



POLITECNICO
MILANO 1863

**SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE**

EXECUTIVE SUMMARY OF THE THESIS

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

LAUREA MAGISTRALE IN BIOMEDICAL ENGINEERING - INGEGNERIA BIOMEDICA

Author: ILARIA STAIANO

Advisor: PROF. ALESSIA PAGLIALONGA

Co-advisor: PROF. ANIA WARZYBOK-OETJEN, MARTA LENATTI

Academic year: 2023-2024

1. Introduction

Hearing loss, increasingly prevalent among adults, is a major cause of disability, affecting approximately 466 million people worldwide. It often goes undiagnosed, leading to cognitive decline, anxiety, and depression. Early detection through widespread screening is crucial to mitigate its impact.

Risk factors for hearing loss include congenital and acquired causes. Congenital causes involve genetic factors and complications during pregnancy, while acquired causes encompass lifestyle factors such as exposure to loud noises, cardiovascular diseases, smoking, and alcohol consumption. Awareness and prevention initiatives are essential to address these risks [1].

Speech-in-noise tests (SIN) assess the ability to understand speech in the presence of background noise, measuring speech reception thresholds (SRT). Different SIN tests offer insights into auditory processing and aid in evaluating hearing aids and auditory training programs. The three SIN tests employed in the present thesis are: Whisper test [2], Oldenburg Sentence Test (OLSA) [3], and Digit Triplet Test (DTT)[4]. The Whisper test uses a stimulus of a series of vocal-consonant-vocal (VCV) sequences that

need to be recognized (the number of stimuli varies according to a one-up-three-down adaptive procedure [1]). OLSA provides 20 grammatically correct sentences (composed of names, verbs, numerals, adjectives, objects), while DTT consists of 27 sequences of 3 digits each.

Hearing loss affects cognitive functions like working memory, crucial for tasks involving information storage and manipulation. Tests like the Digit Span Test (DST) help quantify working memory capacity and identify cognitive decline. In the DST, participants are presented with a sequence of digits one at a time, starting with sequences of 3 numbers. They must recall and type the sequence in the exact order. If correct, the next sequence increases by one digit (max 9 digits); otherwise, it remains the same length. The test stops when two consecutive incorrect responses occur, and the longest correctly recalled sequence length is recorded as the Digit Span Score (DSS).

This thesis aims to develop automated methods for multivariate analyses using machine learning on audiological data from the WHISPER and Virtual Hearing Clinic platforms [5]. The goal of the thesis is to identify hearing impairments and cognitive decline in population

screenings, employing clustering and classification techniques. A novel dataset from collaborative efforts between institutions will be created and used to cross-validate methodologies (comparing the three SIN tests), ensuring robust insights and conclusions.

2. Materials and Methods

2.1. Protocol - phase 1

The first phase of data collection took place at Politecnico di Milano, aiming to expand an existing dataset. Focus has been given on searching for volunteers aged 30-40 years, to enrich this population in the dataset. A total of 33 new subjects were added (53 records in total). Participants underwent Pure Tone Threshold Audiometry (with Amplaid 177+, Amplifon with TDH49 headphones) the DST, a risk factors questionnaire, and the Whisper test. Informed consent was obtained, and the study was approved by the Politecnico di Milano Research Ethical Committee (Opinion No. 13/2022, April 13, 2022).

2.2. Protocol - phase 2

The second phase of data collection was conducted at the University of Oldenburg, targeting normal-hearing German native listeners under 40 years old. In this phase, 17 subjects were tested, yielding a new dataset with 34 records, as each subject's ears were tested separately. The procedure included Pure Tone Threshold Audiometry, DST, risk factors questionnaire and 3 SIN tests (Whisper test, OLSA, and DTT). Audiometer Interacoustics AC40 and headphones HDA200 were employed. The last four tests were administered in a randomized order to prevent fatigue bias. Equipment calibration for DST and Whisper was performed using an artificial ear to ensure consistent output level (65 dB). The study maintained methodological consistency and ethical standards throughout both phases.

2.3. Data analysis

A clustering and classification analysis was performed on the final dataset, which included the original records from previous acquisitions (503 records) plus the records acquired during Phase 1 and Phase 2 (53 and 34 records, respectively),

totaling 590 records.

The clustering and classification analysis initially focused solely on Whisper test features that were present for all the tested subjects. The dataset with the full set of records was used, and K-prototypes clustering method was applied because of the presence of numerical and categorical variables, exploring 3-4-5 clusters based on results of the elbow curve and Silhouette, Davies-Bouldin, and Calinski indices. Clustering was performed on both the entire dataset (590 records) and a subset excluding the worse-performing records in terms of Pure Tone Average (PTA, i.e., the average of the thresholds obtained from Pure Tone Threshold Audiometry for the central frequencies, of 500, 1000, 2000, 4000 Hz) for subjects tested on both ears (434 records).

Subsequently, classification of the above-mentioned clusters was conducted using three algorithms: Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN), for all clusters. Three different splitting methods were employed, with 5-fold cross-validation:

- Train = 80%, Test = 20%;
- Train = best PTA, Test = remaining data;
- Train = data without Oldenburg measurements, Test = records acquired in Oldenburg.

Performance metrics (accuracy, precision, recall, f1-score) were evaluated for each model.

The same procedures were applied to the dataset that also included DST and risk factors features (i.e., DST and risk factor data are only available for the last 266 records acquired).

Also, visualization (t-SNE) and feature reduction (Principal Component Analysis, Factor Analysis of Mixed Data) techniques were exploited to further explore the data.

Additionally, a comparative analysis of DTT, OLSA, and Whisper in terms of SRT, test duration, and symmetry of the ears, was conducted using the dataset acquired during phase 2 (17 subjects).

3. Results

The characterization on the dataset with Whisper features only (590 records) showed a population with average age of 45.5 years. A total of 403 records presented PTA < 20dB (normal hearing), while PTA between 20dB and 35dB

(mild hearing loss) was found in 113 records. Finally, the records with PTA > 35 dB (moderate or severe hearing loss) were 74. The mean value of SRT is -11.75 dB suggesting a prevalence of normal-hearing subjects. Positive correlation between PTA and SRT is observed (value = 0.679). The time duration for Whisper test was, in average, 253.4 sec. The DSS has been, in average, 5.8. Other variables related to Whisper and DST (for example the typing time of the single digits for DST and other numerical variables about the performances in the two tests) were observed and also reprocessed in order to understand the data more deeply before the actual clustering.

3.1. Clustering and classification

Different configurations of clusters and algorithms of classification with also different splitting methods (Section 2.3) are reported in the thesis. Here, one representative case is reported: clustering and classification with Whisper features only, on 5 clusters. From the medoids shown in Table 1 it can be observed that the main distinctive features of each cluster are: PTA, Age and total_time, followed by SRT and #trials. The clusters characterized by groups of young subjects (about 28-36 years) are the ones that exhibit better values in terms of PTA and SRT, and also low total_time, while clusters characterized by older subjects have higher values. The first cluster shows prevalence of male (0) subjects, while the third (even if still characterized by younger subjects) shows more females (1). The second cluster, even if with age lower than the fourth one, shows the worst test performance (higher time, higher #trials, lower %correct).

Results for the classification of the 5 clusters with train = 80%, test = 20% shows high values of accuracy, precision, recall and f1-score for both Random Forest and SVM, with slightly higher values for SVM (KNN is worst performing). Specifically, SVM with best parameters $C = 100$, kernel = 'linear' maintains high accuracy on both training (0.992 ± 0.005) and test (0.979 ± 0.018) data, with stable precision (0.995 ± 0.003 on test) and recall (0.985 ± 0.013 on test) metrics. This suggests robust performance in handling the increased cluster complexity.

Considering other splitting methods with a fixed

test set, and observing their best performance, permits to note some challenges due to the high number of clusters and, on the other hand, low number of points for each cluster, increasing misclassification rate. For example, using the Oldenburg dataset as test set decreases the accuracy on the test for Random Forest (KNN shows 0.56). Conversely, when considering fewer clusters (e.g., 3 clusters), even though they are less distinctive as they are macro groups, they contain more points, making classification easier.

Adding features related to the DST and the risk factors questionnaire permits to discriminate using also the DSS and the average typing time during DST, but the variables that lead the clustering are evidently always the ones related to Whisper test. Classification of clusters using these extended number of features adds more challenges due to the further reduction of the number of data points (from 590 to 266), highlighting the need for more data to increase the reliability of the model.

3.2. Comparative analysis of Whisper test, OLSA and DTT

The dataset acquired during phase 2 (Section 2.2) was used to compare the three SIN tests (OLSA, DTT and Whisper). SRT values were always lower for Whisper than for the other two tests, that presented more comparable values (OLSA: -7.66 ± 1.13 dB, DTT: -9.09 ± 0.54 dB, Whisper: -16.64 ± 2.15 dB). Notably, variability for Whisper was higher in average compared to the other two tests. The right ear reached lower values of SRT for Whisper Test, while DTT and OLSA were almost symmetric in that sense (Figure 1). The difference between right and left ear in SRT for Whisper was in average -1.498 dB, for OLSA -0.147 dB, while for DTT it was -0.069 dB. In terms of test duration (Figure 2), Whisper took always more time than OLSA and DTT, considering also that the number of stimuli were not fixed because it works with an adaptive procedure ([1]), while the other tests are fixed ([4], [3]).

Subjects with higher delta of time between the two ears (second ear different from the first ear of 20% or more) were not found to be related in any particular ways to low SRT or PTA performances.

Medoid table for 5 clusters (Whisper features only).

| | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 |
|------------------------|----------|----------|----------|----------|----------|
| PTA[dB] | 12.5 | 20 | 1.25 | 26.25 | 17.5 |
| srt[dB] | -16.28 | -0.39 | -16.54 | -5.95 | -16.47 |
| Age | 36 | 63 | 28 | 73 | 51 |
| gender | 0 | 1 | 1 | 1 | 1 |
| %correct | 92.65 | 87.83 | 89.47 | 89.33 | 90.70 |
| #trials | 68 | 115 | 76 | 75 | 86 |
| total_time[sec] | 191 | 477 | 249 | 259 | 329 |

Table 1: Medoids coordinates of the five clusters using K-prototypes with Whisper features only. Cluster 1 contains 184 data points, Cluster 2 contains 178 data points, Cluster 3 contains 110 data points, Cluster 4 contains 100 data points, and Cluster 5 contains 18 data points.

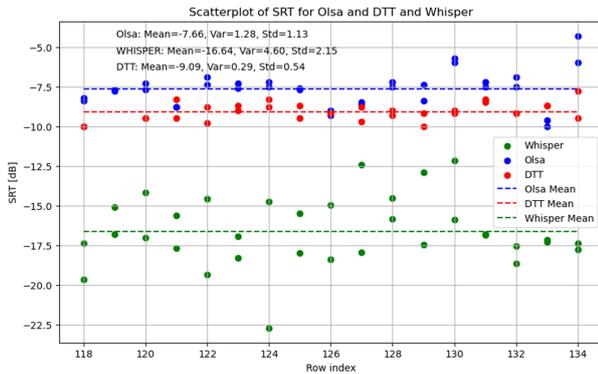


Figure 1: Scatterplot of SRT for OLSA (blue dots), DTT (red dots) and Whisper (green dots). Subjects are identified by their Whisper ID, both ears are reported.

An analysis on the percentage of mistaken answers on each test was made, reporting OLSA as the test with highest number of mistakes. It is interesting to note which consonants were most commonly mistaken in Whisper by the various subjects, for both ears. "M", "F", and "R" were consistently the most critical in every case. 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 easily masked by stationary speech shaped noise

4. Discussions

The study clustered subjects into 3, 4, and 5 groups using the K-Prototypes algorithm with

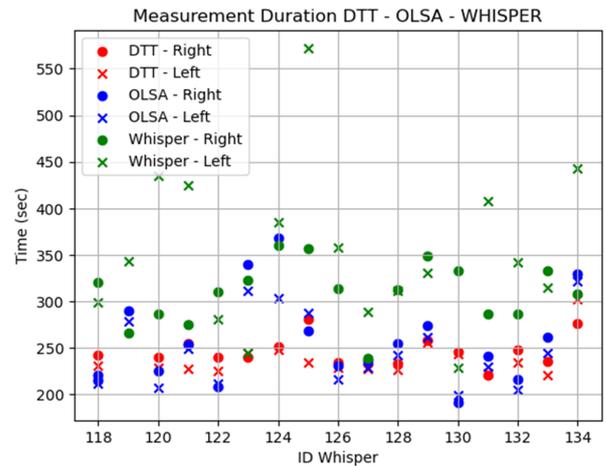


Figure 2: Measurement of test duration for OLSA (in blue), Whisper (in green) and DTT (in red). Dots represent right ears, while crosses represent left ears. Each subject is identified by their Whisper ID and time is measured in seconds. The average duration for right ear is: 245.93 sec (DTT), 256.99 sec (OLSA), 309.40 sec (Whisper). For left ear: 237.56 sec (DTT), 247.73 sec (OLSA), 353.52 sec (Whisper).

the objective of identifying similarities between subjects in terms of auditory and cognitive features. Key findings from this analysis are presented in the following.

4.1. Distinctive features in clustering

Age, PTA, number of trials, and total test time emerged as the most significant variables in distinguishing clusters of subjects. Clusters showed

a general trend identifying groups of subjects characterized by older subjects that tend to have higher PTA values and lower hearing performance. However, some older individuals maintained good test performance, suggesting age alone isn't always the single most important factor for hearing decline.

4.2. Cluster characteristics

With three clusters, the following profiles of subjects can be identified: younger subjects with lower PTA and better test performance (i.e., $PTA \leq 5$ dB, #trials around 77, test time of 215 sec), middle-aged subjects with moderate PTA (15 dB) and slightly worse test performance (i.e., #trials of 84, total time of 292 sec), and older subjects with higher PTA and longer test times ($PTA > 33$ dB, #trials of 97, test time > 430 sec). Increasing to four and five clusters allowed for more distinctions, including profiles of older subjects with high PTA but good test performance, and younger subjects with excellent hearing and cognitive performance (as shown in Table 1).

4.3. Classification performance

Random Forest and SVM algorithms showed robust performance across different cluster configurations, with test accuracy consistently high and no significant overfitting. KNN consistently underperformed compared to Random Forest and SVM, likely due to its sensitivity to the high dimensionality of features and small size of the dataset.

The dataset with more features (Whisper, cognitive test and risk factor features) but less records (266 records), presented more challenges in classification due to the smaller size, leading to some misclassification and instability in the test set, especially increasing the number of clusters.

The current profiles provide a snapshot of the included measures, showing audiological plausible distinctions among subjects based on PTA values and other audiological characteristics. The insights gained can inform clinical decision-making, personalized hearing aid programming, and auditory rehabilitation, emphasizing the need for larger, more diverse datasets to refine these analysis further.

4.4. Comparative analysis

Comparative analysis between Whisper, OLSA, and DTT tests on 17 normal-hearing subjects highlighted that Whisper presents lower values and more variability in terms of SRT, due to the easier speech material and the closed-form three alternatives structure of the test. Despite the lower values and higher variability, Whisper SRT remains informative in predicting PTA even when considering subjects with hearing loss (HL). The correlation between Whisper SRT and PTA, including subjects without HL and those with HL (>0.6), is similar to that reported in other speech-in-noise tests. The asymmetry shown between each ear for single test is also more evident in the Whisper test, even if people made more mistakes in the OLSA test, probably due to the more complex structure of the stimuli. In general, OLSA and DTT shows similar values for SRT.

4.5. Limitations and future research

The study's main limitation is the small dataset size, particularly when including cognitive and risk factor features, which affects the robustness and generalizability of the findings. Future research should focus on expanding the dataset to include more subjects and ensuring a balanced representation of different age groups and hearing loss levels. Addressing these limitations will improve the precision of subject characterization and the overall validity of the clustering and classification models.

5. Conclusions

In conclusion, the study's analyses on clustering and classification highlight the impact of various features on hearing and cognitive performance, confirming the trend of age-related hearing loss. Key variables identified include age, PTA, number of trials, and total test time. Despite the complexity added by incorporating additional features, the models, particularly Random Forest and SVM, maintained high performance. The primary limitation is the small dataset size, affecting robustness. Future research should expand the dataset and explore alternative techniques for small clusters. Additionally, comparative analysis of Whisper, OLSA, and DTT speech-in-noise tests revealed differences in SRT values and error rates due to the structure of the

tests, emphasizing the need for a comprehensive test battery to enhance data reliability and insights. The current analysis can be applied in clinical decision-support systems, mobile assessment tools, personalized hearing aid programming, and auditory rehabilitation programs to enhance patient care and monitoring. Additionally, it can inform public health initiatives for early detection and intervention of hearing and cognitive impairments.

6. Acknowledgements

A special thanks goes to the people who contributed to the development of this thesis: Professor and thesis advisor Alessia Paglialonga, co-advisor Marta Lenatti and Ania Warzybok-Oetjen, Professor Birger Kollmeier and the SP-HEAR team at the University of Oldenburg.

References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] University of Oldenburg, “Virtual hearing centre: Modules and perspectives.” <https://uol.de/vhc/modules-and-perspectives>, 2024. Accessed: 2024-06-24.