over all parameters) = 0.867 ± 0.029 .

Mean performances highlight Random Forest as the most performant algorithm in both training and test set, while SVM is the one showing more std compared to the other two.

Considering only the best performance, Random Forest achieves perfect accuracy (1.00) on the training set and high accuracy (0.99) on the test set. Precision, recall, and F1-score are consistently high across clusters in both sets, indicating strong generalization ability. SVM shows strong best performance with 0.99 accuracy on the training set and 0.96 accuracy on the test set. Precision, recall, and F1-score are generally high for clusters 1 and 2 in both sets, but show a slight drop, especially in recall and F1-score, for cluster 3 on the test set. KNN demonstrates lower consistency compared to the other models, even in the best performance. It achieves 0.96 accuracy on the training set, but this drops to 0.87 on the test set. Particularly, cluster 3 exhibits significantly lower performance metrics on the test set, indicating challenges in generalizing effectively. While Random Forest and SVM maintain robust performance even with this type of data splitting, KNN shows more variability and overall lower performance, especially in handling cluster 3,

that is actually the one with less records.

The best parameters obtained using 5-fold-CV with 4 clusters are:

Random Forest: max_depth=20, min_sample_leaf=1, min_sample_split=10, n_estimators=150.

Mean train accuracy (CV=5 over all parameters) = 0.989 ± 0.010 . Mean test accuracy (CV=5 over all parameters) = 0.960 ± 0.022 .

SVM: C = 100, kernel = 'linear'.

Mean train accuracy (CV=5 over all parameters) = 0.724 ± 0.235 . Mean test accuracy (CV=5 over all parameters) = 0.688 ± 0.215 .

KNN: n_neighbors = 5.

Mean train accuracy (CV=5 over all parameters) = 0.895 ± 0.024 . Mean test accuracy (CV=5 over all parameters) = 0.791 ± 0.028 .

Performance metrics on the best performing set of parameters, considering 4 clusters (precision, recall, F1-score) are high for Clusters 1, 2, and 4 on both sets for Random Forest and SVM, indicating robust generalization. However, Cluster 3 shows lower performance on the test set, particularly in recall and F1-score. KNN demonstrates lower accuracy (0.91) on the training set, and 0.97 on the test set. It shows varied performance across clusters, with Clusters 1, 2, and 4 generally performing well on the test set. However, Cluster 3 again shows lower recall and F1-score.

Lastly, the best parameters obtained using 5-fold-CV with 5 clusters are:

Random Forest: max_depth=20, min_sample_leaf=2, min_sample_split=2, n_estimators=25.

Mean train accuracy (CV=5 over all parameters) = 0.980 ± 0.016 . Mean test accuracy (CV=5 over all parameters) = 0.907 ± 0.021 .

SVM: C = 10, kernel = 'linear'.

Mean train accuracy (CV=5 over all parameters) = 0.711 ± 0.261 . Mean test accuracy (CV=5 over all parameters) = 0.674 ± 0.238 .

KNN: n_neighbors = 7.

Mean train accuracy (CV=5 over all parameters) = 0.883 ± 0.030 . Mean test accuracy (CV=5 over all parameters) = 0.745 ± 0.050 .

Considering the best performing with 5 clusters, Random Forest shows varied performance across clusters on the test set, with some clusters achieving high F1-scores but others showing lower recall and F1-scores, while SVM generally achieves high precision, recall, and F1-scores across clusters, demonstrating robustness in handling different clusters. SVM exhibits the most consistent performance across clusters and sets, followed by Random Forest, whereas KNN struggles more with the test set, especially in maintaining high accuracy across all clusters.

Classification for split: TRAINING = previous data, TEST = Oldenburg data

This subsection deals with classification results on the dataset considering Whisper features only, as in the previous case, but considering a third splitting method.

Specifically, in this scenario, the training set comprises records that have been acquired until December 2023 (556 records), while the remaining records form the test set correspond to the data acquired in Oldenburg (34 records).

For each classification method, the accuracy on the training and test set obtained with 5-fold-CV on all the combinations of parameters is reported. Moreover, the performance of the model with the best parameters has been applied to the training set and evaluated on the test set and is commented.

The best parameters obtained using 5-fold-CV with 3 clusters are:

Random Forest: max_depth=20, min_sample_leaf=1, min_sample_split=2, n_estimators=25.

Mean train accuracy (CV=5 over all parameters) = 0.993 ± 0.007 . Mean test accuracy (CV=5 over all parameters) = 0.974 ± 0.010 .

SVM: C = 100, kernel = 'linear'.

Mean train accuracy (CV=5 over all parameters) = 0.822 ± 0.183 . Mean test accuracy (CV=5 over all parameters) = 0.808 ± 0.176 .

KNN: n_neighbors = 5.

Mean train accuracy (CV=5 over all parameters) = 0.936 ± 0.012 . Mean test accuracy (CV=5 over all parameters) = 0.893 ± 0.045 .

Considering best performing, on the Oldenburg test set, Random Forest and SVM generally maintain strong performance across most clusters (with 1.00 accuracy on training,

0.94 on test), while KNN shows more variability, especially in handling Cluster 3.

The best parameters obtained using 5-fold-CV with 4 clusters are:

Random Forest: max_depth=10, min_sample_leaf=1, min_sample_split=10, n_estimators=20.

Mean train accuracy (CV=5 over all parameters) = 0.990 ± 0.008 Mean test accuracy (CV=5 over all parameters) = 0.969 ± 0.017

SVM: C = 100, kernel = 'linear'.

Mean train accuracy (CV=5 over all parameters) = 0.738 ± 0.236 Mean test accuracy (CV=5 over all parameters) = 0.711 ± 0.220

KNN: n_neighbors = 3.

Mean train accuracy (CV=5 over all parameters) = 0.907 ± 0.020 Mean test accuracy (CV=5 over all parameters) = 0.817 ± 0.042

In the case of best performance, Random Forest achieved 0.99 accuracy on the training set but dropped to 0.76 on the test set. Performance metrics vary across clusters in the test set, with Cluster 1 and Cluster 2 showing strong precision and F1-scores, while Cluster 3 performs less consistently. Cluster 4 lacks test data. SVM reached 1.00 accuracy on the training set and 0.94 on the test set. It demonstrates balanced precision and F1-scores across most clusters in the test set, except for a slightly lower precision in Cluster 3. KNN achieved 0.92 accuracy on the training set but dropped to 0.65 on the test set. The ability of the algorithms to generalize to Oldenburg test data varies. SVM shows the most consistent performance, while Random Forest and KNN exhibit more variability. The less numerosity of the test set (only 34 records) shows some challenge for the classification due to insufficient amount of data. In fact, support for training set is: 214, 124, 13, and 205 records. Support for test set: 9, 20, 5, and 0 records.

Lastly, the best parameters obtained using 5-fold-CV with 5 clusters are:

Random Forest: max_depth=20, min_sample_leaf=1, min_sample_split=2, n_estimators=50.

Mean train accuracy (CV=5 over all parameters) = 0.982 ± 0.014 Mean test accuracy (CV=5 over all parameters) = 0.919 ± 0.033

SVM: C = 100, kernel = 'linear'.

Mean train accuracy (CV=5 over all parameters) = 0.728 ± 0.248 Mean test accuracy (CV=5 over all parameters) = 0.694 ± 0.229

KNN: n_neighbors = 7.

Mean train accuracy (CV=5 over all parameters) = 0.884 ± 0.029 Mean test accuracy (CV=5 over all parameters) = 0.785 ± 0.043

Random Forest achieved 1.00 accuracy on the training set, in the best performing set of parameters. In the test set, overall accuracy is 0.82. In this case for all algorithms, Cluster 1 and Cluster 4 lack of test samples. SVM achieved 1.00 accuracy on the training set and 0.97 on the test set. It demonstrates consistent performance across most clusters in the test set, with strong metrics in precision, recall, and F1-score. KNN, instead, in the test set, has value of accuracy that drops to 0.56. Clusters 2, 3, and 5 exhibit varying levels of performance in precision, recall, and F1-score. SVM shows the most consistent and robust performance across all clusters, followed by Random Forest, whereas KNN exhibits more variability with the addition of a fifth cluster.

3.3.4. Clustering adding DST and risk factors:

This section presents the clustering results obtained by incorporating, in addition to the Whisper features considered previously, features extracted from the risk factor questionnaire and the DST. This was done to observe the influence of the DST and risk factors on clustering and classification, and to assess whether the cluster generation remains consistent with an increased number of features. However, to select these features, the sample size was reduced because not all subjects had performed the DST and risk factors questionnaire. As a result, a dataset of 266 records was considered. In this case, feature reduction techniques such as PCA, FAMD were considered and t-SNE was also used for visualization. Clustering was performed and is presented in the same manner as in the case of Whisper-only features, specifically for 3, 4, and 5 clusters.

Feature reduction

Prior to applying clustering methods, the following feature reduction techniques were analysed: PCA (with only numerical attributes) and FAMD (with mixed attributes) to visualize the dataset after processing and to observe any patterns or clusters (before the actual clustering).

Categorical features (e.g., “ear”, “gender”, “Native_language”, “cardio”, “cholesterol”, “depression”, “diabetes”, “drink”, “ear_infections”, “education”, “family_history_HL”, “head_trauma”, “meningitis”, “obesity”, “smoker”, “stroke”, “high_volume_exposure”, “work_exposure”) were removed before applying PCA.

The remaining 12 features were numerical and were therefore used for PCA analy-

sis. These features included: “Age”, “PTA”, “total_time”, “%correct”, “#correct”, “#trials”, “SRT”, “avg_timecount”, “digitSpanScore”, “cancPressTot”, and “correctPercentagesMean”, “avgSingleDigitTimesMean”.

The data were also standardized prior to applying PCA [41]. The right panel of Figure 3.28 shows the explained variance curve as the number of PCA components varies. It can be observed that as the number of retained PCA components increases, the percentage of variance of the original dataset explained by the components also increases. Specifically, retaining 5 PCA components allows for a drastic reduction in the number of features (from 12 to 5), while preserving most of the information content of the original dataset. Indeed, 5 components are capable of explaining most of the original dataset variance (0.86%); therefore, it was decided to retain only 5 components.

In the left panel Figure 3.28, it can be observed the individual components and their contribution to explaining the variance. The first component contributes the most, while from the second to the fifth component, the explained variance is significantly lower and similar among them. The second plot highlights how there is a drop in explained variance starting from the second component, with a corresponding increase in residual variance.

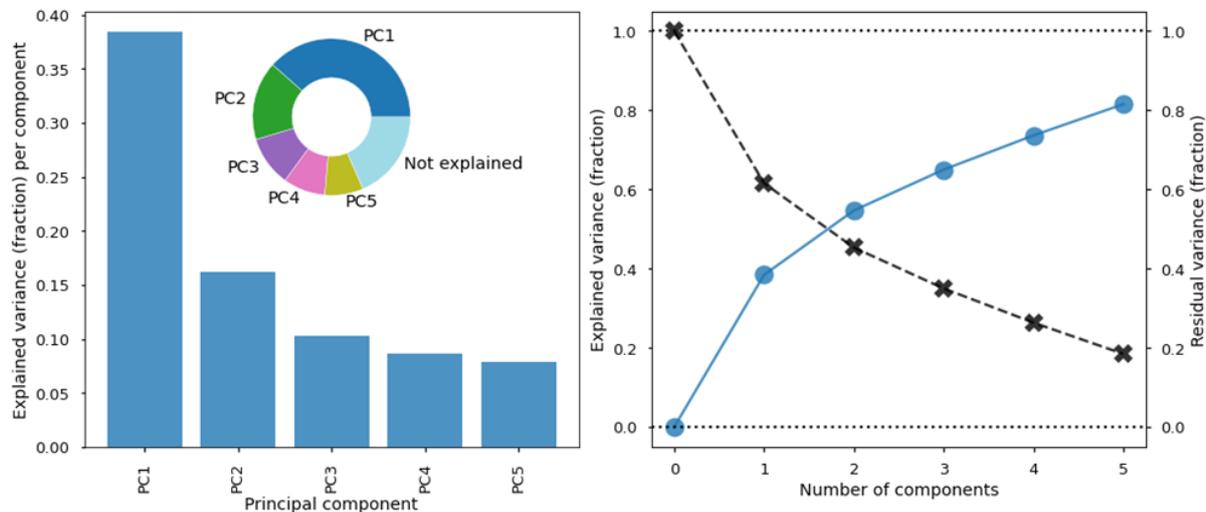


Figure 3.28: Explained variance from PCA as a function of the number of principal components considered—joined plots.

Figure 3.29 shows how the original features of the dataset contribute to the first 5 principal components (PC1, PC2, PC3, PC4, PC5). The size of each bubble represents the (absolute) magnitude of the contribution of a variable to a particular principal component. Larger bubbles indicate a stronger influence of that variable on the component.

For example, it can be observed that the first principal component does not have a single variable that contributes significantly more than others. Variables such as age, PTA, #correct, and SRT all have a fairly high contribution, with none standing out significantly over the others. In the case of PC2, however, it is clear that total_time is the variable that contributes the most. PC3 also has a fairly balanced contribution, although correctPercentagesMean and avgSingleDigitTimes_mean are slightly higher. PC4 and PC5, on the other hand, have notable contributions from %correct and cancPressTot.

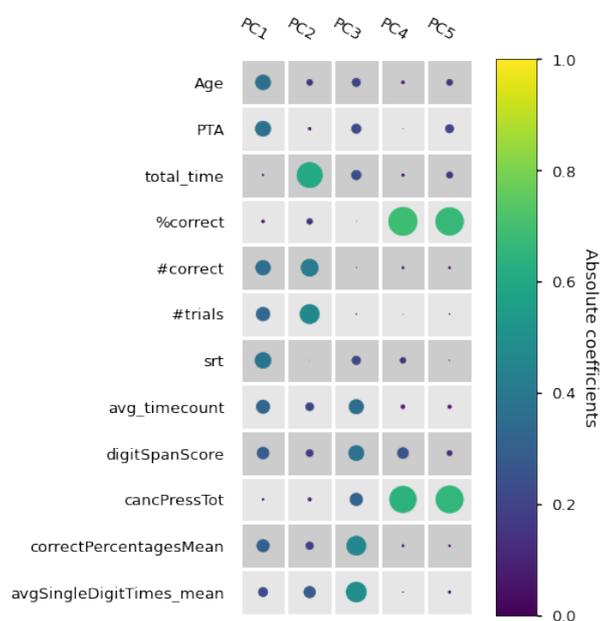


Figure 3.29: PCA Loadings matrix for the retained PCA components (absolute values). The size of the bubble is proportional to the contribution of the feature to that PCA component.

Since only numerical features were involved, K-means clustering was applied after feature reduction. Observing the elbow curve (Figure 3.30), it seemed reasonable to try with 3 clusters. A comparison between the PCA 2D visualization and the t-SNE visualization is reported (Figure 3.31). From the figure, a fairly distinct distribution can be observed among the points of the 3 clusters, with some points at the boundaries showing slight overlap but overall the distinction appears to be noticeable.

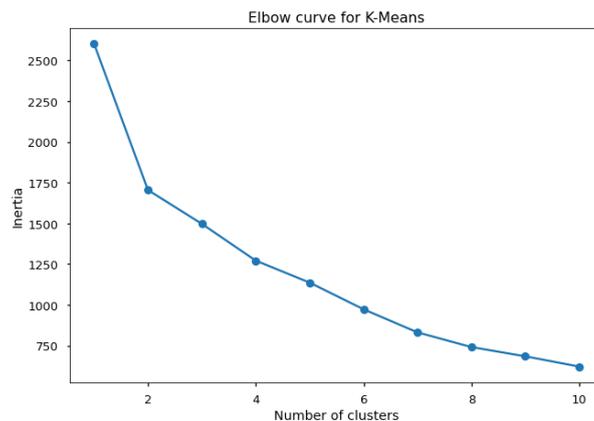


Figure 3.30: Elbow Curve for K-Means after PCA reduction.

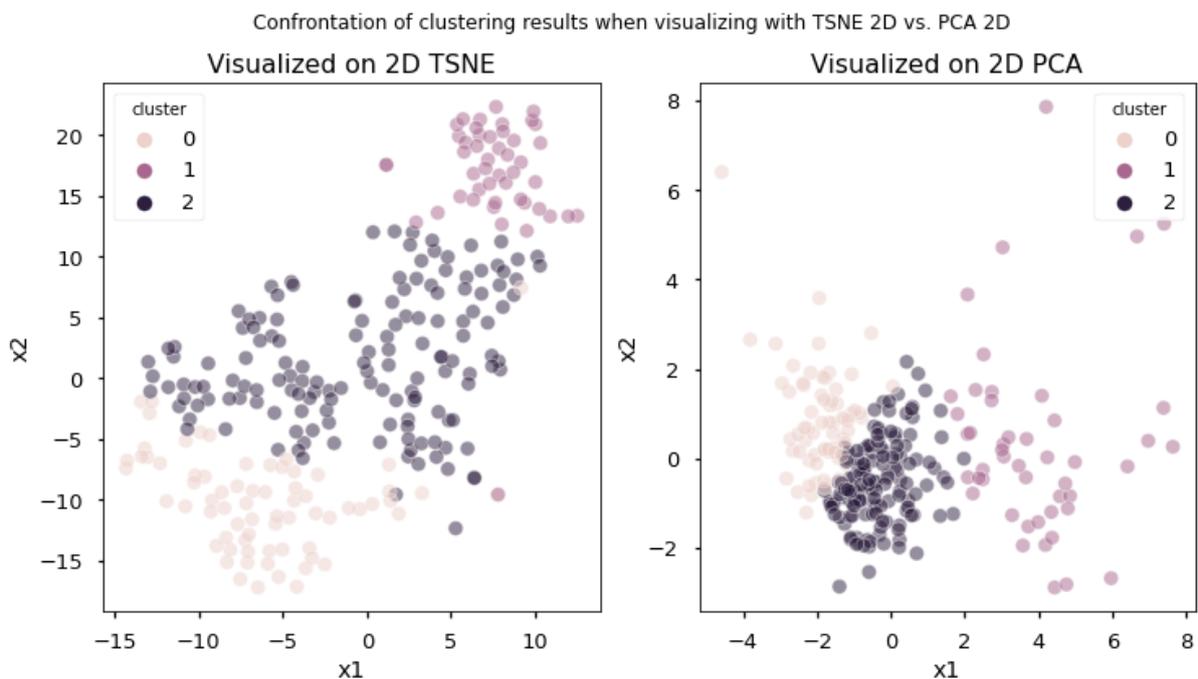


Figure 3.31: TSNE 2D VS. PCA 2D with 3 clusters - using k-Means algorithm.

To incorporate more variables, FAMD was then applied, as it is suitable for both numerical and categorical variables. The non-numerical variables were converted to the “category” type. Specifically, the 18 categorical features considered were: “ear”, “gender”, “Native_language”, “cardio”, “cholesterol”, “depression”, “diabetes”, “drink”, “ear_infections”, “education”, “family_history_HL”, “head_trauma”, “meningitis”, “obesity”, “smoker”, “stroke”, “high_volume_exposure”, “work_exposure”, for a total number of features that was equal to 30.

A graph with the percentage of explained variances based on the number of dimensions (so the variance explained by each component) is reported in Figure 3.32. The first dimension contributes with the highest percentage (15.9%), followed by the second dimension with 6.4% and the third with 5.7%, while higher dimensions (from the 4th to the 10th) contribute with percentages that are from 4.9% to 3.5%.

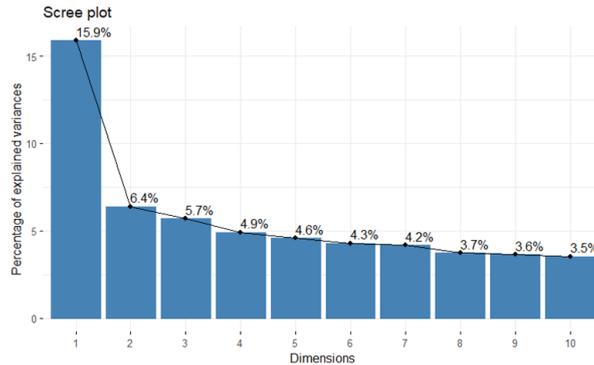


Figure 3.32: Percentage of Variance explained by the first 10 components provided by the FAMD feature reduction technique.

In Figure 3.33 and 3.34 there is the contribution of the variables respectively to the first and second dimension. The red dashed lines on the graphs indicate the expected average value, if the contributions were uniform. For the first dimension, the variables seem to contribute almost in similar percentages, with age, PTA and SRT leading them. On the contrary, the second dimension has a consistent contribution given from total_time, followed by #trials and #correct.

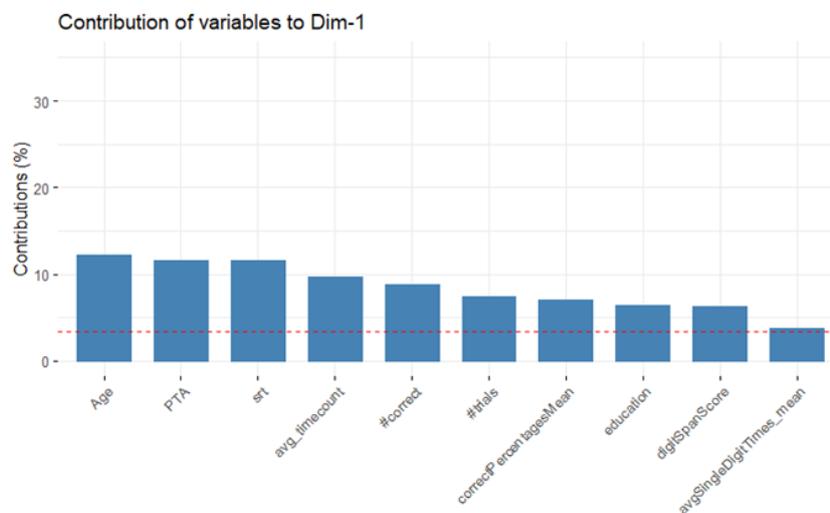


Figure 3.33: Contributions of the variables to the first dimension of FAMD analysis.

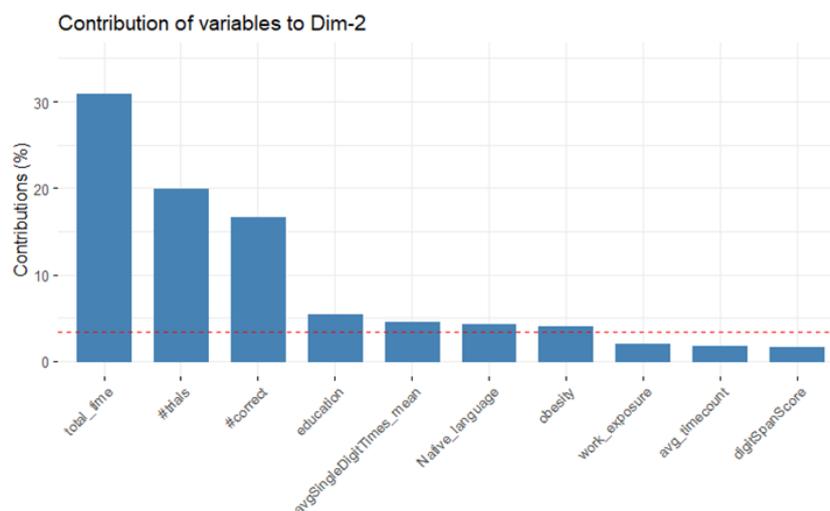


Figure 3.34: Contributions of the variables to the second dimension of FAMD analysis.

Figure 3.35 is a biplot of the quantitative variables on a two-dimensional space created by the first two principal components of the FAMD analysis. Each point in the plot represents a quantitative variable from the dataset. The variables are colored based on their contribution to the principal components. Variables that contribute more to explaining the variance of the principal components are depicted in more intense colors (closer to red/orange), while those with lesser contributions are shown in lighter colors (closer to blue). The color gradient (from blue to yellow to orange) helps quickly identify which variables are more influential. Variables with higher contributions are colored in orange, namely #trials, #correct, total_time (as well as Age, PTA, and SRT, and avg_timecount), while those with lower contributions are colored in blue (cancePressTot and %correct).

The distance of the points from the origin of the plot indicates how much each variable contributes to the variance explained by the two principal components. The farther a variable is from the center, the greater its contribution. Therefore, the variables that are more orange and farther from the center include #trials, #correct, total_time (followed by PTA, Age, SRT, and avg_timecount), while those closer to blue and near the center include cancPressTot and %correct.

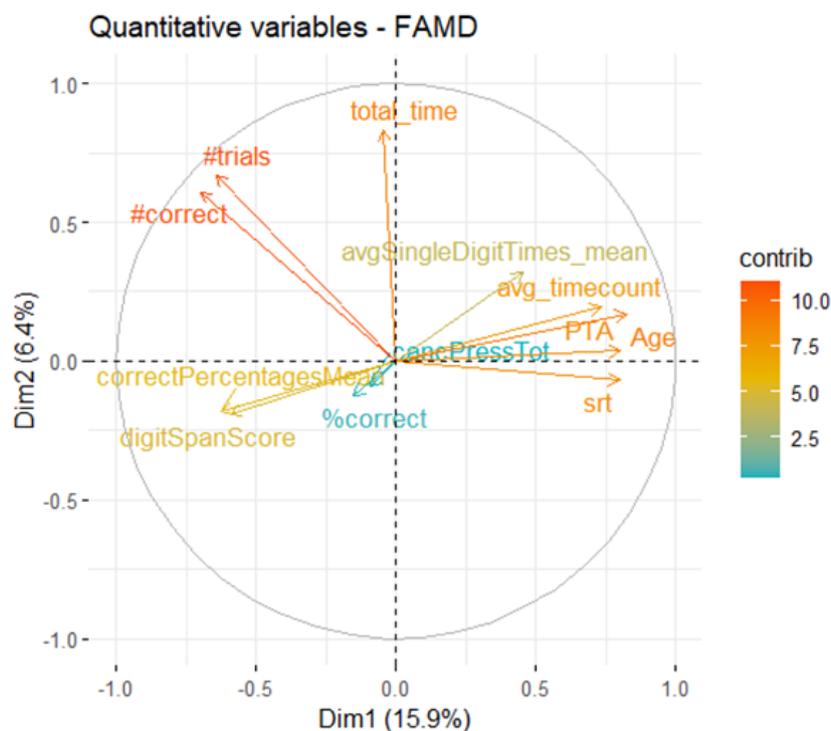


Figure 3.35: Biplot of quantitative variables for contribution to the first two dimensions of FAMD.

Three Clusters:

In this section, the results for clustering into 3 clusters, using not only the features from Whisper but also those from the DST and the risk factor questionnaire (266 records), are shown. Specifically, the selected features include: 'PTA', 'SRT', 'gender', 'Age', 'digitSpanScore', '%correct', '#trials', 'total_time', 'avgSingleDigitTimes_mean', 'family_history_HL', 'cardio', 'high_volume_exposure', and 'education'. The additional features include the DST score, the average single-digit typing time, two risk factors related to hearing impairment (presence of hearing loss in the family, high volume exposure), one feature related cardiovascular problems, and one related to education. Categorical variables were processed using LabelEncoder(), the gender variable was binarized (Female:1, Male:0), while education was coded in this way: Bachelor = 0, Secondary school (High) = 1, Master = 2, Primary school (Middle) = 3. The other cases have the following correspondence: "Yes" = 1, "No" = 0.

The dataset has been grouped in 3 clusters as follows: Cluster 1 contains 147 data points, Cluster 2 contains 9 data points, and Cluster 3 contains 110 data points. Figure 3.36 shows the 3D representation of the clusters using PTA, SRT, and Age, highlighting

centroids and medoids.

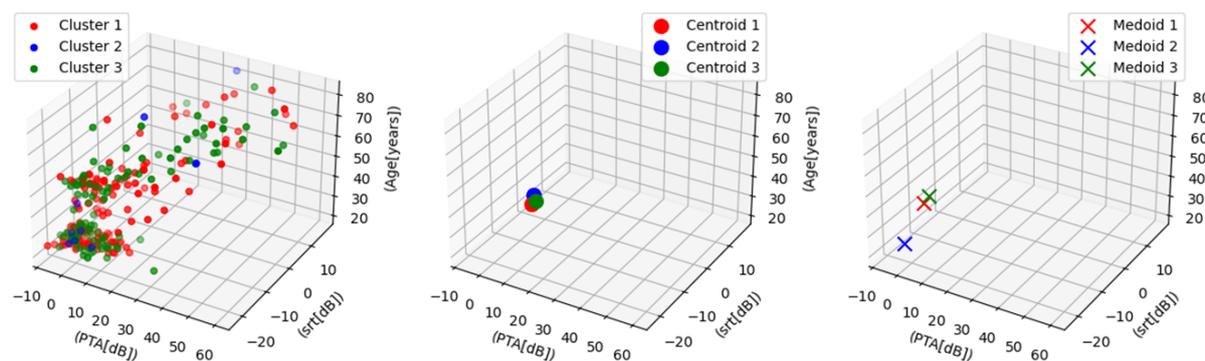


Figure 3.36: 3D representation of 3 clusters, with centroids and medoids highlighted (Whisper + risk factor + DST features).

Table 3.16: Centroids and Medoids table for 3 clusters.

	Centroids			Medoids		
	C11	C12	C13	C11	C12	C13
PTA [dB]	8.52	7.64	10.78	2.50	-3.75	5.00
SRT [dB]	-12.97	-11.58	-13.58	-17.03	-19.18	-17.60
Age	40	43	43	43	23	48
Gender	1	1	1	0	1	1
%correct	88.7	86.4	89.6	91.4	88.5	89.5
#trials	71	102	80	70	104	76
total_time [s]	219	490	296	222	429	293
digitSpanScore	6	6	6	6	5	8
avgSingleDigitTimes_mean	1.15	1.28	1.39	0.94	1.46	2.21
family_history_HL	0	0	0	1	0	1
cardio	0	0	0	0	0	0
high_volume_exposure	0	0	0	1	1	0
education	1	1	1	1	1	1

Cluster 1, with 147 points, is characterized by a PTA of 2.5 dB (considering medoids coordinates) and SRT of -17.03 dB. The average age in this cluster is 43 years, and the gender is predominantly male. The DSS averages at 6, indicating moderate cognitive performance. This cluster shows a high percentage of correct responses at 91.4%, with an average of 70 trials and an average single digit time of 0.94 seconds. The total time

for tasks averages at 222 seconds, and all individuals have the same level of education. Cluster 2, comprising 9 points, has a PTA of -3.75 dB and an SRT of -19.18 dB. The individuals are younger (23 years) and are also predominantly female (the distinction on age is different if we consider centroids, because the second cluster is older). The DSS is lower at 5.0. This cluster has a lower percentage of correct responses (88.5%) but engages in more trials, with an average of 104. The average single digit time is longer, at 1.46 seconds. Similar to Cluster 1, there is no cardiovascular issues, and same behaviour in terms of high volume exposure. The total task time is significantly higher at 429 seconds, yet the education level remains consistent with the other clusters.

Cluster 3, consisting of 110 points, presents the highest PTA at 5 dB (but still normal hearing value) and the lowest SRT at -17.604 dB. The average age in this cluster is 48 years, and the gender is predominantly female. The digit span score is 8.0, the highest. The percentage of correct responses is 89.5%, with an average of 76 trials. The average single digit time is the longest at 2.21 seconds. Similar to the other clusters, there is no reported cardiovascular issues, but high volume exposure has a different distribution. The total task time is 293seconds, and the education level is the same across all clusters.

Cluster 1 tends to represent individuals with mid hearing and cognitive performance, but with a family history of hearing loss and high volume exposure. Cluster 2 is distinctive due to the younger age and a significantly higher number of trials and total task time, which might suggest a more rigorous testing regime or higher task persistence, even with the lowest DSS. Cluster 3, with the highest PTA, indicates slightly worse hearing thresholds and longer cognitive processing times even if the score of the DST is the highest, with a notable family history of hearing loss.

Also here it was decided to re-execute the clustering considering only the records with the lowest value of PTA in case of subjects that have performed the test for both ears. The ears with better PTA dataset is then composed of 153 records.

The dataset with the ears with better PTA records has been grouped in 3 clusters as follows: Cluster 1 contains 69 data points, Cluster 2 contains 79 data points, and Cluster 3 contains 5 data points. Figure 3.37 represents the 3D representation of the clusters for the ears with better PTA records only and the following table (3.17) reports the corresponding full set of centroids and medoids.

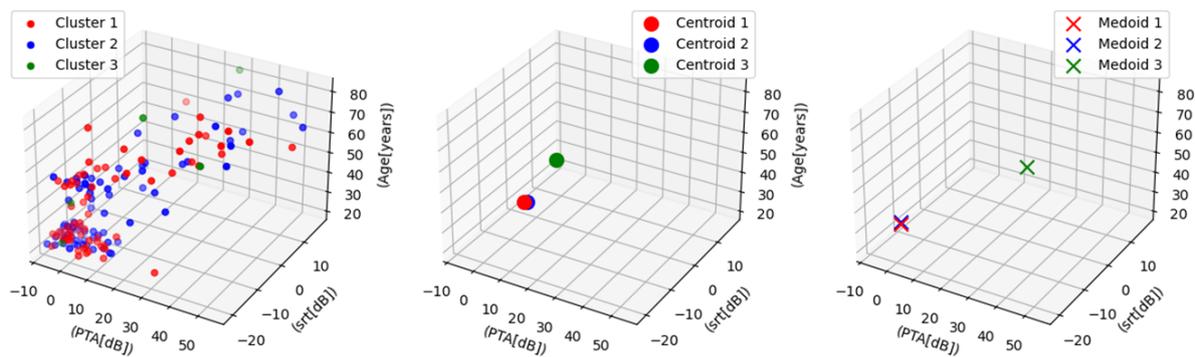


Figure 3.37: 3D representation of 3 clusters, with centroids and medoids highlighted, only ears with better PTA records (Whisper + risk factor + DST features).

Table 3.17: Centroids and Medoids table for 3 clusters, ears with better PTA only.

	Centroids			Medoids		
	C11	C12	C13	C11	C12	C13
PTA [dB]	9.51	8.78	15.75	0.00	-1.25	42.50
SRT [dB]	-13.93	-12.13	-7.92	-17.97	-16.96	-14.71
Age	42.88	40.94	59.60	33	32	74
Gender	1	1	1	0	1	1
%correct	89.68	88.64	84.09	90.91	90.54	90.91
#trials	80.69	69.34	102.60	77	74	77
total_time [s]	291.17	218.07	548.43	289.23	213.00	574.00
digitSpanScore	5.93	5.68	4.80	5.0	6.0	5.0
avgSingleDigitTimes_mean	1.34	1.16	1.50	0.84	1.19	1.63
family_history_HL	0	0	0	0	1	0
cardio	0	0	0	0	0	0
high_volume_exposure	0	0	0	1	1	0
education	1	1	3	2	2	3

Cluster 1 (69 points) indicates modest hearing and cognitive abilities without genetic predispositions to hearing loss. Cluster 2 (79 points) comprises slightly younger individuals with slightly better cognitive performance and shorter task times. Cluster 3 with only 5 elements, though smaller, represents older individuals with poorer hearing thresholds and longer task completion times.

Four Clusters:

The 3D representation of the 4 clusters is the following:

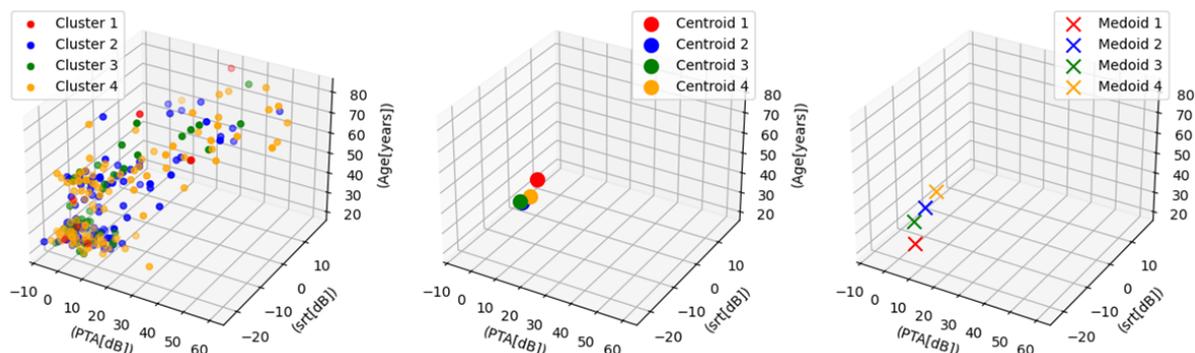


Figure 3.38: 3D representation of 4 clusters, with centroids and medoids highlighted (Whisper + risk factor + DST features).

Table 3.18: Centroids and Medoids table for 4 clusters.

	Centroids				Medoids			
	C11	C12	C13	C14	C11	C12	C13	C14
PTA [dB]	11.43	7.94	8.57	10.98	3.75	1.25	3.75	7.50
SRT [dB]	-10.38	-13.13	-14.49	-12.82	-17.38	-10.64	-17.32	-12.36
Age	49.43	39.73	41.46	43.15	22	32	33	44
Gender	1	1	1	1	1	1	0	0
%correct	85.45	89.20	89.54	88.72	85.25	90.28	89.61	87.50
#trials	101.43	68.58	82.85	76.92	93	72	77	80
total_time [s]	515.75	206.02	328.95	262.58	442.93	207.00	332.98	253.00
digitSpanScore	5.29	5.75	5.92	5.84	5.00	6.00	7.00	8.00
avgSingleDigitTimes_mean	1.43	1.12	1.41	1.31	1.39	1.19	1.77	1.33
family_history_HL	0	0	0	0	1	1	0	1
cardio	0	0	0	0	0	0	0	0
high_volume_exposure	0	0	0	0	0	1	0	0
education	3	1	1	1	1	2	1	2

Clusters 2 (99 points) and 3 (38 points) are quite similar in their medoids coordinates for age, while Cluster 1 (only 7 records) and 3 present the same value of PTA and really close values of SRT (even if Cluster 1 is younger and predominantly female); Cluster 4 is

the biggest, containing 112 elements. The clusters are mainly distinguished by the value of `avgSingleDigitTimes_mean`, `gender` and by the `digitSpanScore`. The inclusion of DST variables allows for a slight distinction between the clusters, although some of them are quite similar in terms of PTA, SRT, and Whisper factors. The total test time is higher for Cluster 1, indicating a subset of people that are young but have low values of cognitive performance (discrete typing time, low DSS), high number of trials and lowest %correct. Here the results using the dataset that contains only the ears with better PTA values.

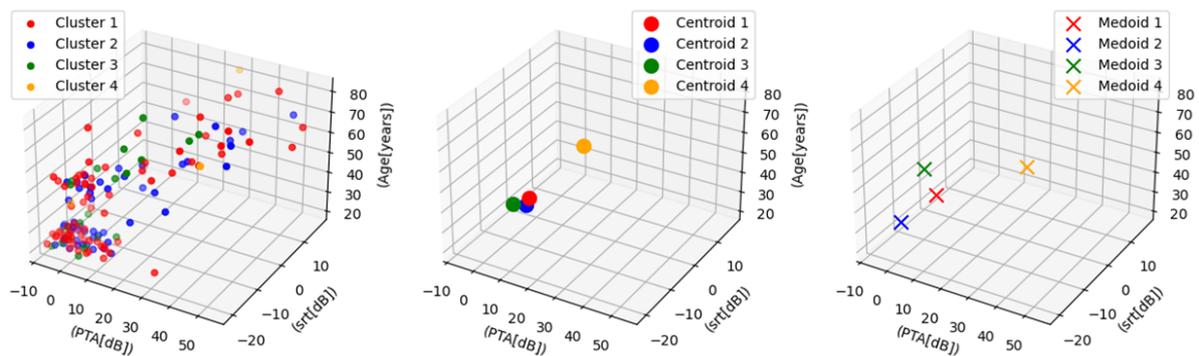


Figure 3.39: 3D representation of 4 clusters, with centroids and medoids highlighted, only ears with better PTA records (Whisper + risk factor + DST features).

Table 3.19: Centroids and Medoids table for 4 clusters, ears with better PTA only.

	Centroids				Medoids			
	C11	C12	C13	C14	C11	C12	C13	C14
PTA [dB]	9.92	8.96	6.74	21.67	7.50	-1.25	3.75	42.50
SRT [dB]	-12.52	-12.61	-15.34	-3.65	-12.36	-16.96	-12.79	-14.71
Age	43.47	39.68	42.52	64.33	44.00	32.00	56.00	74.00
Gender	1	1	1	1	0	1	1	1
%correct	89.39	88.49	89.75	80.83	87.50	90.54	89.29	90.91
#trials	77.16	67.09	85.65	107.67	80.00	74.00	84.00	77.00
total_time [s]	264.98	204.79	335.24	629.05	253.00	213.00	325.00	574.00
digitSpanScore	5.86	5.57	6.00	5.00	8.00	6.00	6.00	5.00
avgSingleDigitTimes_mean	1.32	1.09	1.39	1.42	1.33	1.19	1.79	1.63
family_history_HL	0	0	0	0	1	0	1	0
cardio	0	0	0	0	0	0	0	0
high_volume_exposure	0	0	1	0	0	1	0	0
education	1	1	1	3	2	2	2	3

Cluster 1 presents 74 elements, while Cluster 2 has 53, Cluster 3 contains 23 subject, and Cluster 4 only 3 records. The fourth cluster is observed to correspond to the oldest subjects, who have a low DSS, high PTA and typing time. Their test duration is also higher compared to the others. This cluster can thus be seen as representing older subjects who exhibit greater difficulties both in auditory and cognitive terms.

Five Clusters:

Here the 3D representation for 5 clusters is shown:

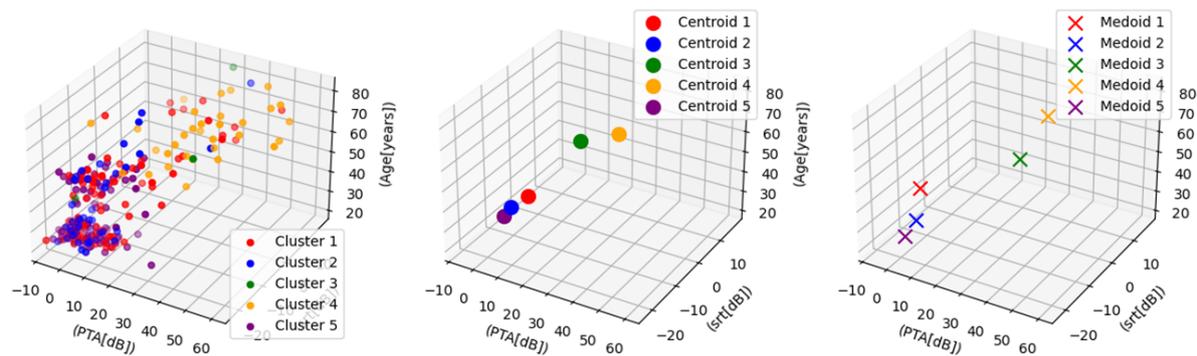


Figure 3.40: 3D representation of 5 clusters, with centroids and medoids highlighted, with Whisper + risk factor + DST features.

Table 3.20: Centroids and Medoids table for 5 clusters.

	Centroids					Medoids				
	C11	C12	C13	C14	C15	C11	C12	C13	C14	C15
PTA [dB]	9.33	5.08	21.67	36.25	2.78	5.00	3.75	42.50	38.75	-1.25
SRT [dB]	-12.65	-15.46	-3.65	-3.29	-15.91	-12.65	-16.89	-12.79	-14.71	-17.25
Age	42.13	38.20	64.33	72.28	32.90	49.00	33.00	56.00	78.00	23.00
Gender	1	1	1	1	1	1	0	1	1	1
%correct	89.11	89.60	80.83	88.28	89.01	91.43	89.61	90.91	87.69	89.74
#trials	66.87	85.31	107.67	65.94	80.15	70.00	77.00	77.00	65.00	78.00
total_time [s]	201.22	341.57	629.05	276.22	255.84	200.00	332.98	574.00	277.00	253.00
digitSpanScore	5.61	6.07	5.00	4.47	6.28	6.00	7.00	5.00	4.00	5.00
avgSingleDigitTimes_mean	1.13	1.32	1.42	1.76	1.18	1.05	1.77	1.63	1.12	1.28
family_history_HL	0	0	0	0	0	1	0	0	0	1
cardio	0	0	0	0	0	0	0	0	0	0
high_volume_exposure	0	0	0	0	0	0	0	1	0	0
education	1	1	3	1	1	2	1	3	3	1

The analysis of the five clusters reveals important insights into the relationships between age, auditory health, and cognitive performance. Cluster 1 (82 records), composed of middle-aged individuals in medoids coordinates, shows balanced auditory and cognitive abilities, suggesting a relatively healthy demographic. Cluster 2 (of 45 people), while also younger (33 years), exhibits higher DSS but lower typing time compared to the first one. Cluster 3 stands out due to its older age group (even if composed of only 3 subjects),

showing significantly poorer auditory health and somewhat diminished cognitive performance. This highlights the impact of aging on these faculties. Cluster 4, the oldest group of 32 people, faces the greatest challenges on the DST and also with hearing, considered that the values of PTA (38.75 dB) and SRT (0.33 dB) are not representative of a healthy profile. None of these last two clusters present family history of hearing loss.

Conversely, Cluster 5, consisting of the youngest individuals (and the biggest cluster, of 104 points), shows the best auditory health, and faster response times and high %correct. Although they have low values for the DSS, and typing time that is discrete.

The inclusion of the fifth cluster seems to illustrate the clear distinctions among different age groups, emphasizing how age-related decline impacts both hearing and cognitive abilities.

Cluster 3 and 4, that seem to be similar in terms of age, gender, also DSS, are mainly distinguished by their typing time, %correct and SRT, other than total time. The fourth cluster is faster even if resulting in lower performance of correct responses and DST score.

Here the results using only ears with better PTA records.

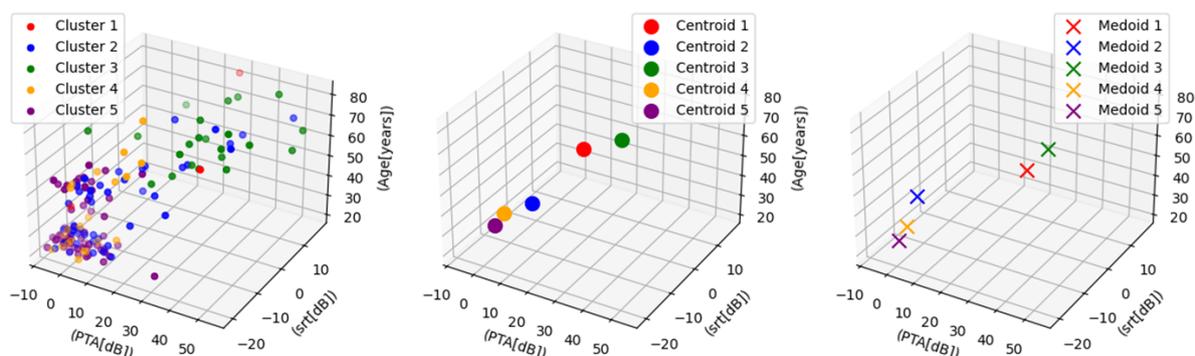


Figure 3.41: 3D representation of 5 clusters, with centroids and medoids highlighted, with Whisper + risk factor + DST features, using only ears with better PTA records.

Table 3.21: Centroids and Medoids table for 5 clusters, ears with better PTA only.

	Centroids					Medoids				
	C11	C12	C13	C14	C15	C11	C12	C13	C14	C15
PTA [dB]	21.67	10.48	34.77	4.17	0.79	42.50	5.00	43.75	1.25	-2.50
SRT [dB]	-12.65	-11.90	-2.97	-16.09	-16.21	-14.71	-16.89	-8.11	-17.17	-16.84
Age	64.33	42.23	72.55	39.29	32.05	74.00	49.00	77.00	33.00	24.00
Gender	1	1	1	1	1	1	1	1	0	0
%correct	80.83	88.07	88.37	89.76	89.91	90.91	91.43	89.04	90.79	88.89
#trials	107.67	65.39	66.09	86.90	80.63	77.00	70.00	73.00	76.00	81.00
total_time [s]	629.05	199.80	274.50	337.70	257.96	574.00	200.00	277.00	314.29	248.00
digitSpanScore	5.00	5.39	4.64	6.14	6.33	5.00	6.00	5.00	7.00	7.00
avgSingleDigitTimes_mean	1.42	1.14	1.62	1.77	1.16	1.63	1.05	1.24	1.77	1.23
family_history_HL	0	0	0	0	0	0	1	0	0	0
cardio	0	0	0	0	0	0	0	1	0	0
high_volume_exposure	0	0	0	0	0	0	0	0	0	0
education	3	1	1	1	1	3	2	1	1	2

Cluster 1 has only 3 elements, Cluster 2 contains 44 records, Cluster 3 has 22, Cluster 4 contains 21 elements, and Cluster 5 is composed of 63 points.

Considering only ears with better PTA records, distinction based on Age, PTA, trials and total_time is still present. Cluster 3 and 5 show similar values of digit time (1.24 vs 1.23), and also total_time (277.000 vs 248.000), even if their auditory performance (PTA and SRT) and age (77 vs 24) are completely far from each other.

The two analysis with 5 clusters (all dataset and ears with better PTA only) highlights how even within groups selected for their good auditory health, age remains a pivotal factor influencing cognitive performance. Younger individuals consistently outperform older ones in both cognitive and auditory tests, and older individuals show greater difficulties and longer test durations, regardless of the PTA values. This underscores the intertwined nature of auditory and cognitive decline with aging, even if there are still cases of discrete performances even for older people.

3.3.5. Classification adding DST and risk factors:

In this section, the results of the classification of clusters derived from the extended dataset considering also DST and risk factor features are reported. Similar to the approach taken for the dataset with only Whisper features (Section 3.3.3), the dataset considering also DST and risk factor features was divided into training and testing sets using three different splitting methods:

- train = 80% (213 records), test = 20% (53 records)
- train = ears with better PTA (153 records), test = worse PTA (113 records)
- train = dataset without Oldenburg data (232 records), test = only Oldenburg records (34 records)

Classification for split: TRAINING = 80%, TEST = 20%

Table 3.22 reports the mean and standard deviation of the performance metrics on the 5 random splits in the external training and external test sets (Nested CV, see Section 3.3.3 for a detailed description).

The parameter grid used for all the algorithm is exactly the same that has been used when only Whisper features were considered. Below the best parameters are reported.

Random Forest: The optimal parameters obtained were: `max_depth=10` (for 3 clusters), `None` (for 4), `10` (for 5), `min_samples_leaf=1` (for 3 clusters), `1` (for 4), `1`(for 5), `min_samples_split=2` (for 3 clusters), `2` (for 4), `10` (for 5), `n_estimators=20` (for 3 clusters), `150` (for 4), `100` (for 5).

SVM: The best parameters were found to be, for all the number of clusters, `C = 100` (only for 4 cluster `C=10`) and linear kernel.

KNN: The best parameter has been `7 n_neighbors` for 3 clusters, and it was `9` for 4 and 5 clusters.

Here, the tables with macro-averaging performance metrics on external training and external test set are reported.

Table 3.22: Classification report: Mean and standard deviation of performance metrics across 5 random splits into training and external test sets. Splitting method: training = 80%, test = 20%, 3 clusters (Whisper + risk factor + DST features).

Metric	Train Mean	Train Std	Test Mean	Test Std
<i>Random Forest</i>				
Accuracy (Macro)	0.999	0.002	0.962	0.032
Precision (Macro)	0.999	0.001	0.705	0.149
Recall (Macro)	0.999	0.002	0.729	0.136
F1-Score (Macro)	0.999	0.002	0.716	0.143
<i>SVM</i>				
Accuracy (Macro)	1.000	0.000	0.991	0.012
Precision (Macro)	1.000	0.000	0.992	0.010
Recall (Macro)	1.000	0.000	0.978	0.032
F1-Score (Macro)	1.000	0.000	0.983	0.023
<i>KNN</i>				
Accuracy (Macro)	0.833	0.029	0.745	0.065
Precision (Macro)	0.757	0.169	0.633	0.181
Recall (Macro)	0.618	0.055	0.569	0.131
F1-Score (Macro)	0.647	0.082	0.570	0.139

Random Forest achieved high accuracy on both training (0.999) and test (0.962) sets. Precision, recall, and F1-score decrease within the test set. SVM demonstrates high accuracy on training (1.00) and test (0.991) sets, showing signs of overfitting to the training data. Precision, recall, and F1-score are consistently high, suggesting strong model consistency. KNN shows lower performance compared to Random Forest and SVM, with accuracy at 0.833 (training) and 0.745(test). Precision, recall, and F1-score metrics are also lower, indicating some challenges in generalizing to the test set.

Adding DST and risk factors features leads to a slight decrease in precision, recall, F1-score metrics, indicating some impact from additional features for Random Forest algorithm. Random Forest exhibits a small decrease in accuracy in the test set but precision, recall and F1-score have higher decrease with some variability, possibly due to the inclusion of new features. SVM shows first signs of overfitting in the training set. KNN also shows decrease in performance adding the new features, probably showing less capability of handling them, suggesting challenges in adapting to the new dataset charac-

teristics. While models trained on the dataset with risk and cognitive factors generally maintain strong performance, there are indications of variability and decreases in performance metrics compared to models trained only on Whisper features. This suggests that incorporating additional features has both positive and challenging aspects for model generalization.

Here the performance when considering 4 clusters (Table 3.23):

Table 3.23: Classification report: Mean and standard deviation of performance metrics across 5 random splits into training and external test sets. Splitting method: training = 80%, test = 20%, 4 clusters (Whisper + risk factor + DST features).

Metric	Train Mean	Train Std	Test Mean	Test Std
<i>Random Forest</i>				
Accuracy (Macro)	0.993	0.007	0.929	0.014
Precision (Macro)	0.992	0.008	0.785	0.138
Recall (Macro)	0.919	0.081	0.800	0.103
F1-Score (Macro)	0.935	0.068	0.789	0.119
<i>SVM</i>				
Accuracy (Macro)	0.988	0.015	0.915	0.035
Precision (Macro)	0.992	0.010	0.880	0.091
Recall (Macro)	0.990	0.013	0.835	0.089
F1-Score (Macro)	0.991	0.012	0.845	0.082
<i>KNN</i>				
Accuracy (Macro)	0.706	0.074	0.538	0.039
Precision (Macro)	0.592	0.144	0.475	0.134
Recall (Macro)	0.547	0.114	0.457	0.081
F1-Score (Macro)	0.560	0.127	0.452	0.098

Random Forest and SVM achieve robust performance with high accuracy (0.929 and 0.915 respectively) on the test set, though Random Forest shows higher variability and especially lower values of precision, recall and F1-score. SVM keeps robust values for these metrics. However, KNN exhibits lower overall accuracy (0.538) and precision metrics, suggesting challenges in distinguishing between the additional cluster categories.

Here the performance when considering 5 clusters (Table 3.24):

Table 3.24: Classification report: Mean and standard deviation of performance metrics across 5 random splits into training and external test sets. Splitting method: training = 80%, test = 20%, 5 clusters (Whisper + risk factor + DST features).

Metric	Train Mean	Train Std	Test Mean	Test Std
<i>Random Forest</i>				
Accuracy (Macro)	1.000	0.000	0.901	0.034
Precision (Macro)	1.000	0.000	0.847	0.096
Recall (Macro)	1.000	0.000	0.828	0.095
F1-Score (Macro)	1.000	0.000	0.828	0.098
<i>SVM</i>				
Accuracy (Macro)	0.993	0.014	0.925	0.023
Precision (Macro)	0.995	0.011	0.890	0.080
Recall (Macro)	0.993	0.014	0.898	0.064
F1-Score (Macro)	0.994	0.013	0.887	0.067
<i>KNN</i>				
Accuracy (Macro)	0.627	0.027	0.495	0.086
Precision (Macro)	0.577	0.021	0.528	0.139
Recall (Macro)	0.459	0.022	0.434	0.139
F1-Score (Macro)	0.483	0.021	0.421	0.128

Random Forest and SVM exhibit strong performance on the test set with accuracy of 1.00 and 0.993 respectively, indicating robust ability to classify across multiple clusters. Here Random Forest has higher values for the other metrics compared to the other two configurations of clusters, and some sign of overfitting on training. However, KNN shows lower overall accuracy (0.495) and macro-averaged metrics, suggesting challenges in distinguishing between the additional cluster categories compared to the other models.

Classification for split: TRAINING = ears with better PTA, TEST = worse PTA

The classification results on the entire dataset are presented, using a different splitting method compared to the previous case. In this method, the training set includes records where the ears with better PTA was selected, while the remaining records (worse PTA) constitute the test set.

For each classification method, the accuracy on the training and test set obtained with 5-fold-CV on all the combinations of parameters is reported. Moreover, the performance of the model with the best parameters has been applied to the training set and evaluated on the test set with the same performance indicators (but not macro-averaged), now based on the best performance with the best set of parameters.

The best parameters here are: *Random Forest*: $\text{max_depth}=5$, $\text{min_sample_leaf}=1$, $\text{min_sample_split}=5$, $\text{n_estimators}=30$.

Mean train accuracy (CV=5 over all parameters) = 0.885 ± 0.092 . Mean test accuracy (CV=5 over all parameters) = 0.766 ± 0.183 .

SVM: $C = 100$, kernel = 'linear'.

Mean train accuracy (CV=5 over all parameters) = 0.815 ± 0.181 . Mean test accuracy (CV=5 over all parameters) = 0.710 ± 0.176 .

KNN: $\text{n_neighbors} = 9$.

Mean train accuracy (CV=5 over all parameters) = 0.843 ± 0.029 . Mean test accuracy (CV=5 over all parameters) = 0.733 ± 0.080 .

With the best performing set of parameters, Random Forest and SVM achieved high accuracies on both training (both 0.99) and test sets (both 0.96), showcasing robustness in classifying. KNN, while slightly less accurate (0.78 on training, 0.82 on test), demonstrates competitive performance, particularly in Cluster 2, but shows challenges in distinguishing Cluster 1 effectively. Also Random Forest has difficulties in classifying Cluster 1, with low value of precision, recall and F1-score (it contains only 3 elements). Overall, SVM maintain strong precision and recall metrics across clusters, indicating their suitability for this split methodology.

The best parameters for 4 clusters are: *Random Forest*: $\text{max_depth}=10$, $\text{min_sample_leaf}=1$, $\text{min_sample_split}=2$, $\text{n_estimators}=150$.

Mean train accuracy (CV=5 over all parameters) = 0.833 ± 0.124 . Mean test accuracy (CV=5 over all parameters) = 0.641 ± 0.175 .

SVM: $C = 1$, kernel = 'linear'.

Mean train accuracy (CV=5 over all parameters) = 0.710 ± 0.235 . Mean test accuracy (CV=5 over all parameters) = 0.481 ± 0.180 .

KNN: $\text{n_neighbors} = 9$.

Mean train accuracy (CV=5 over all parameters) = 0.707 ± 0.049 . Mean test accuracy (CV=5 over all parameters) = 0.520 ± 0.072 .

With the best performance, Random Forest achieves perfect training accuracy across all clusters, demonstrating to the test set an overall accuracy of 0.96. It maintains high precision, recall, and F1-score for most clusters, although it struggles with Cluster 2 in the test set where precision drops to 0. SVM also shows strong performance with 0.85 accuracy on the test set, with consistent precision and recall scores across most clusters, with some variability in Cluster 2. SVM is still the algorithm that seems to have less difficulties in classification with the new features. KNN performs less consistently, achieving 0.55 accuracy on the test set, with notable challenges in achieving high precision across all clusters, particularly in Cluster 2 where it reports 0 precision.

Adding a fourth cluster seems to increase the challenges of classification, as not only the number of records is lower but also the numerosity of each cluster decreases.

The best parameters for 5 clusters are: *Random Forest*: `max_depth=10, min_sample_leaf=1, min_sample_split=5, n_estimators=150`.

Mean train accuracy (CV=5 over all parameters) = 0.909 ± 0.084 . Mean test accuracy (CV=5 over all parameters) = 0.718 ± 0.221 .

SVM: `C = 100, kernel = 'linear'`.

Mean train accuracy (CV=5 over all parameters) = 0.741 ± 0.259 . Mean test accuracy (CV=5 over all parameters) = 0.460 ± 0.192 .

KNN: `n_neighbors = 5`.

Mean train accuracy (CV=5 over all parameters) = 0.667 ± 0.052 . Mean test accuracy (CV=5 over all parameters) = 0.513 ± 0.071 .

With the best performance, Random Forest and SVM achieve perfect training accuracy across all clusters. In the test set, Random Forest achieves an overall accuracy of 0.90, with notable performance in Clusters 3 and 4. SVM performs even better with an accuracy of 0.94, demonstrating robustness across all clusters, particularly excelling in Clusters 4 and 5. Conversely, KNN exhibits more variability in its performance. While achieving 0.88 training accuracy, it struggles in the test set with an overall accuracy of 0.47. KNN shows mixed results across clusters, with notable challenges in maintaining precision and recall, especially in Clusters 2 and 5 where precision drops significantly. Cluster 1 is not present in the test set, because of the low number of records in it, it was not represented in the test. (Support for training set: 3, 26, 55, 50, 19. Support for test

set: 0, 18, 46, 26, 23.)

Classification for split: TRAINING = previous data, TEST = Oldenburg data

Here are the classification results considering the last splitting method. Specifically, the training set comprises records that have been taken until December 2023 only in Italy, while the remaining records form the test set correspond to the data acquired in Oldenburg.

For each classification method, the accuracy on the training and test set obtained with 5-fold-CV on all the combinations of parameters is reported. Moreover, the performance of the model with the best parameters has been applied to the training set and evaluated on the test set and is commented.

The best parameters here are: *Random Forest*: `max_depth=None`, `min_sample_leaf=1`, `min_sample_split=5`, `n_estimators=30`.

Mean train accuracy (CV=5 over all parameters) = 0.985 ± 0.009 . Mean test accuracy (CV=5 over all parameters) = 0.963 ± 0.016 .

SVM: `C = 100`, `kernel = 'linear'`.

Mean train accuracy (CV=5 over all parameters) = 0.811 ± 0.154 . Mean test accuracy (CV=5 over all parameters) = 0.770 ± 0.126 .

KNN: `n_neighbors = 9`.

Mean train accuracy (CV=5 over all parameters) = 0.844 ± 0.027 . Mean test accuracy (CV=5 over all parameters) = 0.733 ± 0.066 .

In the best performing set of parameters, Random Forest achieves perfect training accuracy (1.00) across all clusters, but in the test set, it achieves an overall accuracy of 0.87. It performs struggles in Cluster 2, where all metrics are zero due to the model's inability to classify any instances correctly in this cluster. SVM, on the other hand, also achieves perfect training accuracy (1.00) and performs flawlessly in the test set with an accuracy of 1.00. It maintains high precision, recall, and F1-scores across all clusters, indicating probability some sort of overfitting. KNN exhibits decent training accuracy (0.85) but demonstrates poorer performance in the test set with an accuracy of 0.53. It faces challenges particularly in Cluster 2, where all metrics are zero, suggesting significant misclassification. This could be due to insufficient representation or distinct characteristics of Cluster 2 in the Oldenburg data that weren't well captured by the model during

training.

The best parameters for 4 clusters are: *Random Forest*: $\text{max_depth}=10, \text{min_sample_leaf}=1, \text{min_sample_split}=2, \text{n_estimators}=30$.

Mean train accuracy (CV=5 over all parameters) = 0.982 ± 0.013 . Mean test accuracy (CV=5 over all parameters) = 0.912 ± 0.036 .

SVM: $C = 10, \text{kernel} = \text{'linear'}$.

Mean train accuracy (CV=5 over all parameters) = 0.726 ± 0.222 . Mean test accuracy (CV=5 over all parameters) = 0.648 ± 0.167 .

KNN: $\text{n_neighbors} = 7$.

Mean train accuracy (CV=5 over all parameters) = 0.734 ± 0.061 . Mean test accuracy (CV=5 over all parameters) = 0.561 ± 0.064 .

With the best performance, Random Forest and SVM achieve strong learning from the training data across all clusters. However, in the test set, Random Forest achieves an overall accuracy of 0.87, performing well in Cluster 3 and 4 with high precision, recall, and F1-scores. SVM excels further with 0.97 accuracy in the test set, maintaining high performance across all clusters. Conversely, KNN shows weaker training accuracy (0.69) and struggles in the test set with only 0.23 accuracy. It particularly performs poorly in Cluster 1 and 3, where precision, recall, and F1-scores are notably low, indicating significant misclassification. Cluster 1 and 2 are not in the test because of lack of data points. This suggests that KNN might not generalize well to the Oldenburg data, possibly due to differences in data distributions or insufficient representation of certain clusters in the training data. (Support for training set: 4, 98, 30, 104. Support for test set: 0, 0, 18, 8.)

The best parameters for 5 clusters are: *Random Forest*: $\text{max_depth}=15, \text{min_sample_leaf}=2, \text{min_sample_split}=5, \text{n_estimators}=100$.

Mean train accuracy (CV=5 over all parameters) = 0.977 ± 0.019 . Mean test accuracy (CV=5 over all parameters) = 0.869 ± 0.051 .

SVM: $C = 1, \text{kernel} = \text{'linear'}$.

Mean train accuracy (CV=5 over all parameters) = 0.681 ± 0.234 . Mean test accuracy (CV=5 over all parameters) = 0.603 ± 0.183 .

KNN: $\text{n_neighbors} = 9$.

Mean train accuracy (CV=5 over all parameters) = 0.664 ± 0.045 . Mean test accuracy (CV=5 over all parameters) = 0.518 ± 0.059 .

With the best set of parameters, Random Forest achieves a high training accuracy of 0.98, with strong precision, recall, and F1-scores for Clusters 1, 2, 4, and 5. However, it struggles with Cluster 3. In the test set, Random Forest maintains an accuracy of 0.97, highlighting robust generalization except for Cluster 3, where performance is notably lower, but Cluster 1 and 4 are not represented. SVM, with a slightly lower training accuracy of 0.96, performs consistently well across Clusters 1, 2, 4, and 5, but like Random Forest, it faces challenges with Cluster 3 in the test set and has no Cluster 1 and 4 there. Despite this, SVM achieves an overall test accuracy of 0.77, indicating reasonable generalization across most clusters except for Cluster 3. Conversely, KNN shows weaker performance with a training accuracy of 0.73. It struggles across several clusters in the test set, especially in Cluster 3 where precision, recall, and F1-scores are consistently low.

The presence of the fifth cluster seems to increase the challenges for classification, always due to the small test set.

3.4. Comparison among Whisper, DTT and OLSA:

In this section, the comparative analysis conducted on the Whisper, OLSA, and DTT SIN tests is reported, analyzing the dataset acquired during phase 2 in Oldenburg, which includes 17 subjects (see Section 2.2.2 for a detailed description of the acquisition protocol).

The variables extracted for the Whisper test are the same as those reported in the previous sections, while for OLSA and DTT, an XML file was extracted with numerous variables, and the most significant ones were manually selected for comparison (explanation of the variables in Figure 2.9). Subsequently, the main graphs and tables used for the analysis are presented. It is recognized that the dataset's size is definitely a limitation; therefore, the results obtained and shown in this thesis will need to be confirmed and validated with further studies on a larger sample.

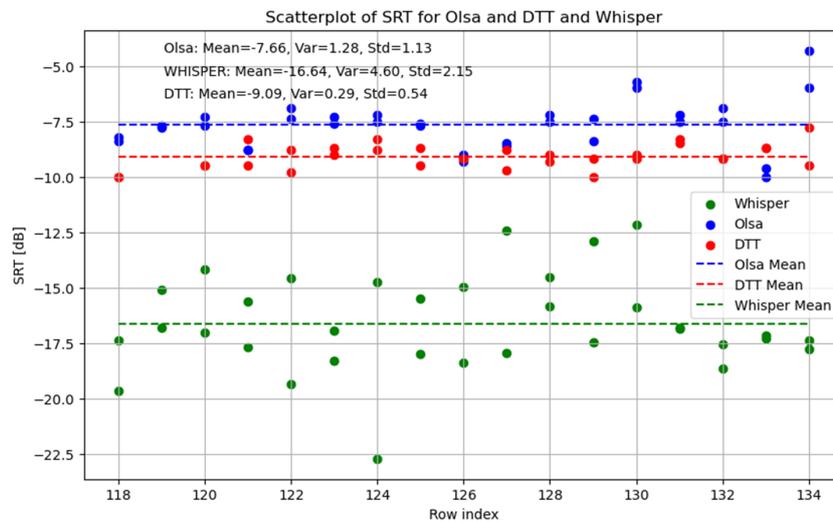


Figure 3.42: Scatterplot of the SRT for the 3 tests, both ears, 17 subjects in total. The x-axis shows the Whisper ID for each subject.

Figure 3.42 shows, for each subject (identified by their Whisper ID on the x-axis) and for both ears, the SRT values (y-axis) for the 3 tests. In green, the Whisper SRT values, which are significantly lower compared to the other two tests. Closer to each other, however, are the red dots (DTT) and blue dots (OLSA). OLSA reports higher average SRT values compared to DTT. Meanwhile, the most significant variance (4.60 dB) and standard deviation (2.15 dB) are observed in the Whisper test.

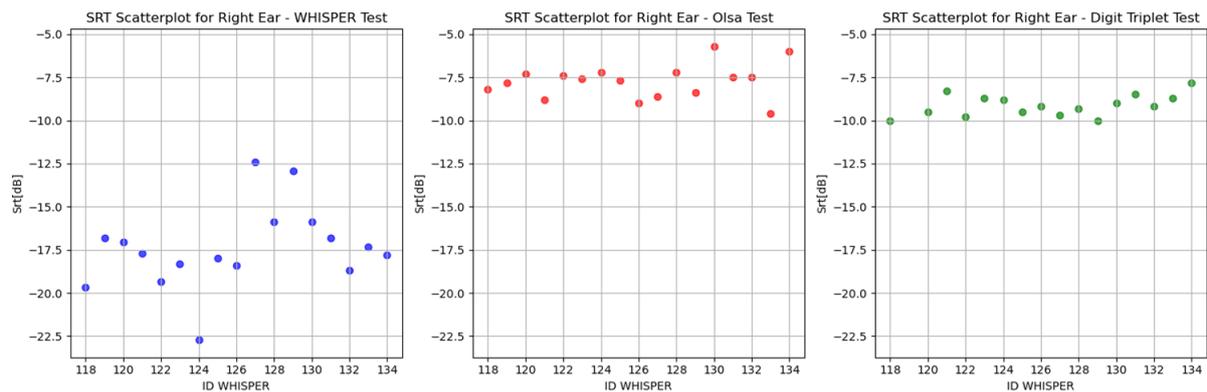


Figure 3.43: Scatterplot of SRT for each single test - right ear, 17 ears.

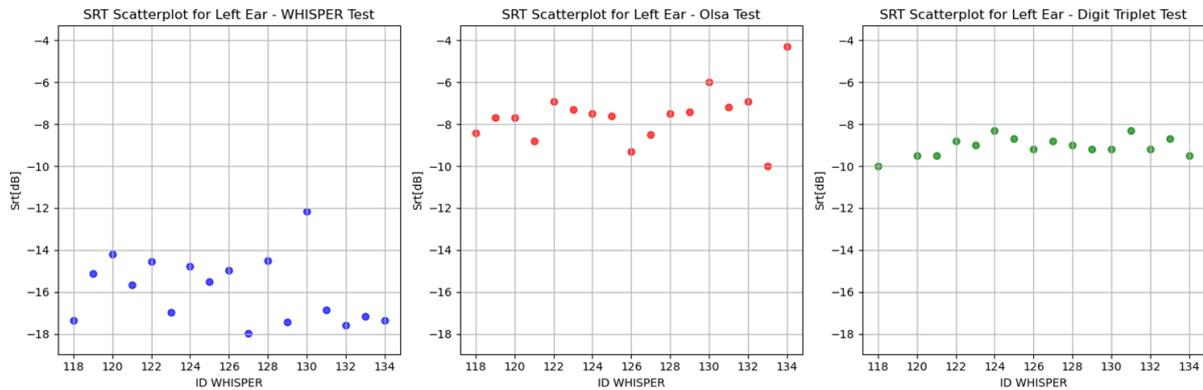


Figure 3.44: Scatterplot of SRT for each single test - left ear, 17 ears.

Figures 3.43 and 3.44 separately depict the SRT values for each individual test and for each ear. From these graphs, it is immediately observable, as previously noted, that Whisper exhibits lower SRT values compared to the other two tests and a significantly higher variability. Furthermore, what this graphs highlight is that lower SRT values, especially in Whisper, are achieved for the right ear, which was generally the first to be tested. Nevertheless, all values confirm that the tested subjects are normal hearing.

It was then decided to examine the Whisper test individually, specifically regarding the presentation of single VCVs during the test for each subject. For each subject, all the presented VCVs (i.e., the presented stimuli) during the adaptive procedure were collected and compared with the option selected by the user. In cases where the selected option differed from the presented stimulus, this mismatch was recorded. The result is what has been defined as the percentage error, which is the percentage of mismatches between the presented stimulus and the selected option relative to the total number of stimuli (which is different for each subject since, as explained in Section 2.1.1, it is an adaptive procedure).

If ($VCV_{\text{presented}} \neq VCV_{\text{selected}}$):

mismatch_count+ = 1

$$\text{Error Percentage} = \left(\frac{\text{mismatch_count}}{\text{total_stimuli}} \right) \times 100$$

In Figure 3.45, 3.46, the error percentages for each subject for the right and left ear are reported. It is observed that the error percentages are slightly higher for the left ear

overall, indicating a higher margin of error.

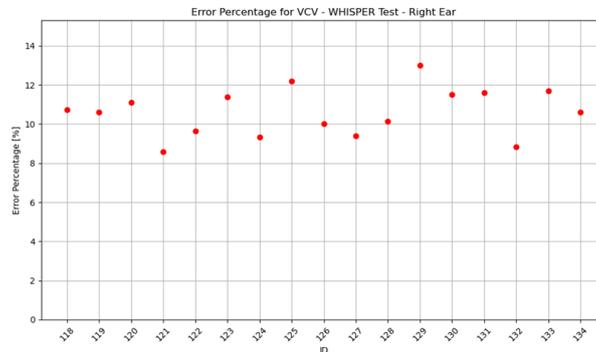


Figure 3.45: Scatterplot of % Error for Right Ear, Whisper.

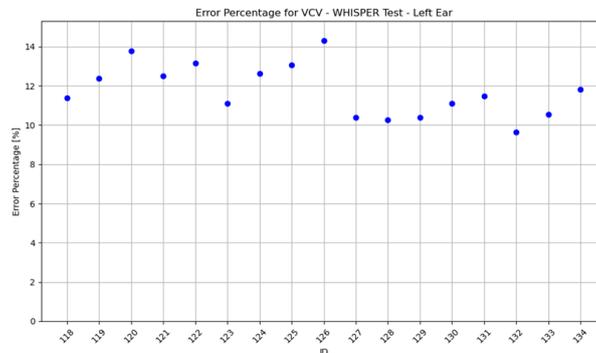


Figure 3.46: Scatterplot of % Error for Left Ear, Whisper.

It is interesting to observe which consonants were most commonly mistaken by the various subjects, for both ears (Figure 3.47, 3.48). "M", "F", and "R" were consistently the most critical in every case (with a simple inversion between "F" and "M" depending on the ear).

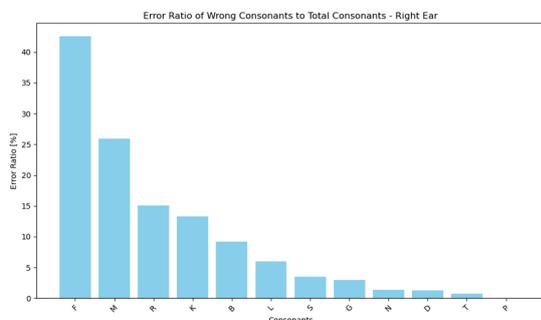


Figure 3.47: Most mistaken consonants Whisper - Right Ear.

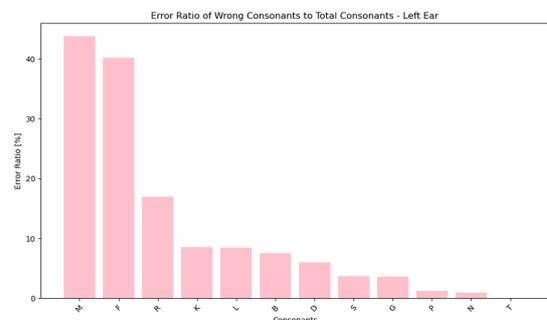


Figure 3.48: Most mistaken consonants Whisper - Left Ear.

In order to perform a similar analysis for the DTT and OLSA tests, the so-called errors were observed in both cases. Specifically, the variable *"SelectedWords"* was analyzed. *"SelectedWords"* contains the "highlighted" (=selected) words at the moment when "OK" was clicked. For OLSA and DTT, this variable contains the number of word items and the "correct" target word when it was selected (recognized correctly), and a minus "-" when it was not selected. When measuring a matrix test in closed format, the selections are "translated" to this format after clicking OK, i.e., it does not contain "wrong" selected words but it contains a minus if the corresponding word was not correctly guessed.

Therefore, the error percentage was calculated by considering both incorrect responses (i.e., a word not matching the provided one) and non-responses (as both cases are indicated with the symbol "-" in the variable). To obtain the error percentage, the total number of stimuli for both tests was considered, i.e., the number of words for OLSA (100, i.e., 5 words per 20 sentences) and 81 digits for DTT (3 numbers per 27 sequences). This was done for each subject, and the values obtained are shown separately for each ear, highlighting the mean value with a line.

A significantly higher error percentage (reaching 50%) was observed in OLSA (Figure 3.49) than in DTT (Figure 3.50).

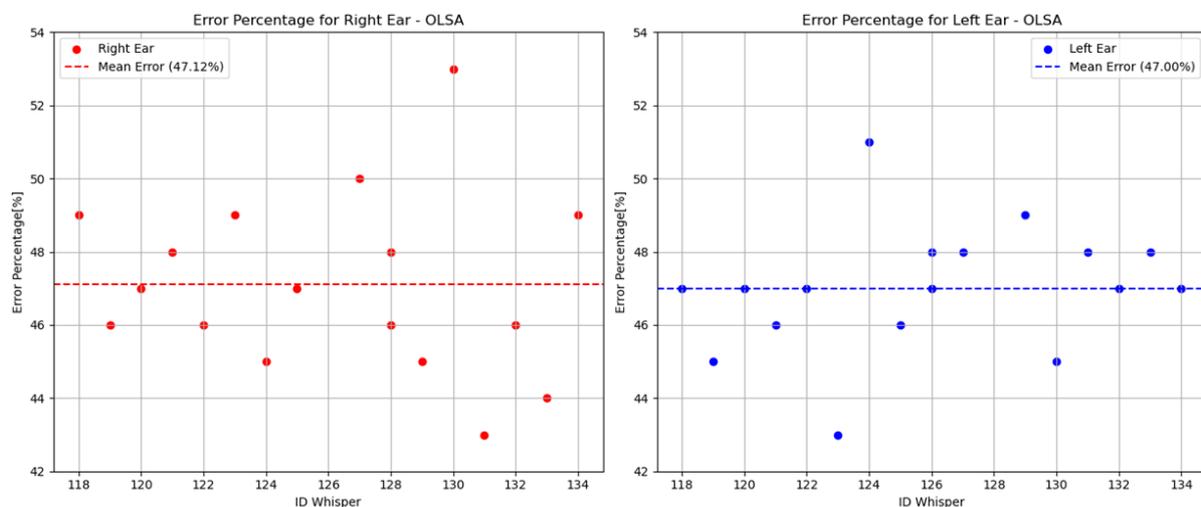


Figure 3.49: Error Percentage for both ears - OLSA

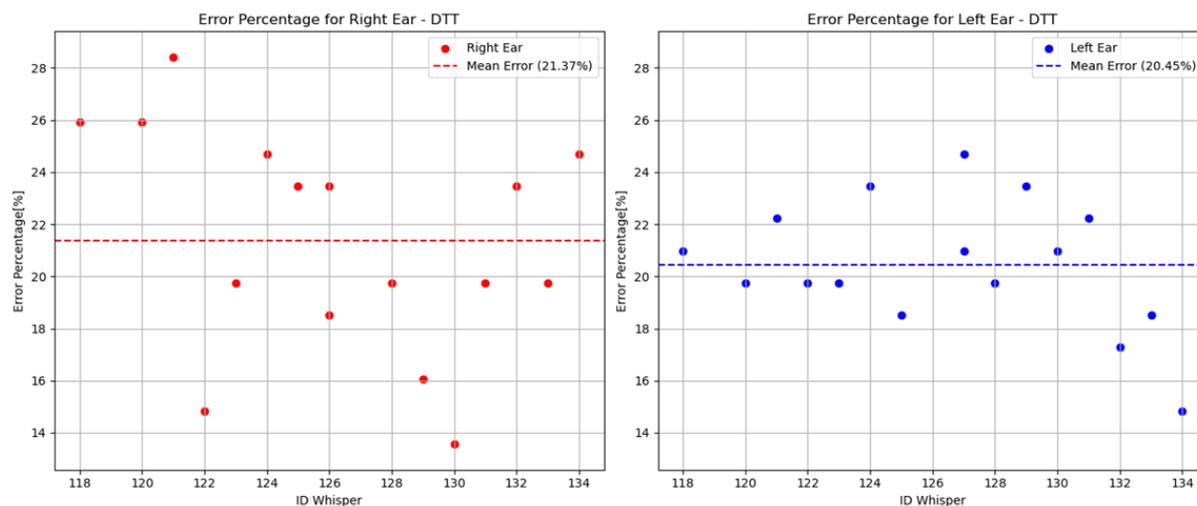


Figure 3.50: Error Percentage for both ears - DTT

Figure 3.51 depicts (for each of the three tests), the subject ID on the x-axis, and on the y-axis the delta of SRT, which was calculated as the difference between SRT for the right ear and SRT for the left ear, for each test and subject. The deltas are shown with sign. Since the dataset includes normal hearing subjects, the SRT is a negative value in all the cases, so positive deltas indicate subjects with a lower SRT value for the left ear compared to the right ear. The majority of subjects show a negative delta, confirming that the SRT for the right ear tends to be lower. This graph illustrates the asymmetry between the right and left ears in the subjects, which is more evident with the Whisper test. Specifically, some subjects, such as 122, 124, and 130, have deltas around 4dB. Subjects who performed Whisper as last test (i.e., 120, 121, 122, 126, 127, 129) were then observed to see if the test order had an effect on SRT performance. As it can be observed, significant delta values are also seen for example in the case of subject 124, who did not perform Whisper as last test, concluding that the test order does not have a particular influence.

The maximum delta value reached for Whisper is 5.56 dB, while for DTT is 1.70 dB, and for OLSA is 0.40 dB. The minimum delta value is -7.97 dB for Whisper, -1.00 dB for DTT, and -1.70 dB for OLSA. The average delta value for Whisper is -1.50 dB, for DTT is -0.069 dB, and for OLSA is -0.15 dB. OLSA and DTT do not show significant asymmetry, compared to Whisper.

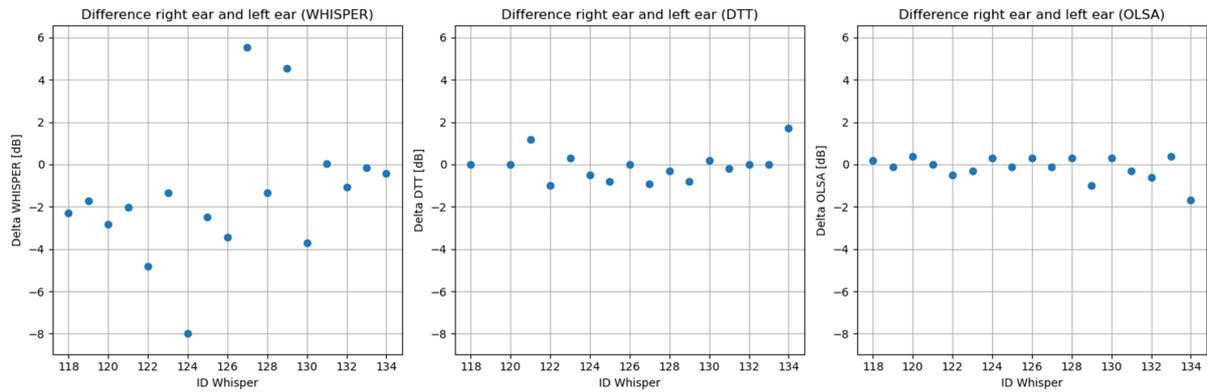


Figure 3.51: Delta (right ear - left ear) of SRT for the three tests.

In Figure 3.52 instead, the deltas of Pure Tone Threshold Audiometry values are shown for individual frequencies, ranging from 500 Hz up to high frequencies (8000Hz). This delta in audiometric thresholds is also calculated as the difference between the right and left ears.

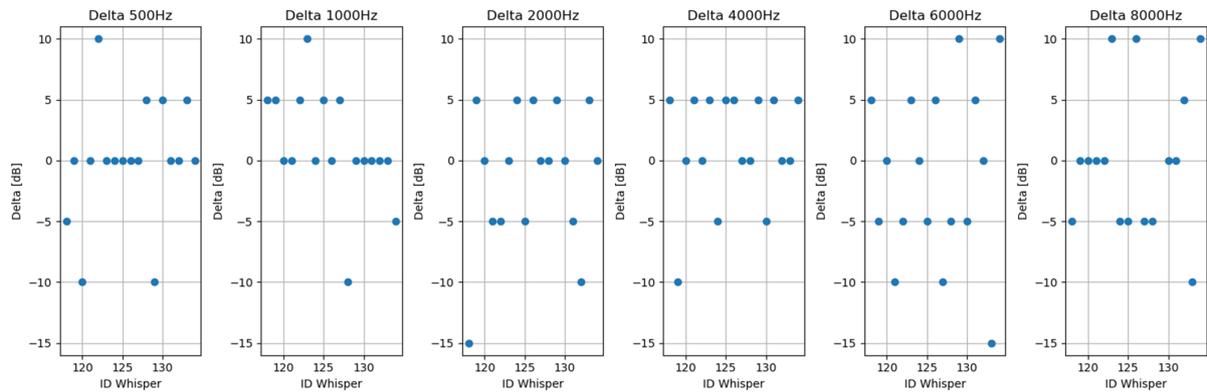


Figure 3.52: Delta of Pure Tone Threshold Audiometry values (right ear - left ear) for single frequencies for the three tests.

In Table 3.25, the minimum, maximum, and average values for each frequency are reported, and it is observed that they are generally negative. Especially in frequencies below 4000Hz, many subjects exhibit null deltas, indicating good symmetry between the right and left ears. However, this becomes less evident as the frequencies increase.

The values of the PTA delta were computed by averaging over the mid frequencies, namely from 500 to 4000Hz (Figure 3.53), and including also high frequencies in the average, up to 8000Hz (Figure 3.54). A more scattered distribution is observed compared to the deltas of individual frequencies, as in that case the values went in increments of 5. The

Table 3.25: Delta of PTA min, max, mean values for all frequencies up to 8000 Hz.

<i>Delta</i>	<i>Frequency (Hz)</i>					
	500	1000	2000	4000	6000	8000
<i>Min</i>	-10.0	-10.0	-15.0	-10.0	-15.0	-10.0
<i>Max</i>	10.0	10.0	5.0	5.0	10.0	15.0
<i>Mean</i>	0.0000	1.1765	-1.1765	1.1765	-1.1765	0.8834

deltas that also consider the high frequencies have higher values (i.e., more asymmetry), while considering only the mid frequencies, they remain more distributed around zero.

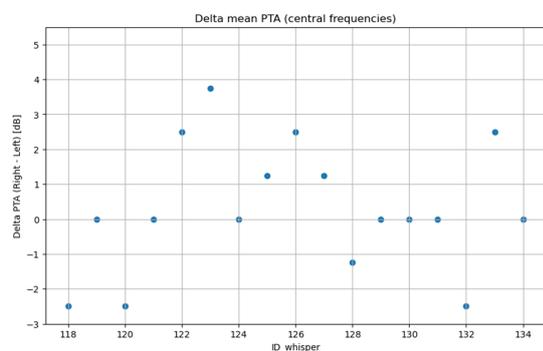


Figure 3.53: Delta PTA for central frequencies (from 500 to 4000 Hz)

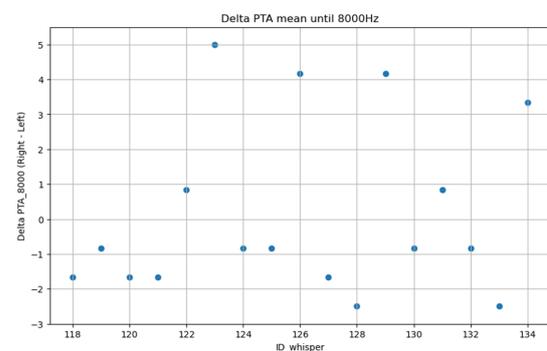


Figure 3.54: Delta PTA until high frequencies (8000Hz)

The correlation between the SRTs extracted from the three tests and also between the SRTs and the PTA was then observed. Due to the non-normal distribution of the data, it was deemed more meaningful to compute Spearman correlation (Figure 3.55). PTA seems to be better correlated to SRT in OLSA but in general there are no correlation values above 0.30. Hence, it can be said that the SRT extracted by the three tests do not show a noteworthy correlation with PTA when considering normal hearing subjects.

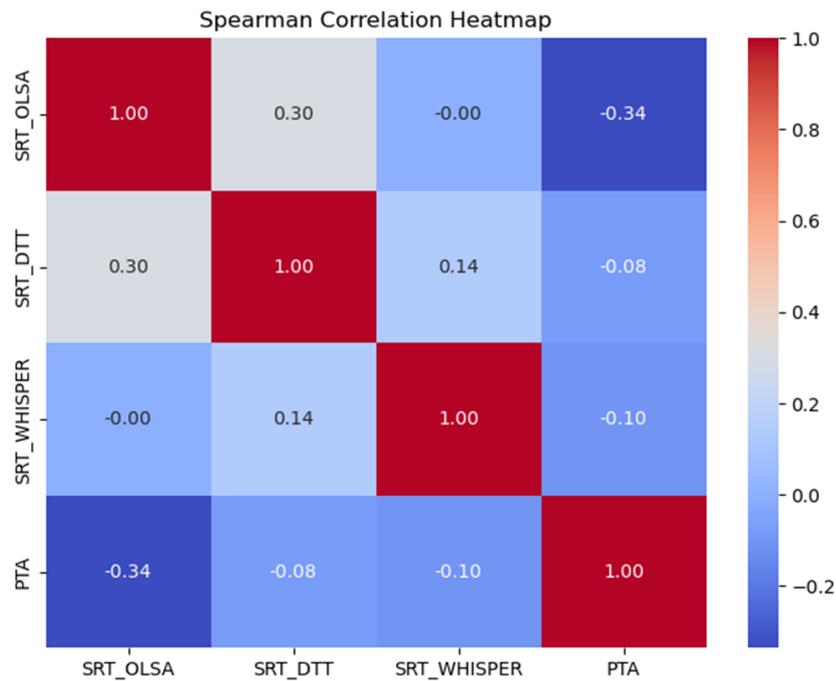


Figure 3.55: Spearman's correlation heatmap for SRT and PTA.

It was then decided to observe, for each subject and for each test, the ear that had the lower SRT value. From the table in Figure 3.56, it is observed that regarding Whisper there are no cases of equality, while these become more frequent in OLSA and DTT, which therefore show greater symmetry. The case of ID = 119 in DTT is described as "None" because the subject did not perform the test for both ears, so only the record related to one is available. For the Whisper test, the 82% of the subjects performed better for the right ear, while 18% was better with left ear. The DTT shows 41% right ear and 24% left ear, with 29% of equal performance. For the OLSA test, 53% of the subject had better values of SRT for the right ear, about 6% were equal and 41% were better with left ear. No tolerance was used to define values as equal.

Better Ear - Each Participant

ID	Better Ear Whisper	Better Ear DTT	Better Ear OLSA
118	Right	Equal	Left
119	Right	None	Right
120	Right	Equal	Left
121	Right	Left	Equal
122	Right	Right	Right
123	Right	Left	Right
124	Right	Right	Left
125	Right	Right	Right
126	Right	Equal	Left
127	Left	Right	Right
128	Right	Right	Left
129	Left	Right	Right
130	Right	Left	Left
131	Left	Right	Right
132	Right	Equal	Right
133	Right	Equal	Left
134	Right	Left	Right

Figure 3.56: Ear with lower SRT value for each participant, in all the three tests compared.

The temporal analysis of the duration of the 3 tests was carried out starting from the following variables: for Whisper, the 'total_time' variable was used, while for DTT and OLSA, the variable called 'MeasurementDuration' was employed for each participant. The graph 3.57 shows the duration in seconds of the three tests, dividing by ear (the cross indicates the left ear, the dot indicates the right ear) and by test based on the color. This is useful to observe the time deltas between the two ears for each test (always right minus left). Firstly, it is already observed from the graph how in the case of DTT, the differences between ears seem to be significantly lower compared to Whisper. Table 3.26 summarizes the average durations per test and per ear.

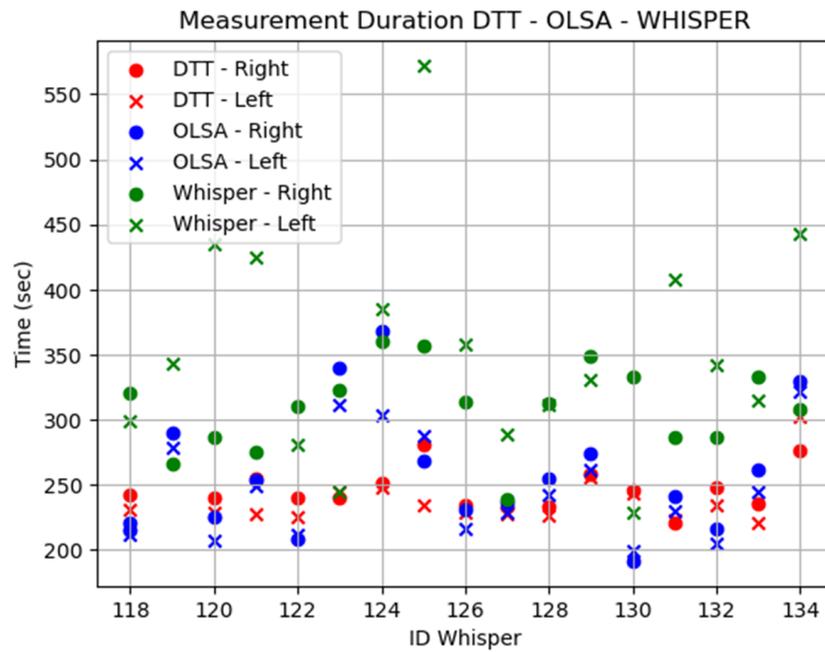


Figure 3.57: Measurement duration for DTT, OLSA and Whisper.

Table 3.26: Table of average test duration.

Mean duration (sec)	DTT	OLSA	WHISPER
Right Ear	245.93	256.99	309.40
Left Ear	237.56	247.73	353.51

Table 3.27 reports the subjects in which a significant time delta between ears was observed, and the test in which this time delta occurs. A significant delta is considered to be a difference between one ear and the other equal to or greater than 20%. It is noted that these high time deltas are observed in Whisper and OLSA, but never in DTT.

Table 3.27: Participants with significant time delta (20% or more)

ID	Delta (sec)	Test
118	-57.454	OLSA
118	-76.686	WHISPER
119	-148.641	WHISPER
120	-103.548	OLSA
121	-215.135	WHISPER
122	-50.527	WHISPER
124	-38.967	OLSA
124	-121.316	WHISPER
125	-60.002	OLSA
126	-134.834	WHISPER

In Figure 3.58 distinctly observed in red are the PTA deltas and in blue the SRT deltas for each subject, separately for each test. What can be observed is, first of all, that the PTA deltas (computed as the mean of the central frequencies) are greater and more variable compared to SRT deltas for OLSA and DTT, whereas in Whisper a similar phenomenon is not observed, as the SRT deltas are also quite variable. The subjects who presented a significant time delta do not always show a significant delta in both SRT and PTA. Specifically, in Whisper we have subject 122 who presents a significant SRT delta and also a PTA delta (although in the opposite direction), subject 124 who shows only a high SRT delta while the PTA delta is null, and subject 126 who has a delta for both between 3 and -3, but none of the other considered subject show significant values of deltas. In DTT and OLSA, where the SRT deltas are already lower, only in OLSA subjects 118 and 129 have a delta that exceeds -2dB. The time factor does not appear to be related to either PTA or SRT. Based on these observations, the greater asymmetry in terms of test duration for one ear and the other does not seem to be linked to a particular asymmetry in terms of PTA and SRT, hence to hearing-related factors.

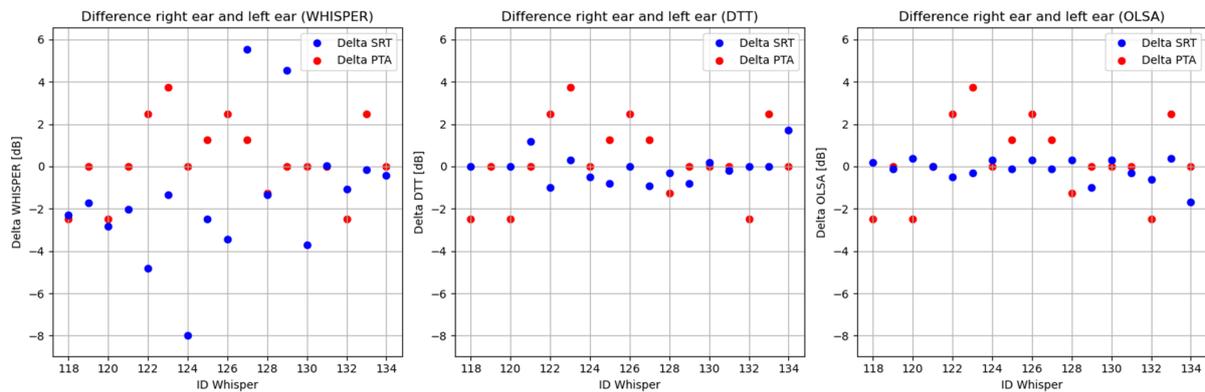


Figure 3.58: Delta of PTA and SRT for each test (right - left).

In the following analysis (Figures 3.59, 3.60, 3.64, 3.63, 3.61, 3.62), the data is categorized based on the quartiles of the SRT values from the entire dataset. The decision to calculate the quartiles and divide the data into bands was made to allow for a more detailed examination of the subjects' performances and helps to identify any groups or patterns of interest. The performers are classified into three groups: 'Best', 'Mid', and 'Worse'. The 'Best' performers are those with SRT values in the lower quartile, the 'Mid' performers fall between the lower and upper quartiles, and the 'Worse' performers are those with SRT values in the upper quartile. Separated scatter plots have been created for the right ear and the left ear, highlighting the SRT values for each subject and test. The points on the scatter plots are color-coded according to their performance group: green for 'Best', orange for 'Mid', and red for 'Worse'. Additionally, horizontal lines are drawn to indicate the separation between these performance groups. The y-axis is uniformly set to a range from -23 to -4 dB for consistency across both plots. It has been performed for Whisper, OLSA and DTT and it is shown here separately for the right and left ear.

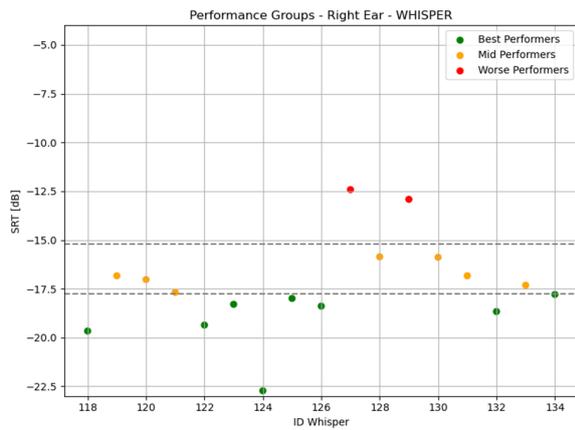


Figure 3.59: Performance Group based on quartile distinction - Whisper Right Ear.

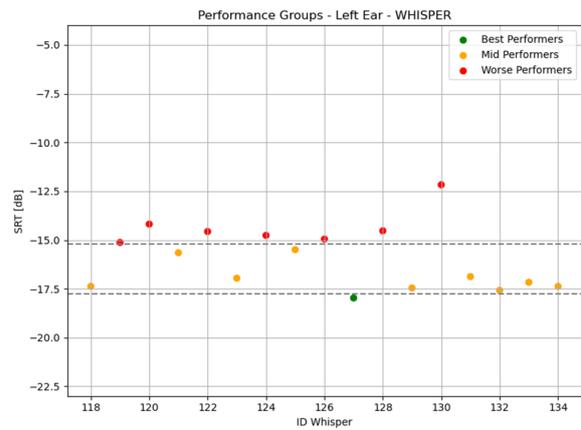


Figure 3.60: Performance Group based on quartile distinction - Whisper Left Ear.

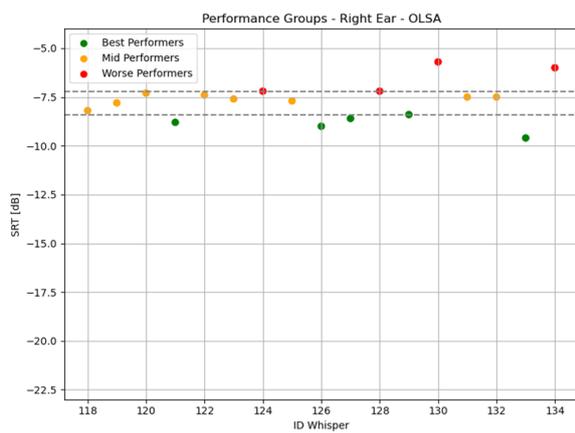


Figure 3.61: Performance Group based on quartile distinction - OLSA Right Ear.

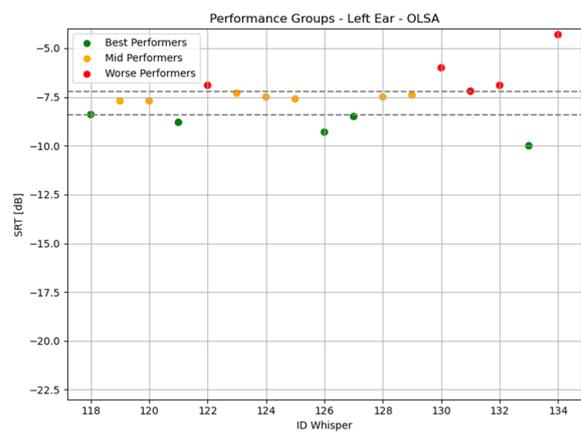


Figure 3.62: Performance Group based on quartile distinction - OLSA Left Ear

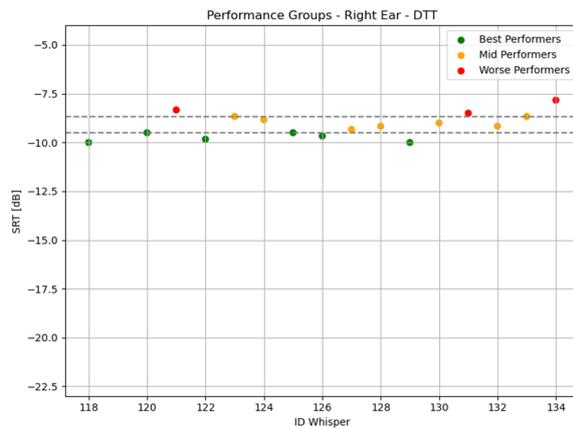


Figure 3.63: Performance Group based on quartile distinction - DTT Right Ear.

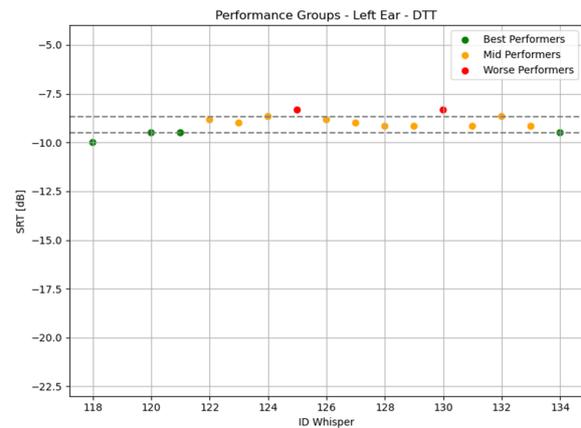


Figure 3.64: Performance Group based on quartile distinction - DTT Left Ear.

In the case of Whisper, it is quite clear that subjects classified as the best performers for the right ear are placed in lower categories for the left ear. Only the subject with ID = 127 appears in green for the left ear, while for the right ear they are in the red category. This further highlights how the left ear generally performed worse. In OLSA, however, the number of subjects in green is the same (although slightly differently distributed) for both ears. In DTT, a significant decline is not observed as in Whisper, although the performances for the right ear are slightly better.

Thanks to the data acquired during the stay in Oldenburg, it was possible to lay the foundation for this comparative analysis between the various Speech in Noise tests, obtaining the results reported here and which will be further discussed in the next chapter. Aware of the limitations especially due to the small sample analyzed, it has been tried to draw conclusions that are consciously temporary and for which further analysis on a larger sample is needed.

4 | Discussions

The aim of this thesis was to employ clustering techniques to identify auditory profiles from a dataset of audiological features derived from laboratory testing, screening initiatives and measurements performed at Politecnico di Milano and at the university of Oldenburg and classify them using machine learning methods. The primary goal was to gain a deeper understanding of the data, beyond mere classification based on the gold standard (PTA) and facilitate the identification of potential hearing impairments and cognitive challenges within population screening contexts. Another objective of the thesis was to collaborate with the Oldenburg team to create a novel dataset comprising subjects who underwent testing using two platforms, the WHISPER Platform (comprising Whisper test, DST, risk factor questionnaire) and the Virtual Hearing Clinic Platform (comprising OLSA and DTT). Furthermore, a comparative analysis of three speech-in-noise tests (Whisper, OLSA, and DTT) was conducted to understand the main differences and how they will be relevant in different applications, enrich the test batteries and eventually as a basis for cross-validating profiling and classification methodologies, leveraging the distinct attributes of each dataset.

In this section, the results presented in the previous chapter will be discussed. The clustering and classification analysis involved segmenting the data into distinct clusters based on various audiological and demographic features, followed by a detailed examination of the characteristics of these clusters. Additionally, classification techniques were employed to predict outcomes based on the identified patterns. The aim of this discussion is to interpret the findings, explore the significance of the identified clusters, and evaluate the performance and implications of the classification models used. Through this, we will gain a deeper understanding of the underlying structure of the audiological data and the potential clinical relevance of these findings. Additionally, the main results from the comparative analysis between the three SIN tests on 17 subjects will be discussed.

4.1. Dataset characterization

Firstly, as observed in Figure 3.1, despite the careful selection of subjects aged between 30 and 50 years, and the contribution of this thesis to try to reduce the lack of subjects in the dataset, there remains a gap that will need to be filled in the future. In general, it is observed that increasing age corresponds in most cases to subjects with higher PTA values (above 35 dB), supporting the thesis of increased hearing loss with aging [42]. The choice of the PTA thresholds for hearing loss distinction is supported by extensive scientific literature highlighting the relationship between PTA values and hearing function, and that is the reason why it has been adopted here.[2]

An average PTA value of around 13.53 dB is observed (Figure 3.2), which falls, as per literature [2] (Figure 1.3), within the range of normal hearing. Therefore, the majority of the sample is normal-hearing, although there are also subjects with moderate and severe hearing loss (maximum PTA of 80dB). The mean SRT values are quite low (-11.75dB, Figure 3.4), confirming the presence of a high amount of normal-hearing subjects.

It can be observed from the SRT histogram (Figure 3.4) that higher SRT values correspond to higher PTA values (correlation value = 0.679, comparable to what have been obtained in literature for different SIN tests, as in [18], where the correlations were around 0.60). This trend suggests a relationship between hearing loss and the ability to recognize spoken language in the presence of background noise (SRT). In other words, when hearing loss (measured through PTA) is more severe, it is expected that the ability to understand spoken language in noisy conditions is also compromised.

This relationship makes sense because hearing loss can affect the ear's ability to gather and transmit sound signals to the brain clearly and accurately. Consequently, compromised hearing due to hearing loss can make it more difficult for an individual to perceive and interpret spoken language correctly, especially in noisy or competitive environments.

In practice, this means that individuals with greater hearing loss (higher PTA values) may find it more challenging to understand spoken language in the presence of background noise (higher SRT values) compared to those with less severe hearing loss.

For subjects over 50 years old, Figure 3.9 shows that there is a slight trend towards increased average typing time during the execution of DST, which could be due to a decline in cognitive and motor abilities with age. While most subjects maintain relatively quick typing times, advanced age tends to be associated with an increase in average typing time, with some extreme cases indicating much higher times than the average. There are, however, adult subjects who perform well in these terms, indicating that this is not always

the case. However, it is noted that among the subjects who experience a slowdown, these correspond to older ages, making it more likely (as has been studied in literature, e.g.,[43]) that cognitive decline, when developed, goes hand in hand with aging.

4.2. Clustering of subjects' profiles

Regarding clustering into different subjects' profiles according to audiological, cognitive-related and risk factors features, the use of the elbow method and the indices utilized to define the optimal number of clusters led to the decision to analyze three cases: 3, 4, and 5 clusters. K-Prototypes has been applied because it is a valuable algorithm for clustering datasets with mixed data types (i.e., both numerical and categorical), offering the ability to uncover meaningful patterns and insights [34]. However, it is important to be aware of the limitations of this method, such as initialization sensitivity and computational complexity.

The dual analysis, conducted at first considering Whisper features only and then adding features related to risk factors and DST, showed interesting results from different perspectives. On the dataset of 590 records considering only Whisper features (Section 3.3.2), both 3, 4, and 5 clusters yield interpretable and distinguishable profiles. Indeed, when considering only the features of the Whisper speech-in-noise test, clustering into 3 clusters seems to be predominantly driven by PTA, number of trials, and total test time, distinguishing into 3 profiles: the first characterized by younger subjects (42 years) with lower PTA values compared to the other two clusters, with fewer trials and shorter test time (profile corresponding to cluster 2 in Table 3.2). The second profile represents a more adult population (49 years), with higher PTA (15dB), higher number of trials, and longer test time (cluster 1). Even more adults (76 years) with higher values of all these variables are represented by cluster 3.

It is interesting to observe that, considering only the records with better PTA, the distinction in age and PTA between the last two profiles is more pronounced also for centroids and not only medoids (as seen in Table 3.3). In medoids' coordinates, cluster 1 represents a population of about 76 years, with PTA values above 20dB, i.e., beyond the normal hearing threshold). Therefore, using only data of ears with better PTA proved useful to represent the older age group with hearing performance outside normal hearing. The other two clusters remain in the normal hearing range but represent younger age groups (49 and 42 years) with test performance decreasing with increasing age.

Increasing the number of clusters to 4 (Table 3.6), the features that continue to drive

the profiling are the same. However, this clustering configuration further distinguishes subjects who, although adults (60 years, cluster 3) with slightly higher PTA values than the other clusters (although still within normal limits), have a high percentage of correct answers, on more trials and longer test times. Considering only the better ear in terms of PTA (Table 3.8) makes it possible to distinguish subjects with PTA around 20dB and those with lower values, as expected because the records with higher values of PTA have been excluded. As expected, the clusters with higher PTA correspond to older subjects, where a decline in hearing is more common.

The addition of a fifth cluster allows for a broader distinction regarding age. In the case of 5 clusters, profiles representative of subjects around 28 years old are obtained, who have excellent PTA, SRT, high %correct, and low test times. Cluster 4 (Table 3.10) also highlights the population around 73 years old who, despite high hearing threshold values, have a good number of trials, a high percentage of correct responses (comparable to the ones of the group of 28 years old subjects), and rather good test times, which is interesting.

In the case of 5 clusters for only the records with better PTA, the distinction is less clear in terms of age compared to the case where all values are taken, because clusters 2-4 and 1-5 report subjects of about the same age, but they are distinguished by features related to the number of trials and total time (cluster 2 performs better than 4, and cluster 1 performs better than 5), as the others are very similar to each other.

Therefore, we observe that the most distinctive variables for the clusters are surely the number of trials, the total test time, the PTA, and the age (this last one especially when increasing the number of clusters), followed by the %correct which in some situations allows for greater distinction. Overall, the clusters observed show that, with increasing age, subjects tend to have lower hearing performance compared to younger subjects, although there are cases where the profiles manage to distinguish older subjects with good test performance. The use of 5 clusters seems very useful in distinguishing by age, thus creating well-defined profiles. In general, however, the clusters and their interpretations are consistent, despite the awareness of the need for more data acquisition to strengthen the analysis.

It is interesting to compare the medoid table for all records with the one for only the ears with better PTA records and note that the vast majority of coordinates remain the same, whereas this consistency does not hold for the centroids. A possible interpretation of this observation lies in the inherent differences between centroids and medoids. Medoids are actual data points and represent the central point in each cluster, making them less

sensitive to variations. Thus, when only the ears with better PTA records are considered, the medoids often remain the same if the distribution is similar to the overall dataset. This suggests the core characteristics of the clusters don't change significantly when excluding higher PTA records. Centroids, being the average of all points in a cluster, are more sensitive to changes. Excluding records with worse PTA values alters the cluster's average characteristics, resulting in different centroid values. This indicates that the overall distribution and central tendencies of the clusters shift when only the ears with better PTA records are used. The consistency of the medoid coordinates suggests that the fundamental structure and core characteristics of the clusters remain largely unchanged when considering only the ears with better PTA records. In contrast, the variation in centroids highlights how the average properties of the clusters are influenced by the inclusion or exclusion of records with different PTA values.

The addition of features related to the cognitive test and risk factors (Section 3.3.4), with analysis performed in the same way as in the case of only Whisper features, led to interesting results, confirming that the features of the speech-in-noise test are the ones driving the profiling. Certainly, the reduction in the number of data (from 590 to 266) resulted in a loss of robustness in the analysis. For this reason, one of the future developments is to expand the dataset to achieve a higher number of subjects who have also completed the DST and the risk factor questionnaire.

From the clustering analysis with 3 clusters (Table 3.16), subjects with lower PTA values are observed compared to the case of only Whisper, thus profiles with higher hearing performance. In this case, Cluster 1 comprises subjects of 43 years, with close values of #trials and %correct compared to Cluster 3 (respectively 70 vs 76, and 91.4% vs 89.5%) that contains subjects of 48 years. Their main difference is in the gender (Cluster 1 has a majority of male, Cluster 3 of female), and PTA (2.5 dB vs 5.0 dB), but especially in two features regarding the cognitive test and one regarding risk factor. In fact, Cluster 1 contains subjects that are exposed to high volumes, they present lower average typing time (0.94 sec) and lower DSS (6.0), while Cluster 3 presents people with no volume exposure, higher average typing time (2.21 sec) and higher DSS (8.0).

The second cluster shows slightly higher average typing times than the first one, confirming that even from the cognitive test perspective, the second cluster represents subjects who, although quite young (23 years), show lower cognitive and test performance (higher values of #trials, lower %correct, higher total time and DSS).

The addition of a fourth cluster (Table 3.18), shows a slightly greater distinction in terms of SRT, but always with trials and test time as dominant variables. The DSS seems to

contain also relevant information to distinguish the clusters. Clusters 2 and 3, although similar in age and %correct, have different typing times (the second is faster than the third) as well as a number of trials and test time. The important difference between the second and third cluster is also the opposite gender type (female for Cluster 2 and male for Cluster 3), the DSS (higher for the third) and total time (still higher in the third cluster).

The fifth cluster once again (Table 3.20), allows distinguishing younger (23 years) and older age groups (78) from intermediate age groups. PTA is also more discriminating in this case, as well as the average typing time, in addition to the usual test time. Therefore, profiles with adult subjects (78 years) are identified who have significantly lower cognitive test performance than average (DSS is 4.0), although in terms of test time, trials, and %correct they perform quite well. PTA values, however, are above 20dB. The younger subjects (23 years) show good hearing and cognitive performances, with low PTA, relatively low test time and good %correct. Interestingly, Cluster 1 (subjects of 49 years old) is the one that shows higher %correct, lower trials and test time, lower typing time and also good hearing values, showing how even mid-aged people can perform really good.

Cognitive and risk factor variables can be useful to provide further specifics and understanding of the generated profiles, confirming the relationship between hearing loss and cognitive decline [19], although they do not appear to be the variables driving the profiling.

The currently available profiles naturally only offer a snapshot of the included measures. It is reasonable to assume that incorporating additional measures will enhance the precision of subjects characterization. The profiles seem to be audiological plausible, and each one can be differentiated from the others by at least one audiological characteristic, thereby allowing them to be regarded as distinct groups in terms of audiological measures.

4.3. Classification of clustered subjects

Regarding classification, starting from the analysis performed only on Whisper features (Section 3.3.3), slightly better performance metrics are observed compared to the case with additional features (Section 3.3.5), especially in the test set, thus showing that adding features certainly introduces greater complexity and variability. Despite the slight decrease in performance, they are still very good, so the addition of features does not cause a deterioration significant enough to rule out the option of performing profiles

classification. However, the reduction in the number of records is certainly a problematic issue for classification, as it results in a loss of model generalization. Specifically, there have been cases where the reduction in the number of records has led to misclassification of test data or even the inability to classify because there were no data in the test set for a certain cluster.

In the case of three clusters, no classification issues were encountered with the complete dataset with only Whisper features (in any of the three splitting methods), whereas there were some in the one with fewer records (adding DST and risk factors). In the case of 3 clusters with the only Whisper features dataset, both Random Forest and SVM had very satisfactory performance, with test accuracy always slightly lower than that of the training (thus not showing overfitting) and always above 0.93. The same can be said in the case of 3 clusters with the reduced dataset, where performance was unfortunately influenced by the small size of the test set, especially in the splitting method where training = without Oldenburg data, test = Oldenburg data, confirming the need to revisit this analysis once the dataset is expanded. The method that performed the worst was always KNN.

In the case of 4 clusters and only Whisper dataset, SVM achieves better test performance than Random Forest for each of the splittings, although Random Forest still performs well. KNN always has the lowest performance. Adding DST and risk factor features, thus reducing the number of data, Random Forest and SVM are always comparable, while KNN performance drops further. The third type of splitting (without Oldenburg data - with Oldenburg data) shows always the most difficulties in classification, due to low number of records in the test set.

Increasing to 5 clusters, however, SVM keeps being slightly more robust than Random Forest even in the reduced dataset (the one with DST and risk factor features), while remaining fairly similar in the only Whisper dataset.

Overall, Random Forest and SVM are certainly valid and robust methods for classification in all the cases shown, while KNN consistently stands out for its lower performance.

On the dataset including also DST and risk factor features, having 4-5 clusters results in at least one cluster with very few elements, and in some cases, after splitting for classification into training and test sets, the least numerous clusters do not appear in the test set. Therefore, it is more reasonable to distinguish only 3 clusters until more data will be added, in view of the classification that allows for better results.

The fewer the clusters, the more they represent macro-groups. As the number of clusters increases, the points within each cluster become more similar and fewer, which can reduce

the classifier's accuracy. As it can be observed in different cases (especially with higher number of clusters) the training of a ML method for the classification of clusters with small number of data does not lead to reliable results and its generalizability is not assured, but the results were included for the sake of completeness. Further information is needed in order to provide a large enough sample size for classification purposes.

Due to the awareness of the limitation caused by the few data available, techniques were attempted to minimize this problem: firstly, an attempt was made to set a minimum number of elements for each cluster during clustering, so that the algorithm would converge only when each cluster had at least that number of points, but unfortunately, the results were not satisfactory as convergence was not reached within a reasonable time range for the application. For the 80-20% split cases, stratification was chosen, i.e., trying to maintain the proportion of each original cluster in both the training and test sets for a more robust analysis. This was not able to solve the problem as there were already very small clusters in the clustering itself, and thus the proportion in the test set became roundable to 0. In the future, it will certainly be of particular interest to address this issue with different techniques, and in parallel, to increase the dataset to obtain a more robust analysis.

The reasons why Random Forest and SVM are performing better than KNN in this scenario could be different, for example their robustness against noise, ability to handle non-linear and complex relationships, and effectiveness in high-dimensional features spaces. These attributes make Random Forest and SVM powerful tools for achieving better performance in complex classification problems such as those presented by this dataset. Random Forest is particularly effective in managing noisy or complex data because each tree is constructed using a random subset of the dataset. This method helps balance the biases and variances of individual trees, making the overall model more robust. On the other hand, SVM tends to generalize well on test data and reduce overfitting. Additionally, SVM can utilize non-linear kernels, such as the Gaussian or polynomial kernel, to transform the feature space and separate classes that are not linearly separable in their original space. Given that SVM is designed to maximize the margin between classes, it helps in generalizing well even with smaller datasets. By focusing on the decision boundary and the most critical points (support vectors), SVM can often produce more accurate classifications with less data. The use of kernel functions allows it to handle non-linear relationships effectively without requiring a large number of data points to learn complex patterns. This adaptability is particularly beneficial in cases with small number of data. In contrast, KNN can be less performant because it suffers from the curse of dimensionality, and has high computational complexity. KNN's performance

heavily depends on the availability of sufficient data to identify accurate neighbors. With fewer data points, the distances calculated may not be representative of the true class boundaries, leading to less reliable classifications. These limitations make it less suitable for this audiological dataset, where Random Forest and SVM can provide more accurate and reliable classification results.

Furthermore, an interesting point for future analysis could be to investigate whether the difference in calibration for the Whisper test between the dataset developed in Italy and the one acquired in Oldenburg might be a relevant issue that influences classification performance. It would be worthwhile, in future research, to explore this aspect before performing clustering and classification. Indeed, the calibration of devices is an important factor to consider when deciding to combine test batteries.

As already reported, a limitation of the current classification is the small number of patients in each profile. To achieve more robust validation, larger and more balanced datasets that also include more severely affected patients are needed. This is likely to enhance predictive accuracy. Increasing the size of the training set will improve the classifier's training, while a larger test set will boost the reliability of predictions. Currently, test performance might be overestimated for some profiles due to the small test set size. However, further reducing the training set size is not advisable as it would increase the bias in the classification models. Therefore, further evaluation using a larger number of patients is required.

The aim was to generate several plausible profiles based on the dataset in order to capture differences among subjects besides their grouping in terms of PTA value. Incorporating specific objectives or practical guidelines for clustering tailored toward a certain outcome such as diagnostics, hearing aid fitting, or some other application field could enable a slightly different analysis more tailored to the defined purpose.

Some application of the current analysis could be to summarize patient information for a clinical decision-support system, aiding healthcare professionals in making informed decisions. Another potential use is in mobile assessment tools that allow for real-time evaluations of subjects' auditory and cognitive functions, improving accessibility and efficiency in audiological and cognitive health monitoring. Additionally, this analysis can be utilized in personalized hearing aid programming, ensuring devices are tailored to individual auditory profiles [44]. It can also support auditory rehabilitation programs by identifying specific patient needs and tracking progress over time. Moreover, the insights gained from this analysis can inform public health initiatives aimed at early detection and intervention for hearing and cognitive impairments.

4.4. Comparative analysis between SIN tests.

Regarding the comparative analysis between the speech-in-noise tests in normal hearing subjects, it is immediately evident from observing Figure 3.42 how the distribution of points for SRT values in Whisper is more scattered and significantly lower compared to DTT and OLSA; this is observed in all analyzed subjects. The results in terms of average SRT for OLSA are comparable to the ones in [27], as the mean value of SRT for OLSA is -7.66 dB against -6.8 dB reported in the cited study (on German listeners with closed-form and adaptive measuring). DTT shows a mean value of -9.09 dB in this analysis, even more similar to -9.3 dB (15 German normal-hearing listeners tested), the value reported in [29]. Whisper mean value for SRT is instead -16.64 dB, comparable to $-15.3dB \pm 1.87$, obtained from 26 normal hearing Italian young adults in [45].

The reason why the SRT values in Whisper are significantly lower compared to the other two tests is due to the fact that this test is a multiple-choice test with 3 alternatives, and also because the vocal material used (i.e., VCVs) is simpler, thus shifting the psychometric intelligibility curve to lower values. Additionally, the different type of noise used in the three SIN tests could influence the various performances.

The variability, on the other hand, is linked to the type of task required by the test, resulting in greater uncertainty in the SRT estimation for Whisper. This is because the psychometric curves of the VCV are less steep compared to those of the digit triplets of the DTT, thus generating greater uncertainty in the estimation.

When observing the two ears separately, it is evident that the right ear reaches lower SRT values, especially in Whisper, while this is less evident in OLSA and DTT, thus showing greater symmetry. Interestingly, the consonants "M", "F", and "R" are the ones that are most difficult for all subjects to distinguish during the Whisper test; this could be due to a linguistic factor, but also to a simple phonetic factor, related to the sound emitted when pronouncing "AFA," "AMA," "ARA," which could be more similar to potential noises and therefore mistaken for such. The error rate for the DTT and OLSA tests is quite different, with OLSA reaching an average of 47% compared to DTT, which is around 21%. This disparity is likely due to OLSA's greater variability and complexity, as it involves sentences with different types of words, whereas DTT consists of a sequence of numbers that is typically easier to remember and recognize. The asymmetry between ears is always more evident in Whisper rather than in the other tests, also because Whisper's SRT values are already more distributed compared to those of DTT and OLSA. It was then decided to observe the subjects who performed Whisper as the last test to evaluate if the phenomenon of fatigue had any effect in terms of performance, but this

phenomenon was not considered observable. Nor was a correlation between the various tests particularly noted, as expected. From Figure 3.57 and Table 3.26 it is possible to observe the measurement duration for the three tests. The mean duration of the test for OLSA and DTT is higher for the right ear than for left ear, while for Whisper is the opposite. The disparity in ears for OLSA and DTT is not so significant as for Whisper and the other two tests (p-value of the Levene's test is 0.00542 for Whisper and OLSA, 0.0110 for Whisper and DTT, and 0.5582 for DTT and OLSA, suggesting no significant variability between DTT and OLSA), showing not only more symmetry in SRT but also in the duration of the tests. In fact, left ear for Whisper test is more than 45 sec longer in average than for right ear, that is usually the first ear tested. This could be due to some kind of tiring of the subject after having performed the test already for the first ear, or concentration, or again familiarity with the specific test. Another factor that could contribute to the apparent asymmetry of Whisper in time is the variable number of trials compared to the other two tests. In fact, while OLSA and DTT have always the same number of stimuli, Whisper has an adaptive procedure meaning that the number of stimuli (and so the total time of the test) can vary and therefore it could lead to a less symmetry. This difference between the tests is noteworthy. It has been decided to look more deeply into the participants that had a notable delta of time (defined as higher than 20%) and it has been noted that these participants show this delta especially in Whisper and never in DTT. As observed in Figure 3.58, the time factor does not appear to have a relationship with the delta of PTA or SRT, as the same behavior for delta PTA and SRT is not observed in subjects with a high delta time of test. It can therefore be deduced that an asymmetry in terms of auditory values is not necessarily linked to an asymmetry in terms of test time. Combining the information that, on average for Whisper, the better ear in terms of SRT (thus with lower values) is the right ear which also takes less time on average, it might indicate that the first ear tested is generally the one for which individuals tend to be more focused, resulting in shorter test times and better auditory performance. However, this relationship between SRT and time is observed within each individual test. In the comparison across the three tests, the scenario is slightly different because we see that subjects perform differently and the asymmetry observed for Whisper is less pronounced in the other two tests.

Subsequently, the data were divided into three groups based on quartiles (0.25, 0.75), identified as Best, Mid, and Worse performers, to observe how many subjects fell into a class (based on their SRT value) for one test and then changed class in the others. This analysis differs from the one based on deltas, as here the thresholds for moving from one class to another is defined by quartiles in SRT.

For OLSA: The SRT performances between the right and left ears are notably symmetrical, as there are no subjects who transition from Best to Worse between ears; transitions are only observed from Best to Mid and vice versa, or from Mid to Worse and vice versa.

For DTT: Subject 121 transitions from Worse to Best between ears, as does 125 and 134 (no observations for 119 as they did not complete the test for the second ear).

For Whisper: Subjects 122, 124, 126, and 127 transition from Best to Worse or vice versa, more frequently compared to the other two tests.

Overall, it is observed that the class into which the same subject falls can vary significantly among the three tests, rather than consistently remaining the same. This highlights how test performances differ across the three tests, owing to their inherent nature. Again, it should be noted that the comparison involved normal hearing subjects, but more evident results are expected when screening also hearing impaired subjects, hence expanding both PTA and SRT range.

In conclusion, the comparative analysis has shown the different nature of the three tests. From the perspective of variability in terms of auditory performance, it highlights the greater variability of SRT in the Whisper test compared to OLSA and DTT. The percentage of error, on the other hand, is more variable for OLSA and DTT (with OLSA reaching higher average values). The timing also varies significantly due to the structure of the tests, with Whisper consistently showing more significant time deltas.

Despite the fact that all three tests are speech-in-noise tests, their different nature (such as having an adaptive procedure in terms of the number of stimuli) also leads to differences in the results obtained. Overall, however, the values obtained for all three tests are considered consistent in the sense that they are associated with normal-hearing subjects. The choice of a test should be made based on the field of application and the subjects being tested. Nonetheless, combining the methods into a single test battery has been enriching and provides additional information about the tested participants, allowing for further analysis.

Certainly, the limited number of subjects is an important limitation, so it would be necessary to extend the dataset composed by subjects who underwent both Whisper, DTT and OLSA, and repeat the analyses to verify greater consistency and to make further observations.

5 | Conclusions

The analyses conducted in this study provide valuable insights into the impact of various features on hearing and cognitive performance, as well as the efficacy of different clustering and classification methods. The relationship between age and PTA values confirmed the expected trend of age-related hearing loss, with the majority of subjects exhibiting normal hearing. The clustering results, whether using solely Whisper features or incorporating additional cognitive and risk factor data, consistently highlight the influence of age, PTA, number of trials, and total test time as key variables to identify homogeneous groups of subjects. While the inclusion of additional features introduces complexity and slightly reduces performance metrics, the overall robustness of the models remains high, with Random Forest and especially SVM consistently outperforming KNN. The primary limitation identified is the reduced dataset size, which affects the robustness of both clustering and classification outcomes. Future work should focus on expanding the dataset to enhance the analysis and exploring alternative techniques to address the challenges of small clusters and data stratification. Despite these limitations, the findings affirm the relationship between aging, hearing loss, and cognitive performance, and underscore the importance of comprehensive data collection and methodological rigor in advancing our understanding of these interrelated domains.

Comparative analysis of the Whisper, OLSA, and DTT speech-in-noise tests showed that Whisper had significantly lower and more variable SRT values due to simpler vocal material and different noise types. OLSA had a higher error rate compared to DTT, attributed to the complexity of sentence-based stimuli. The analysis also revealed consistent asymmetry in SRT performance, with the right ear generally performing better. In this case as well, it will be necessary to expand the dataset and reformulate the analysis to achieve more robust and reliable results. However, this thesis has laid the groundwork for what could be a more in-depth future analysis, exploring the use of a combined test battery to enrich the amount of obtainable information.

Bibliography

- [1] A. Sheikh, B. e Zainab, K. Shabbir, and A. Imtiaz, “Structure and physiology of human ear involved in hearing,” in *Auditory System* (S. Naz, ed.), ch. 1, Rijeka: IntechOpen, 2022.
- [2] World Health Organization, “Deafness and hearing loss.” <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>, 2024. Accessed: 2024-06-16.
- [3] G. Rocco, G. Bernardi, R. Ali, T. van Waterschoot, E. M. Polo, R. Barbieri, and A. Paglialonga, “Characterization of the intelligibility of vowel–consonant–vowel (vcv) recordings in five languages for application in speech-in-noise screening in multilingual settings,” *Applied Sciences*, vol. 13, no. 9, p. 5344, 2023. Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, 20133 Milan, Italy; Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, 3001 Leuven, Belgium; Dipartimento di Ingegneria Informatica Automatica e Gestionale Antonio Ruberti (DIAG), Università La Sapienza di Roma, 00185 Rome, Italy; Cnr-Istituto di Elettronica e di Ingegneria dell’Informazione e delle Telecomunicazioni (CNR-IEIIT), 20133 Milan, Italy.
- [4] G. A. Gates, J. L. Cobb, R. B. D’Agostino, and P. A. Wolf, “The relation of hearing in the elderly to the presence of cardiovascular disease and cardiovascular risk factors,” *Arch Otolaryngol Head Neck Surg.*, vol. 119, pp. 156–161, February 1993.
- [5] R. Davies, “Audiometry and other hearing tests,” in *Handbook of Clinical Neurology*, vol. 137, ch. 11, pp. 157–176, Elsevier, 2016.
- [6] M. Jennifer Junnila Walker, MD, A. Leanne M. Cleveland, A. Jenny L. Davis, and A. Jennifer S. Seales, “Audiometry screening and interpretation,” *American Family Physician*, vol. 87, pp. 41–47, 2013. Available at: <https://www.aafp.org/afp/2013/0101/p41.html>.
- [7] R. C. Bredfeldt, “An introduction to tympanometry,” *Am Fam Physician*, vol. 44, pp. 2113–2118, Dec 1991.

- [8] Y. Motoyama, “Auditory brainstem response: ABR,” *No Shinkei Geka*, vol. 51, pp. 425–429, May 2023.
- [9] W.-L. Wang, Y.-R. Bai, Q. Zheng, S. Zheng, X.-Y. Liu, and G.-J. Ni, “Otoacoustic emission and its application in anesthesia,” *Eur Rev Med Pharmacol Sci*, vol. 26, no. 15, pp. 5426–5435, 2022.
- [10] V. A. Sanchez, M. L. Arnold, D. R. Moore, O. Clavier, and H. B. Abrams, “Speech-in-noise testing: Innovative applications for pediatric patients, underrepresented populations, fitness for duty, clinical trials, and remote services,” *Journal of the Acoustical Society of America*, vol. 152, pp. 2336–2356, 2022.
- [11] C. J. Billings, T. M. Olsen, L. Charney, B. M. Madsen, and C. E. Holmes, “Speech-in-noise testing: An introduction for audiologists,” *Semin Hear*, vol. 45, no. 1, pp. 55–82, 2024.
- [12] R. A. Roy, “Auditory working memory: A comparison study in adults with normal hearing and mild to moderate hearing loss,” *Glob J Oto*, vol. 13, p. 555862, February 2018.
- [13] F. Ostrosky-Solís and A. Lozano, “Digit span: Effect of education and culture,” *International Journal of Psychology*, vol. 41, p. 333–341, Oct. 2006.
- [14] M. K. Mielicki, R. H. Koppel, G. Valencia, and J. Wiley, “Measuring working memory capacity with the letter–number sequencing task: Advantages of visual administration,” *Applied Cognitive Psychology*, vol. 32, no. 6, pp. 805–814, 2018.
- [15] L. H. Sweet, *N-Back Paradigm*, pp. 1718–1719. New York, NY: Springer New York, 2011.
- [16] R. A. A. Teixeira, E. C. Zachi, D. T. Roque, A. Taub, and D. F. Ventura, “Memory span measured by the spatial span tests of the cambridge neuropsychological test automated battery in a group of brazilian children and adolescents,” *Dementia & Neuropsychologia*, vol. 5, no. 2, pp. 129–134, 2011.
- [17] P. Souza and K. Arehart, “Robust relationship between reading span and speech recognition in noise,” *International Journal of Audiology*, vol. 54, no. 10, pp. 705–713, 2015.
- [18] T. S. Redick and D. R. B. Lindsey, “Complex span and n-back measures of working memory: A meta-analysis,” *Psychonomic Bulletin & Review*, vol. 20, no. 6, pp. 1102–1113, 2013.

- [19] A. Paglialonga, E. M. Polo, M. Lenatti, M. Mollura, and R. Barbieri, “A screening platform for hearing loss and cognitive decline: Whisper (widespread hearing impairment screening and prevention of risk),” *Stud Health Technol Inform*, vol. 309, pp. 170–174, Oct 20 2023.
- [20] University of Oldenburg, “Virtual hearing centre: Modules and perspectives.” <https://uol.de/vhc/modules-and-perspectives>, 2024. Accessed: 2024-06-24.
- [21] M. R. Leek, “Adaptive procedures in psychophysical research,” *Perception & Psychophysics*, vol. 63, pp. 1279–1292, Nov. 2001.
- [22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2125–2136, Sept. 2011.
- [23] P. Lyregaard, “Towards a theory of speech audiometry tests,” in *Speech Audiometry* (M. Martin, ed.), pp. 34–62, Chichester (UK): John Wiley & Sons, 2nd ed., 1997.
- [24] J. B. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proc. 5th Berkeley Symp Mathematical Statistics and Probability*, pp. 281–297, 1967.
- [25] M. A. García-Pérez, “Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties,” *Vision Res*, vol. 38, pp. 1861–1881, Jun 1998.
- [26] M. Zanet, E. M. Polo, G. Rocco, A. Paglialonga, and R. Barbieri, “Development and preliminary evaluation of a novel adaptive staircase procedure for automated speech-in-noise testing,” *EMBC*, 2019.
- [27] B. Kollmeier, A. Warzybok, S. Hochmuth, M. A. Zokoll, V. Uslar, T. Brand, and K. C. Wagener, “The multilingual matrix test: Principles, applications, and comparison across languages: A review,” *Int J Audiol*, vol. 54, no. Suppl 2, pp. 3–16, 2015.
- [28] T. S. K. Cas Smits and T. Houtgast, “Development and validation of an automatic speech-in-noise screening test by telephone,” *International Journal of Audiology*, vol. 43, no. 1, pp. 15–28, 2004.
- [29] M. A. Zokoll, K. C. Wagener, T. Brand, M. Buschermöhle, and B. Kollmeier, “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*, vol. 51, no. 9, pp. 697–707, 2012.

- [30] scikit-learn developers, “Principal component analysis (pca).” <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html#pca>, 2023. Accessed: 2024-06-05.
- [31] S. Lê, J. Josse, and F. Husson, “FactoMineR: A package for multivariate analysis,” *Journal of Statistical Software*, vol. 25, no. 1, pp. 1–18, 2008.
- [32] scikit-learn developers, “t-distributed stochastic neighbor embedding (t-sne).” <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>, 2023. Accessed: 2024-06-05.
- [33] A. Vidhya, “K-modes clustering algorithm for categorical data,” 2021. Accessed: 2024-06-05.
- [34] Z. Huang *Data Mining and Knowledge Discovery*, vol. 2, no. 3, p. 283–304, 1998.
- [35] Analytics Vidhya, “In-depth intuition of k-means clustering algorithm in machine learning,” 2021. Accessed: 2024-06-14.
- [36] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
- [37] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.
- [38] T. Caliński and J. Harabasz, “A dendrite method for cluster analysis,” *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.
- [39] M. Bennett, M. Nekouei, A. Prieditis, R. Mehta, and E. J. Kleczyk, “Methodology to create analysis-naive holdout records as well as train and test records for machine learning analyses in healthcare.” License CC BY 4.0, May 2022.
- [40] E. AI, “Multi-class classification metrics.” Available online, Accessed on June 21, 2024.
- [41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *scikit-learn: Machine Learning in Python - StandardScaler*, 2011. Accessed: 2024-06-25.
- [42] J. A. Deal, N. S. Reed, E. C. Pedersen, and F. R. Lin, *Causes and Consequences of Age-Related Hearing Loss*, pp. 173–198. Cham: Springer International Publishing, 2020.

- [43] M. Dexter and O. Ossmy, “The effects of typical ageing on cognitive control: recent advances and future directions,” *Frontiers in Aging Neuroscience*, vol. 15, p. 1231410, 2023.
- [44] R. Sanchez-Lopez, T. Dau, and W. M. Whitmer, “Audiometric profiles and patterns of benefit: a data-driven analysis of subjective hearing difficulties and handicaps,” *International Journal of Audiology*, vol. 61, no. 4, pp. 301–310, 2022.
- [45] E. M. Polo, M. Zanet, A. Paglialonga, and R. Barbieri, “Preliminary evaluation of a novel language independent speech-in-noise test for adult hearing screening,” in *8th European Medical and Biological Engineering Conference* (T. Jarm, A. Cvetkoska, S. Mahnič-Kalamiza, and D. Miklavcic, eds.), (Cham), pp. 976–983, Springer International Publishing, 2021.

List of Figures

1.1	The main elements of the auditory system.	2
1.2	Audiogram of a flat conductive hearing loss from [6]. The horizontal axis represents sound frequency, ranging from low to high pitch. The vertical axis represents sound intensity. Thresholds for the right ear are marked with a red circle, while thresholds for the left ear are marked with a blue X.	4
1.3	Classification of Hearing Thresholds by WHO. Grades of hearing loss and related hearing experience in quiet and noisy environments.	5
2.1	Whisper first page: age, gender, language and thresholds of Pure Tone Threshold Audiometry are inserted.	14
2.2	Whisper second technical information page, to enter the measurement device, select ear and set comfortable volume level.	14
2.3	Interface for Whisper test, where 3 alternatives closed-form questions are presented. English male speaker with background noise pronouncing VCVs.	15
2.4	Average psychometric curves for the four clusters of VCVs.	16
2.5	Example of the Matrix Test in italian. The OLSA test is the German version of this Matrix Test, where sentences are presented in this way but in german language.	17
2.6	Interface of the DTT, where digits are inserted.	18
2.7	Example of questions from the risk factors questionnaire. There are in total 18 questions.	19
2.8	Summary of the main features extracted from Whisper, risk factors questionnaire, and DST.	20
2.9	Summary of the main features extracted from the OLSA and the DTT tests.	21
2.10	Summary of the main features that have been subsequently elaborated for the analysis, all the tests.	22
2.11	Artificial Ear (Brüel & Kjær (B&K) 4153)	26
2.12	Pipeline of data preprocessing, clustering and classification analysis.	28
2.13	Illustration of k-fold cross validation with extension for hold outs (scikit-learn, 2021)[39]	43

3.1	Age histogram with highlighted PTA bands; minimum, maximum and mean values are reported.	49
3.2	Histogram of PTA	50
3.3	Scatterplot of PTA and Age	50
3.4	Histogram of SRT with highlighted PTA values.	51
3.5	Histogram of Total Time for Whisper Test.	51
3.6	Scatterplot of Total Time for Whisper test VS Age.	51
3.7	Histogram of Digit Span Score.	52
3.8	Histogram of Single Digit Average Typing Time	53
3.9	Scatterplot of Age and <i>avgSingleDigitTimes_mean</i>	53
3.10	Total time required to insert the first sequence of 3 digits.	54
3.11	Scatterplot of Time for first sequence and Age, highlighthed with PTA values.	55
3.12	Scatterplot of Time for the first sequence of DST and the Average time single digit typing on all digits, highlighted with Age.	56
3.13	Histogram of Average Single Digit Typing Time on Correct trials.	57
3.14	Histogram of Average Single Digit Typing Time on Wrong trials.	57
3.15	Scatterplot of correct trials VS wrong trials average single digit typing time highlighted by Age.	57
3.16	Number of Yes/No responses for each binary risk factor from the questionnaire.	58
3.17	Elbow Curve illustrating the reduction in the cost function of K-Prototypes as the number of clusters increases, using exclusively Whisper features.	59
3.18	Silhouette Score, Davies-Bouldin Score, Calinski-Harabasz Index as the number of clusters increases, using exclusively Whisper features.	59
3.19	3D representation of 3 clusters, with centroids and medoids highlighted (Whisper features only).	60
3.20	3D representation of 3 clusters only for ears with better PTA records, with centroids and medoids (Whisper features only).	61
3.21	3D representation of 4 clusters with centroids and medoids highlighted (Whisper features only).	63
3.22	3D representation of 4 clusters only for ears with better PTA records, with centroids and medoids (Whisper features only).	64
3.23	3D representation of 5 clusters with centroids and medoids highlighted (Whisper features only).	66
3.24	3D representation of 5 clusters only for ears with better PTA records, with centroids and medoids (Whisper features only).	67
3.25	t-SNE, perplexity = 50, learning_rate= 500, 3 clusters.	68

List of Figures	139
3.26 t-SNE, perplexity = 50, learning_rate= 500, 4 clusters.	69
3.27 t-SNE, perplexity = 50, learning_rate= 500, 5 clusters.	69
3.28 Explained variance from PCA as a function of the number of principal components considered-joined plots.	79
3.29 PCA Loadings matrix for the retained PCA components (absolute values). The size of the bubble is proportional to the contribution of the feature to that PCA component.	80
3.30 Elbow Curve for K-Means after PCA reduction.	81
3.31 TSNE 2D VS. PCA 2D with 3 clusters - using k-Means algorithm.	81
3.32 Percentage of Variance explained by the first 10 components provided by the FAMD feature reduction technique.	82
3.33 Contributions of the variables to the first dimension of FAMD analysis.	82
3.34 Contributions of the variables to the second dimension of FAMD analysis.	83
3.35 Biplot of quantitative variables for contribution to the first two dimensions of FAMD.	84
3.36 3D representation of 3 clusters, with centroids and medoids highlighted (Whisper + risk factor + DST features).	85
3.37 3D representation of 3 clusters, with centroids and medoids highlighted, only ears with better PTA records (Whisper + risk factor + DST features).	87
3.38 3D representation of 4 clusters, with centroids and medoids highlighted (Whisper + risk factor + DST features).	88
3.39 3D representation of 4 clusters, with centroids and medoids highlighted, only ears with better PTA records (Whisper + risk factor + DST features).	89
3.40 3D representation of 5 clusters, with centroids and medoids highlighted, with Whisper + risk factor + DST features.	91
3.41 3D representation of 5 clusters, with centroids and medoids highlighted, with Whisper + risk factor + DST features, using only ears with better PTA records.	92
3.42 Scatterplot of the SRT for the 3 tests, both ears, 17 subjects in total. The x-axis shows the Whisper ID for each subject.	103
3.43 Scatterplot of SRT for each single test - right ear, 17 ears.	103
3.44 Scatterplot of SRT for each single test - left ear, 17 ears.	104
3.45 Scatterplot of % Error for Right Ear, Whisper.	105
3.46 Scatterplot of % Error for Left Ear, Whisper.	105
3.47 Most mistaken consonants Whisper - Right Ear.	105
3.48 Most mistaken consonants Whisper - Left Ear.	105
3.49 Error Percentage for both ears - OLSA	106

3.50	Error Percentage for both ears - DTT	107
3.51	Delta (right ear - left ear) of SRT for the three tests.	108
3.52	Delta of Pure Tone Threshold Audiometry values (right ear - left ear) for single frequencies for the three tests.	108
3.53	Delta PTA for central frequencies (from 500 to 4000 Hz)	109
3.54	Delta PTA until high frequencies (8000Hz)	109
3.55	Spearman's correlation heatmap for SRT and PTA.	110
3.56	Ear with lower SRT value for each participant, in all the three tests compared.	111
3.57	Measurement duration for DTT, OLSA and Whisper.	112
3.58	Delta of PTA and SRT for each test (right - left).	114
3.59	Performance Group based on quartile distinction - Whisper Right Ear. . .	115
3.60	Performance Group based on quartile distinction - Whisper Left Ear. . .	115
3.61	Performance Group based on quartile distinction - OLSA Right Ear. . . .	115
3.62	Performance Group based on quartile distinction - OLSA Left Ear	115
3.63	Performance Group based on quartile distinction - DTT Right Ear. . . .	116
3.64	Performance Group based on quartile distinction - DTT Left Ear.	116

List of Tables

3.1	Centroid table for 3 clusters (Whisper features only).	60
3.2	Medoid table for 3 clusters (Whisper features only).	60
3.3	Centroids table for 3 clusters - only ears with better PTA records (Whisper features only).	62
3.4	Medoids table for 3 clusters - only ears with better PTA records (Whisper features only).	62
3.5	Centroid table for 4 clusters (Whisper features only).	63
3.6	Medoid table for 4 clusters (Whisper features only).	63
3.7	Centroid table for 4 clusters - only ears with better PTA records (Whisper features only).	65
3.8	Medoid table for 4 clusters - only ears with better PTA records (Whisper features only).	65
3.9	Centroid table for 5 clusters (Whisper features only).	66
3.10	Medoid table for 5 clusters (Whisper features only).	66
3.11	Centroid table for 5 clusters - only ears with better PTA records (Whisper features only).	67
3.12	Medoid table for 5 clusters - only ears with better PTA records (Whisper features only).	68
3.13	Classification report: Mean and standard deviation of performance metrics across 5 random splits into training and external test sets. Splitting method: training = 80%, test = 20%, 3 clusters (Whisper features only).	71
3.14	Classification report: Mean and standard deviation of performance metrics across 5 random splits into training and external test sets. Splitting method: training = 80%, test = 20%, 4 clusters (Whisper features only).	72
3.15	Classification report: Mean and standard deviation of performance metrics across 5 random splits into training and external test sets. Splitting method: training = 80%, test = 20%, 5 clusters (Whisper features only).	73
3.16	Centroids and Medoids table for 3 clusters.	85
3.17	Centroids and Medoids table for 3 clusters, ears with better PTA only.	87

3.18	Centroids and Medoids table for 4 clusters.	88
3.19	Centroids and Medoids table for 4 clusters, ears with better PTA only. .	90
3.20	Centroids and Medoids table for 5 clusters.	91
3.21	Centroids and Medoids table for 5 clusters, ears with better PTA only. .	93
3.22	Classification report: Mean and standard deviation of performance metrics across 5 random splits into training and external test sets. Splitting method: training = 80%, test = 20%, 3 clusters (Whisper + risk factor + DST features).	95
3.23	Classification report: Mean and standard deviation of performance metrics across 5 random splits into training and external test sets. Splitting method: training = 80%, test = 20%, 4 clusters (Whisper + risk factor + DST features).	96
3.24	Classification report: Mean and standard deviation of performance metrics across 5 random splits into training and external test sets. Splitting method: training = 80%, test = 20%, 5 clusters (Whisper + risk factor + DST features).	97
3.25	Delta of PTA min, max, mean values for all frequencies up to 8000 Hz. .	109
3.26	Table of average test duration.	112
3.27	Participants with significant time delta (20% or more)	113

List of Abbreviations

Abbreviation	Description
WHISPER	Widespread Hearing Impairment Screening and Prevention of Risk
OLSA	Oldenburg Sentence Test
DST	Digit Span Test
DTT	Digit Triplet Test
RF	Random Forest classification algorithm
SVM	Support Vector Machine classification algorithm
KNN	k-Nearest Neighbors classification algorithm
WHO	World Health Organization
Std	Standard Deviation
Var	Variance
DSS	Digit Span Score

Acknowledgements

Grazie alla professoressa Alessia Paglialonga e all'ingegnere Marta Lenatti, per avermi seguita e accompagnata con tanta attenzione ed entusiasmo nello svolgimento di questa tesi, e per avermi sostenuta nel desiderio di espanderla al di fuori dell'Italia. Grazie anche a Ania Warzybok-Oetjen e tutto lo SPHEAR team, per avermi accolta e guidata ad Oldenburg, permettendomi di arricchire il mio bagaglio tecnico e culturale.

“Per tutto c'è il suo tempo, c'è il suo momento per ogni cosa sotto il cielo: un tempo per nascere e un tempo per morire, un tempo per piantare e un tempo per sradicare ciò che è piantato” diceva l'Ecclesiaste, ed anche *“Dio ha fatto ogni cosa bella al suo tempo: egli ha perfino messo nei loro cuori il pensiero dell'eternità, sebbene l'uomo non possa comprendere dal principio alla fine l'opera che Dio ha fatta.”*

Per me questo è il tempo di concludere questo percorso, ma soprattutto il tempo per fermarmi ed essere grata. Ed è per questo che, in queste pagine, voglio dire grazie a tutte quelle persone che se lo meritano non solo per essermi state accanto nel percorso universitario, ma ancor di più per esserlo nella vita per intero.

Grazie a mia madre e mio padre, per avermi concesso l'opportunità e i mezzi per studiare e formarmi, avendo avuto fiducia e soprattutto investendo su di me. Grazie per i sacrifici che avete fatto e che fate per me, anche quando non li riconosco come dovrei. Grazie a mio fratello, perché in questi anni di crescita mi ha mostrato come sia così facile restare due bambini che discutono per le cose più futili ma si sanno voler bene in tanti modi mentre si sta, allo stesso tempo, diventando due adulti che imparano a pedalare da soli. E grazie a tutto il resto della mia famiglia, le nonne, gli zii e i cugini, perché la qualità del tempo insieme è inversamente proporzionale alla sua quantità.

Grazie alle due amiche che so di poter chiamare così per la vita. Grazie Ori, perché so di poter contare sul tuo sostegno e sulla tua comprensione sempre, anche quando non sei d'accordo con me. Perché il tuo mix di cervello e cuore è un'arma a doppio taglio ed è proprio quello che in questi anni mi ha permesso di crescere insieme a te, e che so che continuerà a farlo, in ogni stagione della vita che sono sicura divideremo (e grazie per aver portato Ati nella mia vita, senza il pao de queijo e il fricassè non sarei la stessa

persona). Grazie Fab, perché mi ispiri con il tuo equilibrio, il tuo coraggio nella vita, e perché condividi con me la passione per le ricette fallimentari e gli infiniti tentativi di riconoscere se un colore è caldo o freddo, tanto quanto so che condividiamo le cose più profonde, con la fiducia che si dà a pochi.

Grazie a tutti gli amici che tra Itri, Fondi e Milano mi hanno incoraggiata, sostenuta e aiutata a ricordare che oltre il computer e i libri c'è davvero tanto altro, so che ognuno di voi mi insegna qualcosa e sono grata di poter crescere così. Vi voglio bene.

Grazie alle compagne e i compagni di università che in questi anni sono stati indispensabili per me per arrivare a questo punto. Con voi ho imparato a riconoscere i miei limiti, le mie capacità, e l'importanza di saper chiedere aiuto. Sono sicura che non sarei arrivata fin qui oggi senza di voi. E grazie a Vittoria, alla coinquilina con la C maiuscola, l'unica con cui avrei potuto condividere una stanza per 4 anni così, a partire da bastoncini findus e le sfoglie, fino ad arrivare a pizze fatte in casa ed avocado toast, passando per fughe di gas scampate, porte cigolanti e tanto ancora che porterò con me sempre.

E grazie a te Davide, il mio team-mate preferito per ogni cosa, che sei entrato nella mia vita con questa magistrale e mi hai mostrato cosa significa la complicità, la premura e la bellezza di avere accanto prima di tutto un amico, e poi un compagno non solo di squadra ma per la vita, che di meglio non avrei potuto desiderare. Grazie per la fiducia e la stima che hai in me, per l'impegno che metti nel valorizzarmi e nel prenderti cura di me nel modo giusto, senza mai cercare di starmi avanti o dietro, ma sempre accanto. Non so davvero esprimere quanto io sia grata a Dio per una persona come te. Ci sarebbe davvero tanto altro di cui ringraziarti, ma so che ho una vita per continuare a farlo e ne sono decisamente entusiasta.

Grazie a quel pallone e quella rete che in questi anni sono stati non solo un mezzo di sfogo potente, ma che hanno anche insegnato a conoscermi in un modo diverso, inaspettato, più profondo e anche scomodo a volte(oltre a procurarmi tante prese in giro per la mia fissazione).

Ed in conclusione, proprio perché è questo che vale davvero la pena che resti impresso, voglio ringraziare Dio. Un Dio che ha saputo aspettarmi e curarmi in questi anni, un Dio che mi ha amata anche quando io non volevo saperne niente, un Dio che si rivela sempre più reale e che attraverso questo percorso ha saputo farsi vedere e che so che continuerà a mostrarsi ancora e ancora. Quel Dio che conosce i miei limiti, la mia testa dura e la mia incapacità, ma che allo stesso tempo sceglie di amarmi e di trasformarmi. Quello che so è che voglio sperimentarlo in ogni nuovo capitolo della vita che mi aspetta.

“Ecco, io sto per fare una cosa nuova; essa sta per germogliare; non la riconoscerete? Sì, io aprirò una strada nel deserto, farò scorrere dei fiumi nella steppa.”

Isaia 43:19 (NR06)

Solamente, grazie.

