QUANTIFYING THE DISCRIMINATIVE VALUE OF AUDIOLOGICAL MEASUREMENTS FOR HEARING LOSS SEVERITY CLASSIFICATION

Marta Campi
Université Paris Cité
Institut Pasteur, AP-HP, INSERM, CNRS,
Fondation Pour l'Audition, Institut de l'Audition
IHU reConnect, F-75012, Paris, France
mcampi@pasteur.fr

Gareth W. Peters
Department of Statistics and Applied Probability
University of California Santa Barbara
Santa Barbara, CA 93106, United States
garethpeters@ucsb.edu

Perrine Morvan
Université Paris Cité
Institut Pasteur, AP-HP, INSERM, CNRS,
Fondation Pour l'Audition, Institut de l'Audition
IHU reConnect, F-75012, Paris, France
pmorvan@pasteur.fr

Mareike Buhl[†]
Université Paris Cité
Institut Pasteur, AP-HP, INSERM, CNRS,
Fondation Pour l'Audition, Institut de l'Audition
IHU reConnect, F-75012, Paris, France
mbuhl@pasteur.fr

Hung Thai-Van[†]
Université Paris Cité
Institut Pasteur, AP-HP, INSERM, CNRS,
Fondation Pour l'Audition, Institut de l'Audition
IHU reConnect, F-75012, Paris, France
htaivan@pasteur.fr

November 4, 2025

ABSTRACT

Current hearing loss classification relies predominantly on Pure-Tone Average (PTA), yet clinicians recognize that speech recognition abilities—particularly in noise—vary substantially among patients with comparable PTAs. Whether speech recognition testing provides discriminative information beyond pure-tone audiometry remains poorly quantified. We present a statistical framework to systematically identify which audiological measurements provide optimal discrimination between severity categories and quantify the added value of speech recognition testing. Analyzing 48,144 adults (ages 40-90) with symmetric hearing loss from Amplifon France databases, we applied multivariate hypothesis testing, advanced statistical methods for capturing complex dependencies between measurements, and bootstrap resampling to construct discriminative features from pure-tone audiometry (125-8000 Hz) and speech recognition testing in quiet (SRT$_Q$) and noise (SRT$_N$). Unsupervised clustering evaluated whether engineered features naturally separate established severity categories. Speech-in-noise testing emerged as the strongest discriminator, with features capturing complex dependencies achieving exceptional discriminative power (Silhouette score 0.94) compared to PTA-based features alone (0.50)—an 88% improvement. Mid-frequency pure-tone thresholds (1000-4000 Hz) showed highest discriminative value among audiometric measures. Adjacent severity categories showed limited discrimination through univariate measures but clear separation through multivariate feature combinations. These findings provide empirical evidence that speech recognition testing, particularly in noise, contributes substantial discriminative information beyond pure-tone audiometry, supporting extension of hearing loss classification systems to incorporate suprathreshold measures alongside threshold-based categories.

Keywords Hearing loss classification, Speech-in-noise testing, Pure-tone audiometry, Multivariate analysis, Clinical decision-making

[†]These authors contributed equally to this work.

1 Introduction

Hearing loss affects over 430 million people worldwide who require rehabilitation, with projections rising to 700 million by 2050 [1, 2]. Clinical assessment primarily relies on the Pure-Tone Average (PTA)—the mean hearing threshold at 500, 1000, 2000, and 4000 Hz—to classify impairment severity. Systems such as those proposed by the WHO, BIAP, and ASHA [3, 4] define discrete categories ranging from slight to severe hearing loss based on these PTA cutoffs.

PTA-based classification remains widely adopted because it offers a simple and interpretable summary of hearing sensitivity in the frequency range most critical for speech understanding. The derived severity categories facilitate clinical communication and treatment planning: a "moderate" loss conveys immediate meaning to both clinicians and patients. This approach has been validated through decades of practice [5, 6] and is easily standardized across clinics and populations.

However, the reliance on PTA as a single, average measure has intrinsic limitations. PTA treats hearing loss as one-dimensional—captured solely by detection thresholds—while auditory function is inherently multidimensional, encompassing both sound detection and speech comprehension, particularly under noisy conditions. Moreover, applying discrete category boundaries to a continuous scale introduces artificial divisions [7, 8]. Most importantly, PTA-based grading provides no principled method to integrate suprathreshold measures such as speech recognition, even though these tests capture processing abilities beyond simple audibility [9, 10]. Individuals with similar PTAs may show large differences in speech understanding [11, 10], underscoring that these measurements reflect complementary and partly independent auditory dimensions.

The World Health Organization's *World Report on Hearing* (2021) acknowledges this explicitly: although WHO hearing loss grades are defined by PTA cutoffs, the report notes that speech understanding cannot be inferred from PTA alone [4]. This recognition highlights a persistent gap between threshold-based classification and functional hearing ability—one that suprathreshold measures may help to bridge.

To address this gap, researchers have developed composite indices and data-driven auditory profiles. Early work combined multiple test results [12, 13], while recent approaches apply unsupervised learning to identify hearing subtypes directly from clinical data [14, 15, 16, 17]. [18] developed Common Audiological Functional Parameters (CAFPAs) demonstrating the need for comprehensive test batteries beyond PTA [19, 20, 21]. Extended assessment frameworks have explored high-frequency audiometry [22, 23] and suprathreshold processing measures [24, 25].

Despite these advances, a central question remains: do speech recognition tests provide discriminative information beyond pure-tone audiometry, and which measurements most effectively differentiate hearing loss severity levels? Existing studies often combine measures without quantifying their relative contributions, identify clusters without linking them to established severity grades, or refine categorical boundaries without determining which variables best support them. The discriminative value of each measure has not been systematically quantified.

This question has both scientific and clinical relevance. In practice, pure-tone audiometry typically initiates assessment, guiding whether further tests are warranted [5]. Identifying which additional measures contribute the most discriminative information could optimize testing protocols—especially for speech recognition, which may reveal suprathreshold deficits not reflected in detection thresholds alone.

The present study systematically quantifies the discriminative power of audiological measurements—individually and in combination—across established severity categories. We focus on whether speech recognition tests provide additional information beyond PTA and on which frequency ranges or dependencies most enhance severity discrimination.

Answering this question requires moving beyond traditional univariate approaches to capture complex measurement interdependencies. Standard analyses examining measurements in isolation cannot reveal whether, for example, the combination of speech-in-noise performance and mid-frequency hearing thresholds provides discriminative information that neither measurement alone captures. Patients with similar average thresholds but different patterns of speech-threshold dependency may belong to functionally distinct severity groups—relationships that traditional correlation-based methods would miss entirely.

To this end, we integrate complementary statistical methods: univariate and multivariate hypothesis testing, copula-based dependency analysis, and unsupervised clustering. Copula methods capture complex relationships between measurements that traditional correlation cannot detect—including whether relationships differ across severity levels and whether extreme values (very poor or very good performance) on one measure predict extreme values on another. For example, speech recognition and pure-tone thresholds might show different dependency patterns in severe hearing loss than in mild loss, or certain combinations of deficits might be characteristic of specific severity categories. Bootstrap procedures address class imbalance, and clustering validation tests whether identified features naturally align with clinical severity categories.

We analyze data from a large clinical database of 48,144 adults assessed at Amplifon France centers, including pure-tone audiometry (11 frequencies, 125–8000 Hz) and speech recognition in quiet (SRT$_Q$) and in noise (SRT$_N$). SRT$_N$ is of particular interest as it reflects real-world listening challenges and suprathreshold auditory processing.

Our objectives are fourfold:

1. Quantify the added discriminative value of speech recognition tests beyond pure-tone audiometry.
2. Identify which measurements and frequency regions provide the most diagnostic information for severity differentiation.
3. Assess whether multivariate combinations capturing interdependencies outperform individual measures, particularly between adjacent severity categories.
4. Develop and validate a generalizable statistical framework to quantify discriminative value using multivariate testing, copula-based analysis, and clustering validation.

This work contributes both to methodological development and to clinical practice. By identifying which measures best distinguish hearing loss categories, it informs evidence-based prioritization of audiological tests, supports extended classification systems incorporating suprathreshold information, and ultimately enhances the precision and efficiency of clinical hearing assessment.

2 Methods

This section details the statistical methodology used to identify and quantify the discriminative power of audiological features across hearing loss severity categories. Rather than developing a new classification system, our approach systematically evaluates which measurements—and which statistical transformations of these measurements—most effectively distinguish between established PTA-based severity categories (slight, mild, moderate, moderately severe, severe).

Our methodology comprises three integrated stages. First, we apply a comprehensive battery of statistical hypothesis tests to identify which audiological measurements show significant differences between severity categories. These tests examine means, variances, distributions, and complex multivariate dependencies, providing a principled basis for feature selection. Second, we employ sophisticated feature engineering techniques, including parametric and non-parametric bootstrap methods combined with copula-based dependency analysis, to transform raw measurements into a higher-dimensional feature space that captures discriminative information. Third, we validate the discriminative power of engineered features through unsupervised clustering analysis, objectively assessing whether selected features naturally separate patients into severity categories without forcing predetermined classifications.

Throughout this process, we maintain focus on quantifying discriminative value rather than proposing clinical replacements. By comparing clustering performance across different feature sets—from raw measurements to sophisticated statistical transformations—we establish which audiological information provides the strongest evidence for severity differentiation. This framework bridges rigorous statistical methodology with clinical interpretability, enabling evidence-based assessment of measurement priorities in audiological protocols.

2.1 Notation and Data Structure

We establish formal notation for describing audiological measurements and their transformations. Uppercase notation denotes random quantities such as random variables, while lowercase denotes realizations obtained from measurements. Bold face indicates vectors, and non-bold face represents scalars or matrices. Subscripts index dimensions of arrays or sets.

Denote by $\mathbf{X}_{N\times D}$ the random variables for measurements and its realizations $\mathbf{x}_{N\times D}$ from observed experiments, where N represents total sample size and D the number of attributes collected (13 total: pure tone thresholds at 11 frequencies [125, 250, 500, 750, 1000, 1500, 2000, 3000, 4000, 6000, 8000 Hz], speech recognition thresholds in quiet (SRT$_Q$) and noise (SRT$_N$)).

The observed attributes are mapped into d' features, where $d' \geq D$, obtained from transformations of these D observed attributes. Data comprise $G = 5$ labeled groups corresponding to hearing loss severity categories, with the g-th group data consisting of n_g participants each having D observations, denoted by $\{\mathbf{x}^{(g)}_{n_g \times D}\}$. The j-th participant in group g has observation vector $\mathbf{x}^{(g)}_j = [x^{(g)}_{j,1}, x^{(g)}_{j,2}, \ldots, x^{(g)}_{j,D}]$. Note that $\sum_{g=1}^{5} n_g = N$.

Denote $\mathcal{G} = \{1, \ldots, G\}$ the set of groups such that $g \in \mathcal{G}$ for every g. The five groups correspond to: group 1 ($g = 1$) slight hearing loss (16-25 dB HL); group 2 ($g = 2$) mild (26-40 dB HL); group 3 ($g = 3$) moderate (41-60 dB HL); group 4 ($g = 4$) moderately severe (61-80 dB HL); group 5 ($g = 5$) severe (>81 dB HL).

Population mean and standard deviation for the g-th group attribute d are denoted by $\mu^{(g)}_d$ and $\sigma^{(g)}_d$ respectively, with sample estimators $\widehat{\mu}^{(g)}_d$ and $\widehat{\sigma}^{(g)}_d$. The pooled variance estimator is denoted by S^2_p, and $F_{X^{(g)}_d}(u)$ denotes the cumulative density function of random variable $X^{(g)}_d$, referring to attribute d of the g-th group.

2.2 Statistical Tests for Feature Discrimination

To systematically identify which audiological measurements provide discriminative information between hearing loss severity categories, we employ a comprehensive battery of statistical hypothesis tests. These tests evaluate whether sample quantities differ significantly between categories, thereby identifying potential features for our discriminative framework. Table 1 presents the complete test battery with respect to two general groups (denoted as group i and group j), representing one pairwise combination between hearing loss categories. These tests serve to screen and select relevant test statistics that capture discriminative patterns across severity levels.

The test battery encompasses three primary classes of univariate analyses. The first class examines mean differences through both standard t-Student tests (assuming equal variances) and Welch's t-test [26] (accommodating unequal variances). The second class targets variance differences using the variance ratio test (F-test) [27] and Bartlett Test [28], assessing whether variances across groups can be considered equal. The third class investigates distributional differences through the Kolmogorov-Smirnov test [29], which is sensitive to differences in location and shape of empirical distribution functions, and the Cramer-von-Mises test with different weighting functions [30], which can be particularly effective for detecting differences in heavy-tailed distributions.

Beyond univariate approaches, we employ multivariate tests to examine interdependencies between audiological measurements. The sparse covariance matrix comparison method [31] enables detection of subtle differences between adjacent hearing loss categories where covariance differences may be sparse. Tukey's Honestly Significant Difference (HSD) test [32] provides controlled pairwise mean comparisons across multiple groups while controlling family-wise error rates. For capturing complex dependence structures between different audiological measurements, we employ copula-based tests [33], which evaluate equality between dependence structures while separating these from marginal behaviors.

2.2.1 Dependence and Concordance Measures

While the copula test provides insights into overall dependence structure, additional measures of dependence and concordance offer complementary information about relationships between audiological measurements. These measures capture different aspects of dependence structure between measurement coordinates and remain invariant under monotone transformations of the data [33, 34, 35].

For any pair of measurements $(X_{ld}, Y_{md'})$, where l, m index observations and d, d' index attributes from groups i and j respectively, with corresponding sample sizes n_i and n_j, we transform the data to ranks:

$$U_{ld,n_i} = \frac{\text{rank}(X_{ld})}{n_i + 1}, \quad V_{md',n_j} = \frac{\text{rank}(Y_{md'})}{n_j + 1}.$$

We employ multiple copula-based dependence measures including modified Kendall's tau and Spearman's rho (capturing overall concordance patterns), sign-based association and Gini-based measures (robust to outliers and sensitive to tail behavior), local Gaussian correlation, and tail dependence coefficients (identifying extreme value relationships). These measures collectively provide comprehensive characterization of dependence structures in audiological data, enabling identification of which measurement pairs exhibit the strongest discriminative relationships across severity categories. Complete formulas for all copula-based measures are provided in Supplementary Appendix, Table S1.

Table 1: Statistical tests for feature discrimination. For each test we provide: the quantity tested, test name, null and alternative hypotheses (H_0 and H_1), test statistic, and distribution under the null with degrees of freedom where appropriate.

Univariate Tests					
Feature	Test	H_0	H_1	Test Statistic	Distribution & DOF
Mean	T-test	$\mu_d^{(g)} = \mu_d^{(h)}$	$\mu_d^{(g)} \neq \mu_d^{(h)}$	$T = \frac{\left(\bar{X}_d^{(g)} - \bar{X}_d^{(h)}\right)}{S_p^2 \sqrt{\frac{1}{n_g} + \frac{1}{n_h}}}$	Student's t, DOF: $n_g + n_h - 2$
Mean	Welch T-test	$\mu_d^{(g)} = \mu_d^{(h)}$	$\mu_d^{(g)} \neq \mu_d^{(h)}$	$T = \frac{\left(\bar{X}_d^{(g)} - \bar{X}_d^{(h)}\right)}{\sqrt{\frac{S_d^{2^{(g)}}}{n_g} + \frac{S_d^{2^{(h)}}}{n_h}}}$	Student's t, DOF: Welch–Satterthwaite
Variance	Variance Ratio	$\sigma_d^{2^{(g)}} = \sigma_d^{2^{(h)}}$	$\sigma_d^{2^{(g)}} \neq \sigma_d^{2^{(h)}}$	$F = \frac{S_d^{2^{(g)}}}{S_d^{2^{(h)}}}$	Fisher–Snedecor DOF: $F_{(n_g-1,\ n_h-1)}$
Distr.	Kolmogorov Smirnov	$F_d^{(g)}(x) = F_d^{(h)}(x)$	$F_d^{(g)}(x) \neq F_d^{(h)}(x)$	$D = \sup_x \left\| \hat{F}_d^{(g)}(x) - \hat{F}_d^{(h)}(x) \right\|$	Free (no DOF)
Distr.	Cramer-Von Mises	$F_d^{(g)}(x) = F_d^{(h)}(x)$	$F_d^{(g)}(x) \neq F_d^{(h)}(x)$	$Q = n_g n_h \int_{-\infty}^{\infty} w(x)[\hat{F}_d^{(g)}(x) -$ $\hat{F}_d^{(h)}(x)]^2 d\hat{F}_d^{(h)}(x)$	Free (no DOF)
Multivariate Tests					
Variance	Bartlett Test	$\sigma_d^{2^{(g)}} = \sigma_d^{2^{(h)}}$ $\forall (g,h)$	$\sigma_d^{2^{(g)}} \neq \sigma_d^{2^{(h)}}$ for at least one (g,h)	$T = \frac{(N-G)\ln(S_p^2) - \sum_{l=1}^{G}(n_g-1)\ln(S_d^{2^{(g)}})}{1+\left(\frac{1}{3(G-1)}\right)\left[\left(\sum_{l=1}^{G}\frac{1}{n_g-1}\right) - \frac{1}{N-G}\right]}$	Chi-Square $\chi^2_{(G-1)}$
Covariance	Sparse Cov.	$\mathbf{\Sigma}_g = \mathbf{\Sigma}_h$	$\mathbf{\Sigma}_g \neq \mathbf{\Sigma}_h$	$M_n - 4\log p + \log\log p$	Type I extreme value (no DOF)
Tukey	Tukey HSD	$\mu_d^{(g)} = \mu_d^{(h)}$ $\forall (g,h)$	$\mu_d^{(g)} \neq \mu_d^{(h)}$ for at least one g, h	$W = \frac{\max_{(g,h)}\left(\bar{X}_d^{(g)} - \bar{X}_d^{(h)}\right)}{\sqrt{\frac{1}{2}\frac{S_d^{2^{(g)}} + S_d^{2^{(h)}}}{n_{g,d}+n_{h,d}}}}$	Studentized range DOF: $q_{(G,\ N-G)}$
Copula	Copula Test	$C_g = C_h$	$C_g \neq C_h$	$E_{n_g,n_h} = \frac{\hat{C}_g - \hat{C}_h}{\sqrt{\frac{1}{n_g} + \frac{1}{n_h}}}$	Free (no DOF)

2.3 Feature Engineering

Building upon statistically identified discriminative features, we employ bootstrap-based feature engineering to create robust representations suitable for clustering analysis. This approach addresses two key challenges: (1) class imbalance across severity categories, and (2) the need for sufficient sample sizes to reliably assess discriminative power.

2.3.1 Parametric and Non-Parametric Bootstrap Approaches

Bootstrap methods [36] provide a simulation-based framework for generating additional samples while preserving underlying statistical properties. We implement both parametric and non-parametric approaches to ensure robustness against distributional assumptions.

Parametric Bootstrap The parametric approach fits a model $f(x|\theta)$ to observed data, then generates new samples from this fitted distribution.

```
Parametric Bootstrap Procedure
Input: Parametric model $f(x|\hat{\theta})$ fit to $X_1, \ldots, X_n$
For $i = 1, 2, \ldots, B$:
1. Simulate iid samples $X_1^*, \ldots, X_n^* \sim f(x|\hat{\theta})$
2. Compute statistic $T^* := T(X_1^*, \ldots, X_n^*)$
Output: Empirical distribution of $T^*$ across $B$ simulations
```

Non-Parametric Bootstrap The non-parametric approach uses the empirical distribution, placing mass $\frac{1}{n}$ at each observed value, generating samples through resampling with replacement.

```
Non-Parametric Bootstrap Procedure
Input: $X_1, \ldots, X_n$
For $i = 1, 2, \ldots, B$:
1. Sample $X_1^*, \ldots, X_n^*$ with replacement from $X_1, \ldots, X_n$
2. Compute statistic $T^* := T(X_1^*, \ldots, X_n^*)$
Output: Empirical distribution of $T^*$ across $B$ simulations
```

Copula-Based Bootstrap Extensions To capture multivariate dependencies, we extend bootstrap methods using copula functions, which separate dependence structure from marginal distributions.

Parametric Copula Bootstrap: Assumes joint distribution $F_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = C_\theta(F_1(x_1), \ldots, F_d(x_d))$, where C_θ is a t-copula with parameters $\theta = (R, \nu)$.

```
Parametric Student-t Copula Bootstrap
Input: Data matrix $X_{n \times d}$; Fitted $t$-copula parameters $\theta = (R, \nu)$; Marginal distributions $\hat{F}_1, \ldots, \hat{F}_d$
For $i = 1, 2, \ldots, B$:
1. Sample $(u_{i1}, \ldots, u_{id}) \sim tCopula(R, \nu)$, $i = 1, \ldots, n$
2. Structure as $U_1^*, \ldots, U_n^*$ where $U_i^* = (u_{i1}, \ldots, u_{id})$
3. Compute statistic $T^* := T(U_1^*, \ldots, U_n^*)$
Output: Empirical distribution of $T^*$ based on $B$ replicates
```

Non-Parametric Copula Bootstrap: Estimates a flexible Bernstein copula $\widehat{C}$ from rank-transformed data.

```
Non-Parametric Bernstein Copula Bootstrap
Input: Data matrix $X_{n \times d}$; Bernstein copula fit $\widehat{C}$; Marginal distributions $\hat{F}_1, \ldots, \hat{F}_d$
For $i = 1, 2, \ldots, B$:
1. Fit Bernstein copula $\widehat{C}$ to rank-transformed data
2. Sample $(u_{i1}, \ldots, u_{id}) \sim \widehat{C}$, $i = 1, \ldots, n$
3. Structure as $U_1^*, \ldots, U_n^*$ where $U_i^* = (u_{i1}, \ldots, u_{id})$
4. Compute statistic $T^* := T(U_1^*, \ldots, U_n^*)$
Output: Empirical distribution of $T^*$ based on $B$ replicates
```

2.3.2 Feature Construction from Bootstrap Samples

For each statistically significant feature identified through hypothesis testing, we generate n' new samples for each group g to form balanced feature vectors suitable for clustering analysis. Let $X_{g,(d_k)}$ denote observed data for group $g \in \mathcal{G}$ and attribute d_k.

Bootstrap Setup We generate $N' = 5 \times n'$ new samples (ensuring balanced representation across all 5 groups) under both parametric and non-parametric approaches:

1. **Parametric (Normal):** Model $X_{g,(d_k)}$ as $\mathcal{N}(\hat{\mu}_g, \hat{\sigma}_g^2)$
2. **Non-Parametric:** Sample n' times with replacement from $X_{g,(d_k)}$

Feature Construction for Univariate Tests For each significant test type, we derive corresponding summary statistics as features:

- **Mean Test:** Compute bootstrapped means for each group, concatenating across groups:

$$\widetilde{\mathbf{x}}_{N'\times 1} = \big[\bar{x}^{(1)}_{1,(d_k)}, \ldots, \bar{x}^{(1)}_{n',(d_k)}, \ldots, \bar{x}^{(5)}_{1,(d_k)}, \\ \ldots, \bar{x}^{(5)}_{n',(d_k)}\big]^\top$$

- **Variance Test:** Compute bootstrapped variances:

$$\widetilde{\mathbf{x}}_{N'\times 1} = \big[s^{2^{(1)}}_{1,(d_k)}, \ldots, s^{2^{(1)}}_{n',(d_k)}, \ldots, s^{2^{(5)}}_{1,(d_k)}, \\ \ldots, s^{2^{(5)}}_{n',(d_k)}\big]^\top$$

- **Distribution Test:** Compute robust summaries (median, IQR, 5th/95th percentiles):

$$\widetilde{\mathbf{x}}_{N'\times 4} = \Big[\Big(\text{med}^{(1)}_{1,(d_k)} \quad \text{IQR}^{(1)}_{1,(d_k)} \quad p5^{(1)}_{1,(d_k)} \quad p95^{(1)}_{1,(d_k)}\Big), \ldots\Big]^\top$$

Feature Construction for Multivariate Tests - Copula For multivariate relationships identified through copula tests between attributes d_k of group g and d_l of group h:

- **Rank Transformations:** Transform bootstrapped samples to rank scale:

$$\widetilde{\mathbf{x}}^{\text{rank}}_{N'\times 2} = \Big[\Big(U^{(g)}_{1,(d_k)} \quad U^{(h)}_{1,(d_l)}\Big), \ldots, \\ \Big(U^{(g)}_{n',(d_k)} \quad U^{(h)}_{n',(d_l)}\Big)\Big]^\top$$

- **Dependence Measures:** Compute copula-based dependence measures (Supplementary Appendix, Table S1):

$$\widetilde{\mathbf{x}}^{\text{cop}}_{N'\times M} = \Big[\Big(\tau^{(g,h)}_{\text{cop},1} \quad \rho^{(g,h)}_{\text{cop},1} \quad \cdots \quad \lambda^{(g,h)}_{L,1} \quad \lambda^{(g,h)}_{U,1}\Big), \ldots\Big]^\top$$

Feature Construction for Covariance and Bartlett Tests For significant covariance relationships and variance comparisons across groups, we similarly construct correlation and variance ratio features through bootstrap sampling.

This feature engineering process produces feature matrices:

$$\widetilde{\mathbf{X}}^{\mathcal{N}}_{N'\times D'}, \quad \widetilde{\mathbf{X}}^{\text{NP}}_{N'\times D'}$$

corresponding to Normal parametric ($\mathcal{N}$) and Non-Parametric (NP) bootstrapping, with sample sizes $n' \in \{50, 500, 1000, 5000\}$ to assess the effect of bootstrap sample size on discriminative power.

2.3.3 Feature Space Dimensionality and Combinations

The feature engineering process produces feature spaces of varying dimensionality depending on the type and combination of features employed. Table 2 summarizes the dimensionality of different feature categories.

Table 2: Feature space dimensionality for different feature types and combinations.

Feature Type	**Unscreened Dimension**	**Screened Dimension**
	Individual Features	
Frequency Univariate	11	3
Speech Univariate	2	2
Frequency Copula	165	99
Speech Copula	90	81
	Combined Features	
Univariate (Freq + Speech)	13	8
Univariate Full (Mean + Var + Dist)	39	24
Copula (Freq + Speech)	255	180
All Features (Univariate + Copula)	741	534

For individual features, screening based on statistical significance ($p < 0.05$) substantially reduces dimensionality while retaining discriminative power. Frequency univariate features reduce from 11 to 3 dimensions when screened, focusing on the most significant frequencies (1000Hz, 2000Hz, 4000Hz) identified through hypothesis testing. Speech-related features maintain both dimensions (SRT_Q and SRT_N) as both prove significant.

Copula-based features represent the highest-dimensional category, with screened versions maintaining 81-99 dimensions for speech and frequency measures respectively. These capture complex dependencies between different audiological measurements, providing rich discriminative information.

Combined feature sets demonstrate the trade-off between complexity and information content. Univariate combinations maintain relatively low dimensionality (8-39 dimensions) while integrating multiple statistical measures. Full cross-feature combinations incorporating both univariate and copula measures span higher dimensions (192-741), particularly in unscreened versions.

The screening process plays a crucial role in dimensionality reduction while preserving discriminative power. This hierarchical organization of features, from individual measures to sophisticated combinations, allows flexible adaptation to different clustering scenarios while maintaining interpretability.

2.4 Clustering Analysis for Discriminative Power Assessment

To objectively evaluate the discriminative power of engineered features, we employ unsupervised clustering methods. Unlike supervised approaches that force predetermined classifications, unsupervised methods assess whether identified features naturally separate patients into meaningful groups corresponding to severity categories.

We employ two complementary clustering approaches: K-means clustering provides efficient partitioning based on centroid distances, while hierarchical clustering with Ward's method captures nested relationships reflecting progressive hearing loss patterns. Both methods are configured to identify $k = 5$ clusters, corresponding to the five established severity categories.

The K-means algorithm partitions observations into clusters by iteratively minimizing within-cluster sum of squares. For robustness, we perform multiple restarts with different random initializations, selecting the solution with minimal total within-cluster variance.

Hierarchical clustering with Ward's method builds a dendrogram revealing nested cluster structure. Ward's criterion minimizes total within-cluster variance while merging clusters, making it particularly suitable for detecting gradations in hearing loss severity.

Clustering performance is evaluated using the Silhouette score [37], which quantifies both cluster cohesion (how similar objects are within clusters) and separation (how distinct clusters are from each other). Scores range from -1 to 1, with values exceeding 0.5 indicating well-separated clusters and 1.0 representing perfect separation.

Full technical details of clustering algorithms and additional evaluation metrics are provided in the Supplementary Appendix.

3 Data Description

This section describes the audiological dataset used to evaluate discriminative power of different measurements and feature transformations across hearing loss severity categories.

3.1 Data Acquisition and Testing Procedures

We utilize data from Amplifon France hearing aid fitting practices, provided in pseudonymized form to Institut Pasteur under the BIG DATA AP project. The Commission Nationale de l'Informatique et des Libertés authorized data processing on April 05, 2024.

The dataset includes participants' age, sex assigned at birth, pure-tone audiograms for both ears, and speech recognition thresholds in quiet and noise. Hearing loss severity was categorized using Pure-Tone Average (PTA) based on thresholds at 0.5, 1, 2, and 4 kHz [3], following ASHA classification (Table 3).

We focused on participants aged 40-90 years with symmetric hearing loss, defined by PTA difference less than 15 dB between ears [38]. This age range was selected based on data availability and completeness. The final dataset comprises 48,144 participants. Data on race or ethnicity were not collected per French legal restrictions [39].

Table 3: Pure-tone average (PTA) categories following ASHA classification [3]. Our dataset contains no participants with normal hearing.

Degree of hearing loss	PTA range (dB HL)
Normal	–10 to 15
Slight	16 to 25
Mild	26 to 40
Moderate	41 to 55
Moderately severe	56 to 70
Severe	71 to 90

3.2 Dataset Characteristics

Table 4 summarizes descriptive statistics for the complete dataset. Mean participant age is 73 years (SD = 9.73), ranging from 40 to 90 years. Mean hearing thresholds increase with frequency, from approximately 30 dB HL at 125 Hz to 72 dB HL at 8 kHz, with standard deviations ranging from 13 to 19 dB HL. Mean SRT_N is 4.43 dB SNR (SD = 3.96), while mean SRT_Q is 45.97 dB SPL (SD = 11.56).

Table 4: Descriptive statistics for the complete sample. Variables include age, audiogram frequencies, SRT_N (dB SNR) and SRT_Q (dB SPL).

Statistics	Age	Frequencies (Hz)											SRT_N	SRT_Q
		125	250	500	750	1000	1500	2000	3000	4000	6000	8000		
Mean	72.98	30.48	31.24	33.83	36.40	38.11	45.08	48.48	55.79	61.74	70.35	71.71	4.43	45.97
Median	74.00	30.00	30.00	30.00	35.00	35.00	45.00	50.00	55.00	60.00	70.00	70.00	4.00	45.00
SD	9.73	13.19	14.50	15.14	15.43	15.68	16.18	16.28	16.50	16.97	18.24	18.75	3.96	11.56
Min	40	-10.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	-5.00	3.00	-10.00	5.00
Max	90	120.00	120.00	120.00	120.00	120.00	120.00	120.00	125.00	125.00	125.00	130.00	20.00	80.00

Figure 1 shows sample distribution across severity categories. The majority fall into Moderate (20,246) and Mild (18,979) categories. Moderately severe (4,826) and slight (3,704) categories have fewer individuals, while severe (389) represents the smallest group, indicating lower prevalence of severe hearing loss.

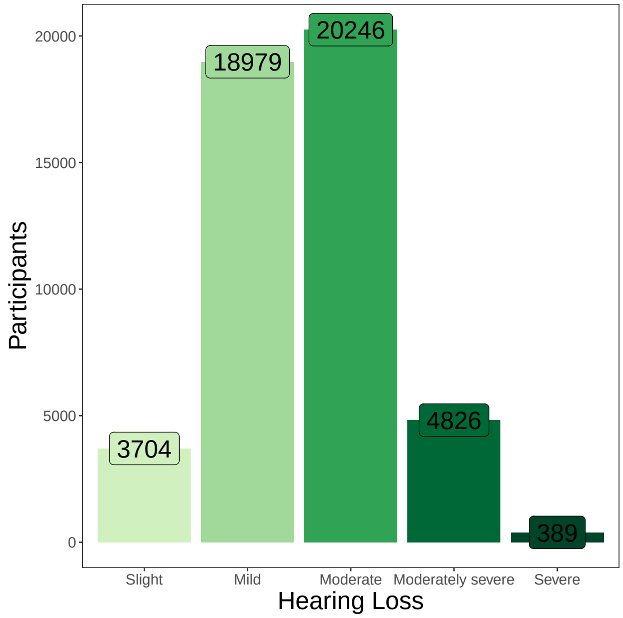

Figure 1: Sample size distribution across PTA-based hearing loss severity categories.

Figure 2 displays distributions via violin plots. Pure-tone thresholds vary across age groups, with lower thresholds from 40-45 to 65-70 years, particularly at lower frequencies (125-1000 Hz). Thresholds increase markedly at higher frequencies with age, rising from 55.44 dB (40-45 years) to 81.85 dB (85-90 years), indicating age-related hearing loss. SRT$_N$ and SRT$_Q$ values also increase with age: mean SRT$_N$ rises from 2.79 dB to 6.94 dB, while SRT$_Q$ increases from 40.70 dB to 53.73 dB.

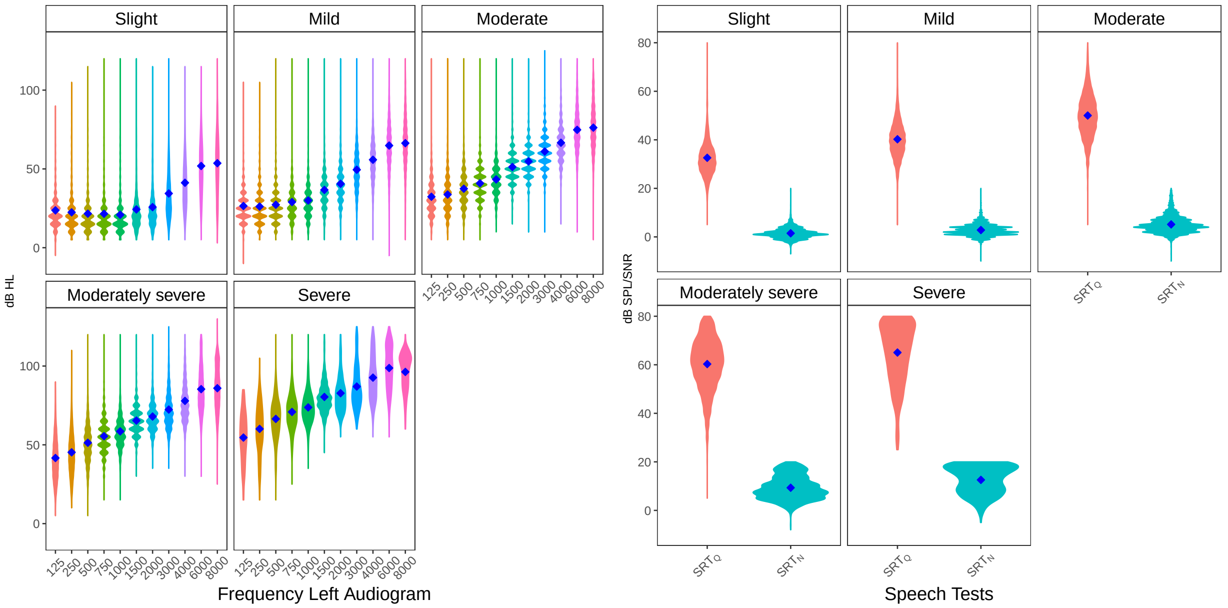

Figure 2: Violin plots of hearing thresholds at different frequencies (left) and SRT$_Q$, SRT$_N$ (right) by hearing loss degree for the left ear. X-axis on left shows frequencies 125-8000 Hz; x-axis on right shows speech tests. Y-axis shows thresholds in dB HL (left) and dB SPL/SNR (right). Due to symmetric hearing loss, left ear was selected; right ear shows equivalent patterns.

This dataset provides robust foundation for evaluating discriminative power of audiological measurements across hearing loss severity categories, with sufficient sample sizes in most categories and comprehensive measurement coverage including both pure-tone and speech recognition assessments.

4 Results

This section presents a systematic evaluation of discriminative power across audiological measurements and their statistical transformations. Rather than proposing a new classification system, we quantify which features—from raw measurements to sophisticated statistical contrasts—most effectively distinguish between established hearing loss severity categories. The analysis progresses through exploratory visualization, rigorous feature screening via hypothesis testing, feature engineering through bootstrap methods, and unsupervised clustering to validate discriminative capacity.

Throughout this section, we refer to speech recognition in quiet as SRT$_Q$ and speech recognition in noise as SRT$_N$. This notation maintains consistency with existing labels in tables and plots.

4.1 Audiological Measurement Patterns Across Severity Categories

Figure 2 presents violin plots for audiological measurements across hearing loss categories, revealing both progressive patterns and substantial overlap that motivate our feature engineering approach. Pure-tone thresholds show clear progression across severity categories (slight to severe), yet with considerable overlap between adjacent groups. Variability increases notably in speech recognition tasks as hearing loss becomes more severe, potentially influenced by ceiling effects in adaptive testing for the most impaired cases.

This increased heterogeneity in performance patterns indicates that simple threshold-based discrimination between categories using raw measurements alone will be insufficient. The complex distributions cannot be adequately characterized by single descriptive statistics (mean, median, variance). This limitation reflects a fundamental characteristic of PTA-based categorization: categories are defined using audiogram thresholds rather than incorporating speech test information, creating inherent challenges for multivariate discrimination.

These observations establish the empirical foundation for our feature engineering approach. The overlapping distributions in raw measurement space demonstrate why sophisticated statistical transformations—capturing contrasts, dependencies, and higher-order relationships—are necessary to achieve effective discrimination between severity levels.

4.2 Evaluating Separability in High-Dimensional Feature Space

To assess whether raw audiological data naturally separate into severity categories, we employ t-Distributed Stochastic Neighbor Embedding (t-SNE) [40], mapping data into an optimal 2-dimensional representation (Figure 3). This dimensionality reduction technique reveals the inherent structure of measurement relationships without imposing predetermined classifications.

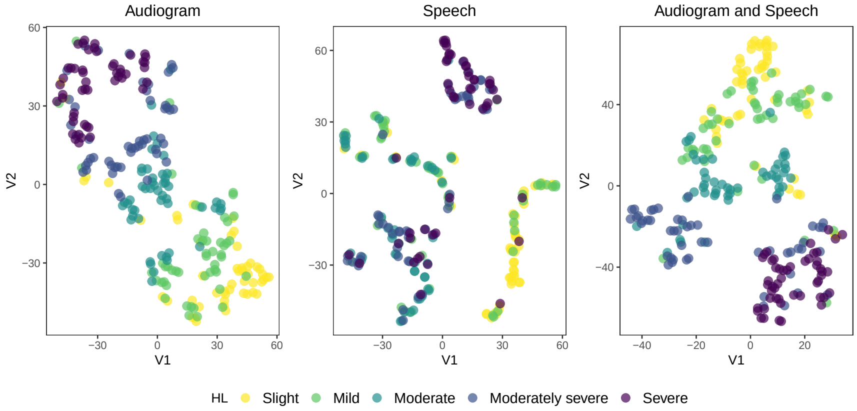

Figure 3: t-SNE visualization of raw audiological data demonstrating limited natural separation between severity categories. Left: Audiogram data alone shows substantial overlap between adjacent categories. Middle: Speech recognition scores (SRT_Q and SRT_N combined) reveal slightly improved differentiation, particularly for severe cases. Right: Combined audiogram and speech data illustrates enhanced but incomplete clustering. Color gradations represent severity levels from Slight to Severe. The x and y axes represent the first and second t-SNE dimensions.

The analysis reveals limited natural separation using raw measurements. Audiogram data alone (left panel) exhibits visible clustering but with substantial overlap between adjacent categories, particularly mild-moderate groups—an expected finding given that PTA categories impose discrete thresholds on continuous scales. Speech recognition scores (middle panel) show clearer separation for severe cases but significant overlap persists. Notably, the speech panel lacks the progressive ordering visible in audiogram data, suggesting these measures capture complementary discriminative information beyond pure-tone thresholds.

Combined analysis (right panel) indicates that integrating measurement types improves category separation but does not achieve complete discrimination without further transformation. This stems from the fundamental tension between continuous hearing function and discrete clinical categories. Raw audiological measurements, even when optimally projected, cannot effectively discriminate between severity levels without sophisticated feature engineering.

This motivates our statistical framework: transforming raw measurements into feature spaces that explicitly capture discriminative contrasts between categories, which we explore in subsequent sections.

4.3 Quantifying Discriminative Power Through Statistical Testing

We systematically evaluate discriminative power by applying the statistical test battery (Section 2.2) to all audiological measurements across severity category pairs. This identifies which measurements and measurement properties show significant differences between categories, providing principled feature selection.

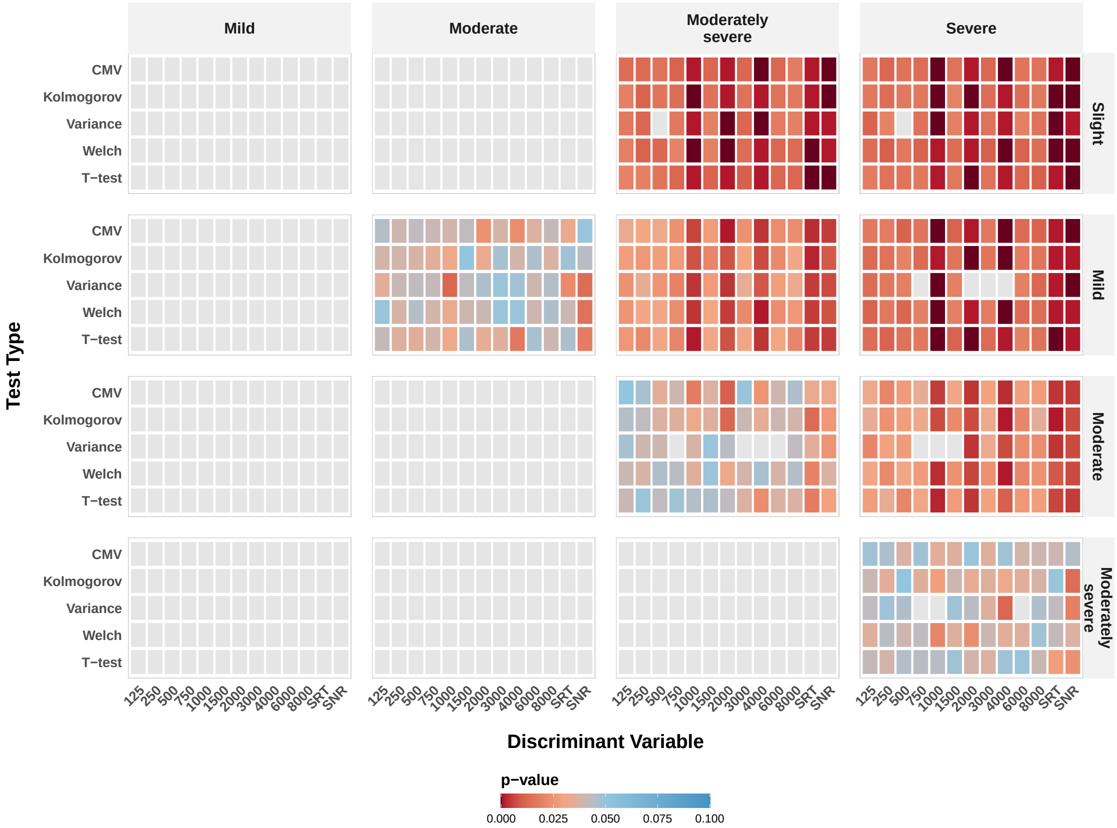

Figure 4: Heatmap of statistical significance (p-values) across severity categories, test types, and audiometric frequencies. Y-axis shows five statistical tests (T-test, Welch, Variance, Kolmogorov, CvM) applied to pure-tone thresholds (125-8000 Hz) and speech measures (SRT_Q, SRT_N) on x-axis. Color intensity indicates significance level: darker red represents stronger significance ($p < 0.001$), grey/blue indicates weaker discrimination. Adjacent categories show limited discriminative power, while non-adjacent categories demonstrate robust statistical separation, particularly in speech-critical frequencies (1000-4000 Hz).

Figure 4 reveals distinct patterns in univariate discriminative power. Discriminative strength increases substantially with severity gap between categories. For adjacent categories (e.g., Slight-Mild), the heatmap shows predominantly grey cells across test types and frequencies, indicating limited discriminative power of individual measurements. This fundamental limitation suggests univariate measures alone cannot distinguish adjacent severity levels, necessitating multivariate feature combinations.

In contrast, non-adjacent categories (e.g., Slight vs Severe) display intense coloring (azure to red) across multiple frequencies and tests, indicating robust discrimination. This pattern concentrates in the speech-critical frequency range (1000-4000 Hz), where measurements consistently achieve $p < 0.001$ across different statistical methodologies.

Analysis of discriminative power across frequencies (Figure 5) reveals a clear hierarchy. Speech recognition tests (SRT_Q, SRT_N) exhibit highest discriminative power, achieving significance in 28-29 comparisons across statistical tests. Mid-frequency pure-tone thresholds (2000-4000 Hz) follow closely with 27-28 significant results. Lower frequencies (125-750 Hz) show comparatively weaker discrimination (13-15 significant tests), suggesting a natural weighting scheme where speech measures and mid-frequency thresholds contribute most to severity differentiation.

However, discrimination challenges intensify for adjacent category pairs, as evident in grey regions of Figure 4 where significance is limited across all test types. This pattern highlights key limitations of univariate approaches: individual measurements alone prove insufficient for distinguishing adjacent severity levels, motivating multivariate analysis.

Multivariate tests reveal that higher-frequency thresholds provide strongest discrimination, while Bartlett tests highlight variance differences in speech-critical ranges, underscoring dispersion's role in classification. Copula test results (detailed in Supplementary Appendix) show strongest discriminatory power emerges from interactions between speech recognition scores and pure-tone thresholds in 1000-4000 Hz range.

Copula analysis reveals hierarchical discriminative patterns across category comparisons. Non-adjacent comparisons (Slight vs Severe, Slight vs Moderately Severe) achieve exceptionally strong discrimination ($p < 0.001$) across multiple frequency pairs. Cross-frequency combinations (125Hz|4000Hz, 500Hz|2000Hz) show significant effects for milder contrasts, while higher frequency pairs (2000Hz|8000Hz) become important for severe cases. Speech measure combinations ($\text{SRT}_Q \mid \text{SRT}_N$) and their pairings with frequency thresholds (2000Hz|SRT, 3000Hz|SRT) consistently achieve significant discrimination ($p < 0.05$ to $p < 0.01$), emphasizing the value of integrating multiple measurement types.

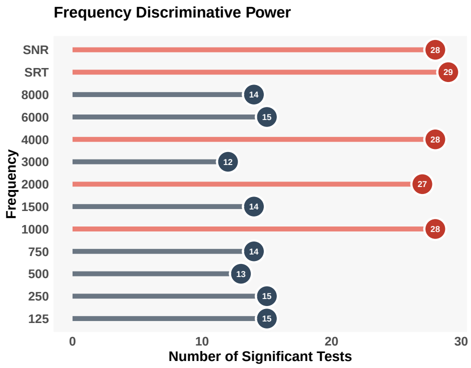

Figure 5: Discriminative power of audiometric frequencies and speech measures based on number of significant statistical tests. Bar colors distinguish high-discriminative (red) and low-discriminative (gray) measures.

Adjacent category comparisons exhibit weaker but still significant discrimination. Mild vs Moderate comparisons show strongest effects in 250Hz|4000Hz and 500Hz|2000Hz combinations ($p \approx 0.01$). Moderate vs Moderately Severe comparisons demonstrate intermediate performance ($p \approx 0.013$-0.025) with high-frequency pairs (2000Hz|8000Hz) and speech-frequency interactions (1000Hz|SNR) providing highest discrimination.

This multivariate analysis establishes that while individual measurements struggle with adjacent categories, specific frequency combinations and speech-score pairings capture subtle variations in hearing loss progression. Strong interactions in speech-critical frequencies indicate multivariate feature embeddings offer superior classification robustness compared to univariate measures, particularly for borderline cases.

Based on comprehensive testing, we identify optimal feature combinations guided by three criteria: statistical robustness across methodologies, discriminative power across severity levels, and clinical relevance aligned with audiological understanding. Table 5 presents top-ranked features for each severity contrast, revealing several key patterns.

Table 5: Top five discriminative features for each severity category comparison based on statistical significance. Columns show: discriminating attribute, test type, significance level α, and test category (univariate/multivariate).

Feature Ranking by Severity Contrast									
#	**Slight vs. Mild**				**#**	**Mild vs. Severe**			
1	125Hz|4000Hz	Copula	< 0.05	Multi.	1	SNR	T-test	< 0.001	Uni.
2	SRT|SNR	Copula	< 0.05	Multi.	2	SNR	Variance	< 0.001	Uni.
3	500Hz|1000Hz	Copula	< 0.05	Multi.	3	1000Hz	Kolmogorov	< 0.001	Uni.
4	1500Hz|4000Hz	Copula	< 0.05	Multi.	4	2000Hz	Kolmogorov	< 0.001	Uni.
5	6000Hz|SNR	Copula	< 0.05	Multi.	5	4000Hz	CMV	< 0.001	Uni.
#	**Slight vs. Moderate**				**#**	**Moderate vs. Moderately Severe**			
1	2000Hz|SRT	Copula	< 0.01	Multi.	1	4000Hz	T-test	< 0.01	Uni.
2	250Hz|2000Hz	Copula	< 0.01	Multi.	2	250Hz|1000Hz	Copula	< 0.01	Multi.
3	SRT|SNR	Copula	< 0.01	Multi.	3	1000Hz	Variance	< 0.01	Uni.
4	750Hz|2000Hz	Copula	< 0.01	Multi.	4	2000Hz	Variance	< 0.01	Uni.
5	750Hz|SRT	Copula	< 0.01	Multi.	5	1000Hz	Kolmogorov	< 0.01	Uni.

Speech-critical frequencies (1000-4000 Hz) consistently emerge as dominant discriminators. These frequencies exhibit strongest discriminative power across statistical tests, particularly in severe impairment comparisons ($p < 0.001$). Speech recognition measures (SRT_Q, SRT_N) provide essential complementary information, especially in moderate-to-severe cases where they significantly enhance discrimination accuracy. Their value increases when paired with frequency-based features, underscoring the importance of integrating different audiological metrics.

Distinguishing adjacent categories (Slight-Mild, Mild-Moderate) requires multivariate approaches, as univariate tests often fail to achieve significance. Copula test results demonstrate that feature pairs—including cross-frequency combinations and speech-frequency interactions—offer superior discriminative power compared to individual features. These findings indicate feature interactions capture nuanced severity distinctions, particularly for borderline cases where univariate measures struggle.

Statistical significance strengthens with increasing severity contrast, with strongest discriminative features emerging in non-adjacent category comparisons. Mid-to-high frequencies (2000-4000 Hz) consistently show highest significance across all test types. Higher frequencies (4000 Hz and beyond) become increasingly important for moderate-to-severe distinctions, whereas lower frequencies (125-750 Hz) contribute more to early-stage differentiation.

Feature pairs like $\text{SRT}_Q \mid \text{SRT}_N$ and their interactions with frequency thresholds (particularly 1000-4000 Hz) demonstrate highest discriminative power, highlighting the importance of combining speech recognition with pure-tone thresholds. Variance-based methods (Bartlett, variance tests) reveal that dispersion in hearing thresholds also plays crucial roles, particularly in mid-to-high frequencies. This suggests variability differences—not just mean threshold shifts—are critical for characterizing severity.

Interaction between speech recognition scores and pure-tone thresholds emerges as key factor in defining severity. Feature pairings such as 1000Hz|SNR and 2000Hz|SRT consistently achieve high significance, demonstrating that integrating speech-based measures with audiometric data provides more robust classification framework.

These findings establish a foundation for feature engineering: the identified discriminative patterns guide transformation of raw measurements into feature spaces that capture statistical contrasts most effectively distinguishing between severity categories.

4.4 Feature Engineering and Discriminative Power Enhancement

Our feature engineering approach combines statistical bootstrapping with systematic dimensionality analysis to create robust discriminative features. This addresses two challenges: class imbalance across severity categories and insufficient sample sizes for reliable discriminative power assessment.

4.4.1 Bootstrap-Based Feature Construction

We employ both parametric and non-parametric bootstrapping (Section 2.3) to generate simulation-based replicates, ensuring robustness against distributional assumptions. The dual approach validates that results remain consistent across different feature simulation methods.

Figure 6 demonstrates effectiveness through t-SNE visualization of engineered features. Compared to raw data (Figure 3), engineered features exhibit markedly improved separation between severity levels. Copula-based measures show particularly strong clustering, while univariate statistics—including means and variances for frequency and speech data—display enhanced separation. These visualizations confirm that feature engineering successfully captures underlying patterns in hearing loss progression.

4.4.2 Feature Space Dimensionality Analysis

Feature space dimensionality varies substantially based on feature type and screening. For individual features, screening based on statistical significance ($p < 0.05$) substantially reduces dimensionality while retaining discriminative power. Frequency univariate features reduce from 11 to 3 dimensions, focusing on most significant frequencies (1000Hz, 2000Hz, 4000Hz). Speech features maintain both dimensions (SRT_Q, SRT_N) as both prove significant.

Copula-based features represent the highest-dimensional category, with screened versions maintaining 81-99 dimensions. These capture complex dependencies between measurements, providing rich discriminative information. Combined feature sets demonstrate trade-offs between complexity and information content. Univariate combinations maintain relatively low dimensionality (8-39) while integrating multiple statistical measures. Full cross-feature combinations incorporating both univariate and copula measures span higher dimensions (192-741). The screening process plays crucial roles in dimensionality reduction while preserving discriminative power. This hierarchical organization enables flexible adaptation to different evaluation scenarios while maintaining interpretability.

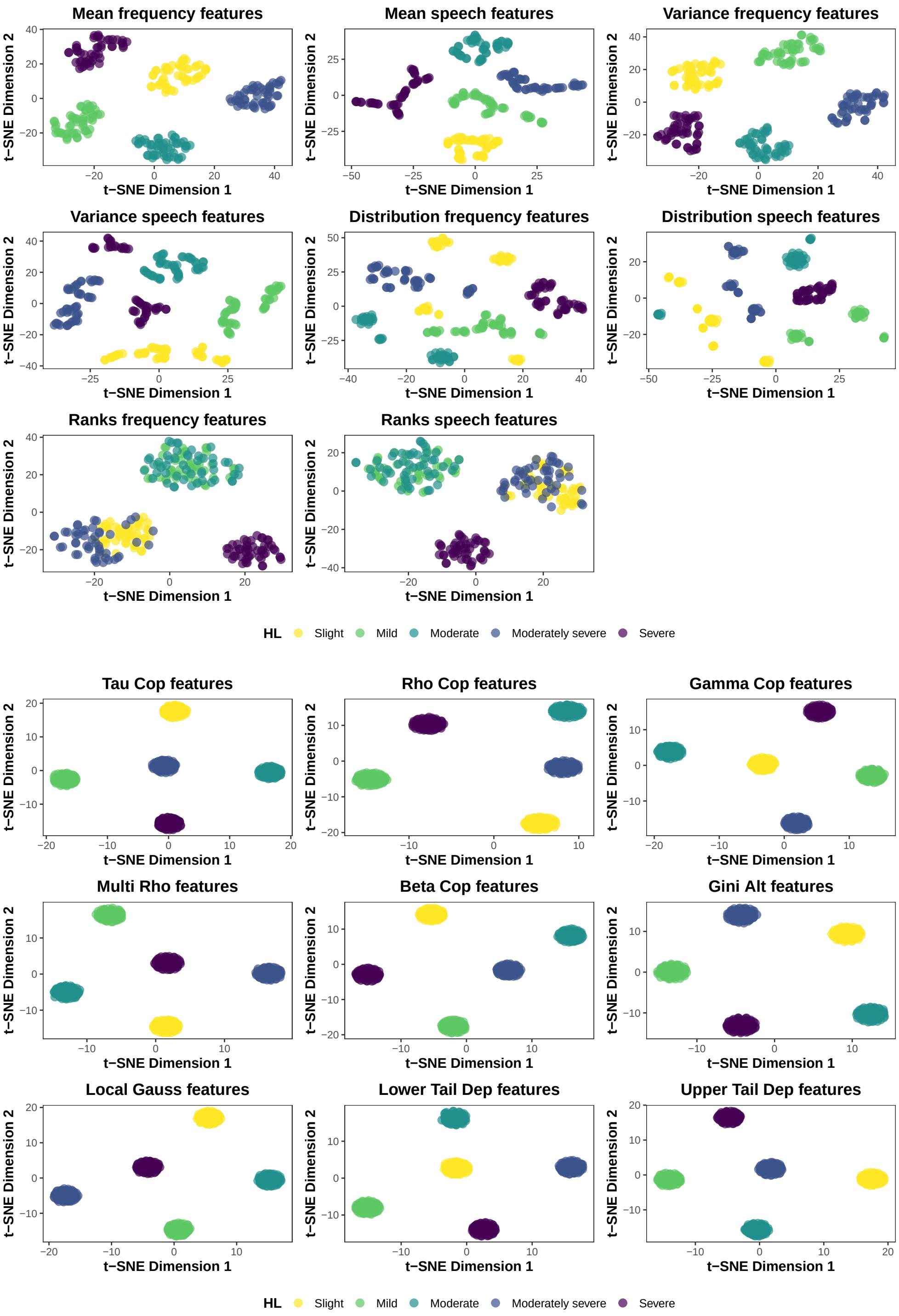

Figure 6: t-SNE visualization of engineered features demonstrating enhanced discrimination. Top: Univariate statistical representations showing improved clustering potential compared to raw data. Bottom: Copula-based measures capturing complex dependencies with superior category separation. Color gradient from yellow (Slight) to purple (Severe) tracks hearing loss progression. Both use parametric bootstrapping (n=50). X and Y axes represent first two t-SNE dimensions.

4.5 Unsupervised Clustering Validates Discriminative Power

To objectively evaluate discriminative power of engineered features, we employ unsupervised clustering methods (K-Means and Hierarchical Clustering with Ward's method), configured to identify five clusters corresponding to severity categories. Performance is assessed using multiple complementary metrics. The Silhouette score [37] quantifies cluster cohesion and separation on a scale from -1 to 1, with scores exceeding 0.5 indicating well-separated clusters and 1.0 representing perfect separation. The Adjusted Rand Index (ARI) measures agreement between predicted and true cluster assignments, corrected for chance, ranging from -1 to 1 with values near 1 indicating strong agreement. Normalized Mutual Information (NMI) quantifies shared information between clusterings on a scale from 0 to 1, with higher values indicating greater similarity. The Calinski-Harabasz Index (CH-Index) evaluates cluster separation through the ratio of between-cluster to within-cluster variance, with higher values indicating better-defined clusters. Stability is measured through bootstrap resampling consistency, ranging from 0 to 1. More details are provided in the Supplementary Appendix.

Analysis reveals strong relationship between sample size and clustering performance, particularly for parametric bootstrap features. As sample size increases from n=50 to n=5000, we observe consistent improvement in Silhouette scores across both methods, with most pronounced gains in mean and variance-based features. Improvement plateaus around n=1000, where performance stabilizes. Speech copula features achieve highest scores (exceeding 0.7) while traditional feature combinations reach moderate scores (0.6-0.66). Comprehensive clustering results across all feature types, sample sizes, and bootstrap methods are provided in Supplementary Tables S6-S8.

Non-parametric bootstrapping demonstrates consistently superior performance compared to parametric methods, achieving Silhouette scores approximately 0.03 higher across all sample sizes and feature types. This advantage likely stems from inherent flexibility in handling complex audiological data distributions.

Individual feature analysis (Supplementary Table S8) reveals varying discriminative power across pure tone thresholds. Thresholds at 2000Hz and 4000Hz emerge as particularly strong discriminators for mean-based features, while speech recognition scores and 1000Hz thresholds show dominance in variance-based discrimination. This aligns with clinical understanding of speech-critical frequencies and their role in hearing loss assessment.

Feature combinations demonstrate complex behavior—simple combinations of two to three features often achieve optimal performance, while more complex feature sets can degrade clustering effectiveness. This suggests careful feature selection may be more valuable than comprehensive feature inclusion.

Comparative analysis reveals similar performance between clustering methods. While K-Means shows marginally higher scores for larger samples (n$\geq$1000), differences are minimal. Both methods demonstrate consistent results across feature types, suggesting either could be appropriate for discriminative power evaluation.

Feature screening emerges as crucial component, with screened features consistently achieving comparable or superior performance despite reduced dimensionality. Optimal feature sets demonstrate clear hierarchy: speech copula features, particularly Upper Tail Dependence features, achieve Silhouette scores of 0.76-0.94 for samples $\geq$1000 with non-parametric bootstrapping. Traditional audiometric measures—combinations of two to three frequency-specific thresholds (1000Hz, 2000Hz, 4000Hz) with speech recognition measures—show moderate performance with Silhouette scores of 0.60-0.66 when optimally combined.

Table 6 presents performance of combined feature sets, demonstrating that speech-only combinations consistently outperform frequency-only combinations. Complete feature combinations (frequency + speech + copula, dimension 534) show lower performance due to curse of dimensionality—where excessive features introduce noise rather than signal—reinforcing the value of feature screening.

Based on comprehensive analysis, the most robust configuration emerges from K-Means clustering with n $\geq$ 1000 samples using non-parametric bootstrapping. Three distinct high-performing approaches emerge: (1) screened speech copula features, particularly Upper Tail Dependence measures (Silhouette 0.94), (2) screened combinations of speech mean, variance, and distribution features (Silhouette 0.88), and (3) traditional feature sets combining frequency thresholds with speech recognition scores (Silhouette 0.73).

Table 7 presents multi-metric evaluation validating these findings. Speech copula features achieve high ARI (0.91) and NMI (0.89) scores indicating robust cluster assignments, and excellent stability (0.88) suggesting reliable reproducibility.

Table 6: Clustering performance (Silhouette scores) for combined feature sets across sample sizes and bootstrap methods. Comprehensive results for all feature types provided in Supplementary Tables S6-S8. Par.: Parametric; NonPar.: Non-Parametric.

Clustering Results - Combined Features Analysis																	
		K-Means								HCW							
		n = 50		n = 500		n = 1000		n = 5000		n = 50		n = 500		n = 1000		n = 5000	
Dim	Features	Par.	NonPar.	Par.	NonPar.	Par.	NonPar.	Par.	NonPar.	Par.	NonPar.	Par.	NonPar.	Par.	NonPar.	Par.	NonPar.
Frequency-Only Combinations																	
33	All Univariate	0.35	0.38	0.40	0.43	0.42	0.45	0.44	0.47	0.33	0.36	0.38	0.41	0.40	0.43	0.42	0.45
495	All Copula	0.25	0.28	0.30	0.33	0.32	0.35	0.34	0.37	0.23	0.26	0.28	0.31	0.30	0.33	0.32	0.35
Speech-Only Combinations																	
3	All SRT	0.45	0.48	0.50	0.53	0.52	0.55	0.54	0.57	0.43	0.46	0.48	0.51	0.50	0.53	0.52	0.55
3	All SNR	0.47	0.50	0.52	0.55	0.54	0.57	0.56	0.59	0.45	0.48	0.50	0.53	0.52	0.55	0.54	0.57
6	SRT + SNR	0.50	0.53	0.55	0.58	0.57	0.60	0.59	0.62	0.48	0.51	0.53	0.56	0.55	0.58	0.57	0.60
Complete Combinations																	
39	Frequency + Speech	0.40	0.43	0.45	0.48	0.47	0.50	0.49	0.52	0.38	0.41	0.43	0.46	0.45	0.48	0.47	0.50
534	All Features	0.15	0.18	0.20	0.23	0.22	0.25	0.24	0.27	0.13	0.16	0.18	0.21	0.20	0.23	0.22	0.25

Table 7: Multi-metric evaluation of best performing feature sets. ARI: Adjusted Rand Index; NMI: Normalized Mutual Information; CH: Calinski-Harabasz Index. Stability measured through bootstrap resampling consistency.

Feature Set	Silhouette	ARI	NMI	CH-Index	Stability
Individual Features					
Best Frequency Univariate (2000Hz Mean)	0.50	0.47	0.45	145.2	0.82
Best Speech Univariate (Mean screened)	0.75	0.72	0.70	187.9	0.85
Best Single Copula (Upper Tail Dep. screened)	0.94	0.91	0.89	235.6	0.88
Feature Combinations					
Best Frequency + Speech	0.73	0.70	0.68	198.4	0.86
Speech Mean & Var & Distr screened	0.88	0.85	0.83	215.7	0.87
Best Overall (Speech Copula Multi-rho screened)	0.91	0.88	0.86	228.3	0.89

Screened feature sets demonstrate consistently higher performance across all metrics (CH-Index improvements $>40\%$), with speech-based copula measures substantially outperforming traditional frequency-based approaches. Traditional audiometric measures provide adequate discrimination (Silhouette ≈ 0.50), while incorporating sophisticated speech-based features dramatically improves discriminative capacity.

5 Discussion and Conclusion

This study developed and validated a rigorous statistical framework for quantifying discriminative value of audiological measurements across hearing loss severity categories. By systematically evaluating which features—from raw measurements to advanced statistical transformations—most effectively distinguish established PTA-based categories, we provide an evidence-based foundation for refining hearing loss classification and optimizing clinical assessment.

5.1 Which Measurements Best Discriminate Hearing Loss Severity?

To address this question, our approach combines three complementary strategies ensuring valid quantification of discriminative value. Rather than relying on a single statistical test, we applied comprehensive hypothesis testing across multiple frameworks—univariate tests (t-tests, variance tests, distribution tests), multivariate tests (Bartlett, Tukey HSD), and copula-based dependency tests—to each measurement pair across all severity categories. The consistency of discrimination hierarchies across these diverse methodologies (Figure 4) indicates identified features show robust, reproducible differences rather than test-specific artifacts.

Copula-based feature engineering addresses fundamental limitations of conventional approaches. Standard correlation methods assume linear relationships and may fail to detect tail dependencies or nonlinear patterns characteristic of audiological data, where threshold-suprathreshold relationships are known to be complex [12, 25]. Copula methods separate dependence structure from marginal distributions, enabling detection of whether patients with similar average thresholds but different speech-threshold dependency patterns belong to functionally distinct groups. These methods nearly doubled discriminative performance relative to univariate measures (Silhouette 0.94 vs. 0.50), demonstrating their capacity to capture information inaccessible to conventional approaches.

Unsupervised validation objectively confirms discriminative capacity. We applied clustering algorithms without providing severity category labels, then evaluated whether discovered clusters aligned with established PTA categories. High clustering performance (Silhouette scores 0.76-0.94) with strong agreement to known categories (ARI 0.85-0.91, NMI 0.83-0.89, Table 7) confirms that engineered features capture genuine severity differences. This triangulation approach—combining multiple hypothesis testing frameworks, sophisticated dependency modeling, and unsupervised validation—ensures the discriminative hierarchies we identify reflect real audiological patterns rather than analytical choices.

A critical methodological choice warrants justification: we validate features against established PTA-based categories rather than discovering novel label phenotypes. This design serves our research question—quantifying which measurements provide discriminative information within existing clinical frameworks. Our goal is identifying which additional measurements enhance discrimination between established categories, thereby informing evidence-based extension of current systems rather than proposing alternative classifications. By using PTA categories as reference, we provide actionable guidance for clinicians while demonstrating empirical justification for system extension.

5.1.1 Discriminative Hierarchies Across Measurements

Clear hierarchies emerged across measurement types through systematic testing of 48,144 adults. Among pure-tone measures, mid-frequency thresholds (1000-4000 Hz) consistently achieved strongest discrimination (27-28 significant comparisons), while lower frequencies (125-750 Hz) showed weaker performance (13-15 comparisons) and highest frequencies (6000-8000 Hz) fell intermediate. This empirically derived hierarchy aligns with speech-critical frequencies [5] but emerged from systematic statistical testing rather than acoustic assumptions.

Speech recognition measures (SRT_Q, SRT_N) achieved 28-29 significant discriminations, matching or exceeding the strongest audiometric frequencies (Figure 5). Critically, speech-in-noise (SRT_N) demonstrated slightly higher discriminative power than speech-in-quiet (SRT_Q), confirming that speech-in-noise testing captures functional severity information particularly relevant to patient complaints [9, 41, 10].

Feature combinations revealed substantial advantages over individual measures. Speech-only combinations achieved Silhouette scores of 0.50-0.62, exceeding frequency-only combinations (0.35-0.47) despite fewer dimensions (Table 6). Speech copula features capturing dependencies between quiet and noise performance achieved the highest discriminative power observed (Silhouette 0.94, stability 0.88-0.89). Integrating speech with frequency-specific thresholds further improved performance (40% CH-Index gains), with optimal pairs like 1000 Hz|SNR and 2000 Hz|SRT showing consistently high significance (Table 5).

Adjacent category discrimination (slight-mild, mild-moderate) essentially failed with univariate approaches but succeeded with multivariate combinations. Patients within adjacent categories show overlapping threshold distributions but exhibit distinct patterns when speech measures are integrated with audiometric data. The finding that patients with similar PTAs show varying speech recognition abilities (Figure 2) reflects multidimensional hearing function where threshold detection captures only one aspect [11, 10, 41, 12, 24]. Speech tests provide orthogonal information about suprathreshold auditory processing not reflected in pure-tone thresholds alone.

5.2 Implications for Clinical Classification

These findings provide empirical foundation for extending classification systems beyond threshold-based categories. While PTA adequately discriminates widely separated severity levels, speech recognition measures capture complementary suprathreshold information that substantially improves severity characterization, reflecting functional dimensions that threshold-only classification misses [12, 25].

The WHO World Report on Hearing [4] acknowledges that speech understanding cannot be inferred from PTA alone, yet provides no quantitative framework for incorporating speech measures. Previous efforts recognized this need [18, 16, 25, 14, 17, 15], but our study quantifies how much added value speech tests provide, which measures contribute most (speech-in-noise combined with mid-frequency thresholds), and which statistical approaches best reveal their discriminative power (copula-based dependency features capturing tail dependencies and nonlinear relationships).

We propose maintaining PTA-based categories as primary framework while systematically incorporating speech recognition measures as complementary characterization. This approach enhances clinical classification without requiring modification of established categorical boundaries. Within any PTA category, speech testing reveals functional variation reflecting differences in suprathreshold processing [11, 10], providing information relevant to treatment planning and outcome prediction. The consistency of discrimination hierarchies across multiple statistical tests and bootstrap implementations indicates these relationships are robust and reproducible.

For clinical assessment protocols, these findings yield specific, evidence-based guidance. When time is limited, prioritize: (1) mid-frequency audiometry (1000-4000 Hz), (2) speech-in-noise testing, and (3) speech-in-quiet testing. This ordering reflects both the speech-critical nature of mid-frequencies and the substantial discriminative value of suprathreshold measures [9, 25].

Comprehensive protocols should evaluate speech recognition in both quiet and noise conditions, as these capture distinct aspects of auditory function [9, 10]. Integration of speech and audiometric data through multivariate combinations reveals dependency patterns invisible to univariate analysis. This enables identification of patients with suprathreshold processing deficits despite comparable thresholds [24]. Consideration of measurement variability also contributes discriminative information beyond mean values.

The substantial improvement achieved through screened feature sets (CH-Index gains >40% with 28% dimensionality reduction) indicates systematic measurement selection maintains discriminative power while improving efficiency—particularly relevant for resource-limited settings where assessment burden must be minimized. Our results specify which measures contribute most and quantify their added value, transforming clinical intuition into evidence-based practice [5, 6, 12, 25].

5.3 Methodological Contributions Beyond Audiology

While motivated by audiological questions, this work contributes statistical methodology applicable to any medical classification context where the goal is quantifying which measurements best discriminate established categories.

Unsupervised validation provides rigorous assessment independent of supervised learning biases. By applying clustering without providing labels, we objectively tested whether engineered features naturally separate groups, with high agreement to known categories (ARI 0.85-0.91, NMI 0.83-0.89) confirming features capture genuine differences. Copula-based feature engineering demonstrates that these methods reveal information inaccessible to conventional approaches—our demonstration that copulas effectively capture tail dependencies and nonlinear patterns in complex, non-normal distributions provides template for other diagnostic contexts where extreme values carry clinical significance. While copulas have been applied in medical prediction [42, 43], their use for diagnostic feature engineering remains limited.

Multiple complementary validation metrics (Silhouette, ARI, NMI, CH-Index, Stability) combined with nonparametric bootstrapping provide robust performance assessment while addressing class imbalance (severe n=389 vs. moderate n=20,246), offering template for validation in medical machine learning contexts where certain diagnostic categories are naturally rare.

5.4 Limitations and Future Directions

This study analyzed adults aged 40-90 years with symmetric hearing loss (PTA difference <15 dB) from a single database. While the sample size (N=48,144) provides robust statistical power, validation in independent cohorts from different healthcare systems and geographic regions would strengthen generalizability claims. Cross-database studies could assess whether discrimination hierarchies remain consistent across populations with different demographic characteristics and testing protocols.

We quantify discrimination between PTA-based categories rather than independent functional outcomes. Future work should evaluate discriminative power relative to patient-reported outcomes (e.g., Hearing Handicap Inventory), self-reported communication abilities, or hearing aid benefit measures to assess whether measurements that best discriminate PTA categories also capture dimensions most relevant to patients.

The small severe category (n=389, 0.8%) limits discriminative power assessment for this group despite bootstrap methods partially addressing imbalance. Results for severe hearing loss should be interpreted cautiously. Additionally, restriction to symmetric cases limits generalizability to asymmetric presentations, which may exhibit different discrimination patterns.

Future work should evaluate whether highly discriminative measurements also predict intervention outcomes such as hearing aid benefit and patient satisfaction. If speech-in-noise measures that distinguish severity categories also predict treatment response, this would strengthen the case for their integration into clinical planning. Extending the framework to longitudinal contexts could evaluate which measurements best discriminate progression rates, adapting the approach to characterize change patterns rather than static severity levels and informing monitoring protocols and intervention timing.

Integration with auditory neuroscience research could elucidate physiological mechanisms underlying observed discrimination patterns [25]. Application to other populations (asymmetric hearing loss, specific etiologies, pediatric cases, cochlear implant candidates) could establish whether the discriminative hierarchy generalizes across clinical presentations. The finding that speech measures provide information orthogonal to pure-tone thresholds may have particular relevance for populations with discordant threshold-function relationships, such as auditory neuropathy spectrum disorder.

5.5 Conclusions

This study provides the first systematic quantification of discriminative value across audiological measurements, validated on a large clinical database (N=48,144). Speech-in-noise testing combined with mid-frequency thresholds provides substantially greater discriminative power than threshold-based measures alone, multivariate feature combinations capturing complex dependencies reveal functional information invisible to univariate analysis, and systematic feature screening enables efficient assessment protocols prioritizing highest-value measurements.

These findings support extending hearing loss classification by maintaining PTA-based categories as primary framework while systematically incorporating speech recognition measures as complementary suprathreshold characterization. The statistical framework developed here—combining hypothesis testing, copula-based dependency analysis, and unsupervised validation—offers generalizable methodology for discriminative value quantification applicable beyond audiology.

By establishing the added value of speech recognition testing with quantitative evidence, this work provides empirical foundation for evidence-based refinement of audiological classification systems—moving toward more comprehensive, functionally-relevant characterization that better serves clinical decision-making and patient care.

Author's Contribution

M.C.: Conceptualization, data curation, formal analysis, investigation, methodology, software, original draft preparation, writing—review and editing.
G.W.P.: Conceptualization, formal analysis, investigation, methodology, supervision, original draft preparation, writing—review and editing.
P.M.: resources, data curation, review and editing.
M.B.: formal analysis, investigation, original draft preparation, writing—review and editing.
H.T-V.:supervision, funding, review and editing.
M.C., P.M., M.B., H.T-V. confirm that they had full access to all the data in the study. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by a grant from Fondation Pour l'Audition (FPA) to the Institut de l'Audition (FPAIDA09 to Hung Thai-Van and FPAIDA10 to Paul Avan and the Ceriah), a French government grant managed by the Agence Nationale de la Recherche under the France 2030 program, reference ANR-23-IAHU-0003 and by Amplifon. Furhter, the authors thank the Clinical Research Coordination Office (CRCO) of the Institut Pasteur for their help with biomedical regulatory and ethical aspects of the project. MB was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project ID 496819293.

Data Statement

The data provider is Amplifon France.

Conflicts of Interest

The authors have no financial relationships relevant to this article to disclose.

References

[1] Blake S Wilson, Debara L Tucci, Michael H Merson, and Gerard M O'Donoghue. Global hearing health care: new findings and perspectives. *The Lancet*, 390(10111):2503–2515, 2017.

[2] Lesley M Haile, Kaloyan Kamenov, Paul S Briant, et al. Hearing loss prevalence and years lived with disability, 1990-2019: findings from the global burden of disease study 2019. *The Lancet*, 397(10278):996–1009, 2021.

[3] John G Clark. Uses and abuses of hearing loss classification. *Asha*, 23(7):493–500, 1981.

[4] World Health Organization. *World report on hearing*. World Health Organization, 2021.

[5] Jack Katz, Larry Medwetsky, Robert F Burkard, and Linda J Hood. *Handbook of clinical audiology*. Wolters Kluwer, Lippincott William & Wilkins Philadelphia, 2009.

[6] Steven Kramer and David K Brown. *Audiology: science to practice*. Plural Publishing, 2021.

[7] Douglas G Altman, Berthold Lausen, Willi Sauerbrei, and Martin Schumacher. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute*, 86(11):829–835, 1994.

[8] Elizabeth L Turner, Jessica E Dobson, and Stuart J Pocock. Problems of categorizing continuous variables in clinical prediction models. *BMC Medical Research Methodology*, 21(1):1–11, 2021.

[9] Mead C Killion and Patricia A Niquette. What can the pure-tone audiogram tell us about a patient's snr loss?. *The Hearing Journal*, 53(3):46–48, 2000.

[10] David R Moore, Mark Edmondson-Jones, Piers Dawes, Heather Fortnum, Abby McCormack, Robert H Pierzycki, and Kevin J Munro. Relation between speech-in-noise threshold, hearing loss and cognition from 40–69 years of age. *PloS one*, 9(9):e107720, 2014.

[11] Andrew J Vermiglio, Sigfrid D Soli, Daniel J Freed, and Laurel M Fisher. The relationship between high-frequency pure-tone hearing loss, hearing in noise test (hint) thresholds, and the articulation index. *Journal of the American Academy of Audiology*, 23(10):779–788, 2012.

[12] Reinier Plomp. Auditory handicap of hearing impairment and the limited benefit of hearing aids. *The Journal of the Acoustical society of America*, 63(2):533–549, 1978.

[13] Harold F Schuknecht and Mark R Gacek. Cochlear pathology in presbycusis. *Annals of Otology, Rhinology & Laryngology*, 102(1_suppl):1–16, 1993.

[14] Judy R Dubno, Mark A Eckert, Fu-Shing Lee, Lois J Matthews, and Richard A Schmiedt. Classifying human audiometric phenotypes of age-related hearing loss from animal models. *Journal of the Association for Research in Otolaryngology*, 14:687–701, 2013.

[15] Raul Sanchez Lopez, Federica Bianchi, Michal Fereczkowski, Sebastien Santurette, and Torsten Dau. Data-driven approach for auditory profiling and characterization of individual hearing loss. *Trends in hearing*, 22:2331216518807400, 2018.

[16] Raul Sanchez-Lopez, Michal Fereczkowski, Tobias Neher, Sébastien Santurette, and Torsten Dau. Robust data-driven auditory profiling towards precision audiology. *Trends in hearing*, 24:2331216520973539, 2020.

[17] Aravindakshan Parthasarathy, Sandra Romero Pinto, Rebecca M Lewis, William Goedicke, and Daniel B Polley. Data-driven segmentation of audiometric phenotypes across a large clinical cohort. *Scientific reports*, 10(1):6704, 2020.

[18] Mareike Buhl, Anna Warzybok, Marc René Schädler, Thomas Lenarz, Omid Majdani, and Birger Kollmeier. Common audiological functional parameters (cafpas): Statistical and compact representation of rehabilitative audiological classification based on expert knowledge. *International journal of audiology*, 58(4):231–245, 2019.

[19] Mareike Buhl, Anna Warzybok, Marc René Schädler, Omid Majdani, and Birger Kollmeier. Common audiological functional parameters (cafpas) for single patient cases: Deriving statistical models from an expert-labelled data set. *International Journal of Audiology*, 59(7):534–547, 2020.

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

[21] Samira Saak, David Huelsmeier, Birger Kollmeier, and Mareike Buhl. A flexible data-driven audiological patient stratification method for deriving auditory profiles. *Frontiers in Neurology*, 13:959582, 2022.

[22] R Badri et al. Auditory filter shapes and high-frequency hearing in adults who have impaired speech in noise performance despite clinically normal audiograms. *J. Acoust. Soc. Am.*, 129:852–863, 2011.

[23] D Moore et al. Benefits of extended high-frequency audiometry for everyone. *The Hearing Journal*, 70:50–52, 2017.

[24] Van Summers, Matthew J Makashay, Sarah M Theodoroff, and Marjorie R Leek. Suprathreshold auditory processing and speech perception in noise: Hearing-impaired and normal-hearing listeners. *Journal of the American Academy of Audiology*, 24(04):274–292, 2013.

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

[26] Bernard L Welch. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[27] David B Duncan. Multiple range and multiple f tests. *biometrics*, 11(1):1–42, 1955.

[28] George EP Box. Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335, 1953.

[29] Indra Mohan Chakravarti, Radha G Laha, and Jogabrata Roy. Handbook of methods of applied statistics. *Wiley Series in Probability and Mathematical Statistics (USA) eng*, 1967.

[30] Marcelo G Cruz, Gareth W Peters, and Pavel V Shevchenko. *Fundamental aspects of operational risk and insurance analytics: A handbook of operational risk*. John Wiley & Sons, 2015.

[31] Tony Cai, Weidong Liu, and Yin Xia. Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501):265–277, 2013.

[32] John W Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114, 1949.

[33] Bruno Rémillard and Olivier Scaillet. Testing for equality between two copulas. *Journal of Multivariate Analysis*, 100(3):377–386, 2009.

[34] Jun Yan. Multivariate modeling with copulas and engineering applications. *Springer handbook of engineering statistics*, pages 931–945, 2023.

[35] Antonio Dalessandro and Gareth W Peters. Efficient and accurate evaluation methods for concordance measures via functional tensor characterizations of copulas. *Methodology and Computing in Applied Probability*, 22:1089–1124, 2020.

[36] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979.

[37] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[38] Position statement: Red flags-warning of ear disease - american academy of otolaryngology-head and neck surgery (aao-hns), 2024. Accessed: 2024-08-11.

[39] Article 8 - loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, 2024. Accessed: 2024-08-11.

[40] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

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

[42] Pranesh Kumar and Mohamed M Shoukri. Copula based prediction models: an application to an aortic regurgitation study. *BMC medical research methodology*, 7:1–9, 2007.

[43] Meng Hu, Kelsey L Clark, Xiajing Gong, Behrad Noudoost, Mingyao Li, Tirin Moore, and Hualou Liang. Copula regression analysis of simultaneously recorded frontal eye field and inferotemporal spiking activity during object-based working memory. *Journal of Neuroscience*, 35(23):8745–8757, 2015.