Reference data for a quick speech-in-noise hearing test in the French language

Julie Bestel, Daniel Pressnitzer, Mathieu Robier, F. Rembaud, Christian Renard, François Leclercq, Christophe Vincent

Audiology and Neurotology

Research Article

Audiol Neurotol
DOI: 10.1159/000537768

Received: October 12, 2023
Accepted: February 7, 2024
Published online: March 18, 2024

Reference Data for a Quick Speech-in-Noise Hearing Test in the French Language

Julie Bestel[a] Daniel Pressnitzer[b] Mathieu Robier[c] Frédéric Rembaud[d]
Christian Renard[e] François Leclercq[e] Christophe Vincent[f]

[a]Audilab Ressources, Saint-Pierre-des-Corps, France; [b]Laboratoire des systèmes perceptifs, Département d'études cognitives, École normale supérieure, PSL University, CNRS, Paris, France; [c]Audilab, Orléans, France; [d]École d'Audioprothèse de Cahors, Cahors, France; [e]Laboratoire Renard, Lille, France; [f]Service Otologie et Otoneurologie, CHU de Lille, Lille, France

Keywords
Audiology · Speech perception · Cohort study

Abstract

Introduction: Difficulty in understanding speech in noise is the most common complaint of people with hearing impairment. Thus, there is a need for tests of speech-in-noise ability in clinical settings, which have to be evaluated for each language. Here, a reference dataset is presented for a quick speech-in-noise test in the French language (Vocale Rapide dans le Bruit, VRB; Leclercq, Renard, & Vincent, 2018). ***Methods:*** A large cohort (N = 641) was tested in a nationwide multicentric study. The cohort comprised normal-hearing individuals and individuals with a broad range of symmetrical hearing losses. Short everyday sentences embedded in babble noise were presented over a spatial array of loudspeakers. Speech level was kept constant, while noise level was progressively increased over a range of signal-to-noise ratios. The signal-to-noise ratio for which 50% of keywords could be correctly reported (speech reception threshold, SRT) was derived from psychometric functions. Other audiometric measures were collected for the cohort, such as audiograms and speech-in-quiet performance. ***Results:*** The VRB test was both sensitive and reliable, as shown by the steep slope of the psychometric functions and by the high test-retest consistency across sentence lists. Correlation analyses showed that pure tone averages derived from the audiograms explained 74% of the SRT variance over the whole cohort, but only 29% for individuals with clinically normal audiograms. SRTs were then compared to recent guidelines from the French Society of Audiology [Eur Ann Otorhinolaryngol Head Neck Dis. 2022;139(1):21–7]. Among individuals who would not have qualified for hearing aid prescription based on their audiogram or speech intelligibility in quiet, 18.4% were now eligible as they displayed SRTs in noise impaired by 3 dB or more. For individuals with borderline audiograms, between 20 dB HL and 30 dB HL, the prevalence of impaired SRTs increased to 71.4%. Finally, even though five lists are recommended for clinical use, a minute-long screening using only one VRB list detected 98.6% of impaired SRTs. ***Conclusion:*** The reference data suggest that VRB testing can be used to identify individuals with speech-in-noise impairment.

karger@karger.com
www.karger.com/aud

Correspondence to:
Julie Bestel, j.bestel@audilab.fr

Introduction

The standard test of the hearing status of individuals is the audiogram, which measures detection thresholds in quiet for pure tones at different frequencies. However, it has long been argued that speech-in-noise intelligibility should also be measured, in spite of the added cost and effort [1–3]. Indeed, difficulty in understanding speech in noise, such as during a lively conversation in a social setting, is the most common subjective complaint motivating a first audiological consultation [3]. Having a direct measure to objectify such complaints and being able to compare an individual's performance to reference data are therefore important for audiologists and patients alike, to decide whether an intervention may be appropriate. Also, improvement for speech in noise is the most important benefit expected by prospective hearing aid recipients [4]. Indeed, speech-in-noise performance is the best predictor for self-reported satisfaction with a newly acquired hearing aid [5]. Routine speech-in-noise testing may thus further help to objectively track progress. Finally, and specifically for the French language, recent regulations have introduced a prescription criterion based on speech-in-noise performance, which allows the prescription of a hearing aid even if the audiogram displays a pure tone average (PTA) better than 30 dB HL [6]. Thus, there is a renewed incentive for reliable and efficient speech-in-noise tests in the French language [7].

Unlike the audiogram, where the appropriate methodology is broadly accepted (e.g., ANSI S3.6-2004), there are several choices to be made when designing a speech-in-noise test [3, 8]. The speech material may consist of naturalistic sentences, or instead grammatically correct sentences with no semantic content, or isolated words, or even nonsense syllables. The masking noise may be stationary, such as speech-shaped noise, or with dynamic properties, such as babble noise. The target speech and masking noise may be spatially colocated or presented from different locations. The intelligibility measure may be the percent correct for a set of stimuli, or the signal-to-noise ratio (SNR) at which a target level of performance is observed, known as a speech reception threshold (SRT). If SRTs are chosen, they may be estimated with an adaptive procedure, where SNR varies according to the correctness of trial-by-trial responses, or with a constant stimuli procedure, where performance is collected over a fixed set of SNRs and psychometric functions are fitted. Finally, the way SNR itself is varied during a test may be through changing the speech intensity while keeping the noise intensity fixed, or instead by changing the noise intensity while keeping the speech intensity fixed. All of these a priori valid choices impact various sorts of trade-offs: the ecological validity of the test versus the contribution of nonauditory factors; the reliability of the test versus its time cost; and the subjective ease of the test for the patient versus the ease of scoring for the clinician [8].

A further issue for speech-in-noise tests is that it is not possible to develop a single international standard because the test material must be localized to each language. In English, several speech-in-noise tests have been developed and formally evaluated (e.g., BKB-SIN, HINT, QuickSIN, and WIN [9]). In French, a recent review has surveyed the options available to clinicians [7]. Briefly, for French from France, tests include digits in speech-shaped stationary noise (FrDigit3, [10]), semantically unpredictable sentences (FraMatrix, [11, 12]), everyday sentences in speech-shaped stationary noise (FIST, [11, 13]), or everyday sentences in babble noise (VRB, [14]; MBAA2, [15]; HINT-5 MIN, [16]). In addition to differences in audiological material, these tests also differ on several aspects of the measurement methodology, such as adaptive versus fixed stimuli paradigms or fixed speech level versus fixed noise level (see Ref. [7] for details).

Here, we investigate the VRB test [14], which is similar to the Quick-SIN test in the English language [17]. VRB measures the percent correct of reported words at different SNRs, from which an SRT is computed. The audio material consists of lists of everyday sentences embedded in babble noise. Each list contains nine sentences, with the speech level fixed and the noise level increased for each new sentence in the list. Participants are asked to repeat the sentences and scoring is made by the clinician on three keywords per sentence. The expected advantages and drawbacks of these choices were as follows. The material was intended to simulate ecological situations, with naturalistic speech and noise background. The fixed stimulus paradigm allows for "easy" trials at the beginning of each list, with a predictable and limited number of "harder" trials, thus encouraging participant's engagement and confidence during the task. The fixed speech level reduces issues related to the audibility of the target signal, instead focusing specifically on the effect of noise on intelligibility. As a trade-off, scoring has to be made manually by the clinician administering the test, unlike for automated closed-set methods [10–12].

The initial development of VRB has been described in Leclercq et al. [14]. This study motivated the design choices of the test but only included a limited dataset (N = 29). Normative data for normal-hearing participants (N = 200) have then been reported [18]. However, a dataset with a large cohort, covering various degrees of hearing impairment, is still lacking for VRB. The present study provides such a dataset. Here, VRB testing is reported for a very large cohort

($N = 641$), the largest for any test in the French language to the best of our knowledge. The cohort included normal-hearing and hearing-impaired participants, as identified by tonal audiometry. The severity of impairment was systematically varied over the cohort. Participants over a broad age range were also included, as aging may contribute to speech-in-noise difficulties with or without an associated tonal hearing loss [19]. We suggest that such a large and diverse cohort, representative of the population typically encountered by clinicians, is essential to a robust evaluation of the VRB test's reliability and sensitivity. The main aim of the present study is thus to present normative data for the VRB test in a clinical setting.

There is also a fundamental interest for large-scale investigations of speech-in-noise performance for participants with various degrees of tonal hearing loss. Obviously, for the unaided ear, the audiogram will provide a hard-limiting factor to intelligibility: any acoustic component below audibility will not contribute to intelligibility. However, there may be additional contributors to intelligibility [2]. For instance, speech in noise involves total or partial masking of the target speech signal by the noise. Masking is, in turn, linked to the frequency selectivity of the cochlea, which is often impaired by hearing losses of a cochlear origin [20]. Speech also contains temporal cues that could be impacted by cochlear hearing loss [20–22]. Finally, the recent discovery of cochlear synaptopathy has revealed a physiological mechanism that could lead to a disconnect between the audiogram and speech intelligibility in noise, resulting in "hidden hearing loss" [23–27].

The relevance of hidden hearing loss to everyday clinical practice remains a matter of debate. In an influential study, Killion and Niquette [28] attempted to predict speech-in-noise intelligibility from audiometric data. They failed to do so and eloquently concluded that "[. . .] the only reliable way to determine a patient's ability to hear in noise is to measure it." Given the far-reaching clinical impact of such a conclusion, quantifying the statistical link between speech-in-noise performance and the audiogram over large cohorts is an ongoing effort [12, 27]. A secondary aim of this study is thus to provide a new large dataset containing audiometric and speech-in-noise measures.

Materials and Methods

Recruitment

Participants were recruited from 8 different hearing aid centers, belonging to one of the two following commercial networks: Audilab and Renard Audiologie. The centers were spread over 8 cities throughout France (Marseille, Orléans, Dinan, La Rochefoucauld, Hénin-Beaumont, Lille, Douai, Maubeuge). Participants with hearing difficulties were all existing patients from either Audilab or Renard Audiologie. For those participants, data were collected during routine clinical visits. VRB testing was part of their clinical examination, whether or not they chose to enroll in the study. The normal-hearing participants were individuals who happened to accompany the patients for the clinical visits, as well as clinical trainees, assistants, and audiologists. No participant was familiar with the VRB material beforehand.

Ethical Approval

The study obtained ethical agreement from the "Comité de Protection des Personnes Sud-Est IV", approval number: 2018-A02948-47 (ID-RCB). Under French law (loi Jardé, "Recherche Impliquant la Personne Humaine"), and as confirmed by the ethical committee, the study qualified as "Category 3: non-interventional research in which all procedures and products are within clinical standard of care, without additional or unusual procedures of diagnosis, treatment, or supervision." Therefore, written informed consent was not required. Potential participants were instead provided with a study information sheet and were informed that they could refuse to participate, without any consequence on their clinical care. Moreover, participants were informed that they could withdraw from the study at any time, without providing any reason.

Inclusion and Exclusion Criteria

Inclusion criteria were: adult participants (>18 years of age at testing time); French speakers; PTA across both ears <65 dB HL; normal otoscopy; and sensorineural hearing loss only (for hearing-impaired participants). Exclusion criteria were: conductive hearing loss; hearing loss due to ototoxicity when known; PTA difference across ears >20 dB HL; and inability to understand the study and description or test procedures because of cognitive or language issues.

The procedure to enforce these inclusion and exclusion criteria was as follows. For new patients (less than 6 months from diagnosis), conductive hearing loss was measured using bone conduction thresholds. For patients with more than 6 months since diagnosis, conductive hearing loss was determined according to the medical history. If air conduction thresholds were available less than 6 months before the appointment, these were used; otherwise they were recollected. A thorough questionnaire was administered to participants to screen for past exposure to ototoxic agents. We acknowledge that this did not exclude participants who may not have been aware of past exposure to ototoxic agents. However, ototoxicity was an exclusion criterion simply to reduce sources of heterogeneity in the cohort [29]. We chose to focus on the more common forms of sensorineural hearing loss, leaving a characterization of ototoxic losses for future study.

Asymmetrical hearing losses (PTA difference across ears >20 dB HL) were exclusion criteria because the test involved spatially diffuse audio presentation, which could have led to "better ear" effects in the outcome. Such effects would also need to be characterized in future studies.

Finally, note that environmental factors such as exposure to noise were not part of the exclusion criteria. Thus, our normal-hearing sample was not as strictly defined as in other studies [18].

Cohort

A total of 644 participants were enrolled in the study following inclusion and exclusion criteria. From these, 3 participants had to be excluded from the analysis because of missing data due to

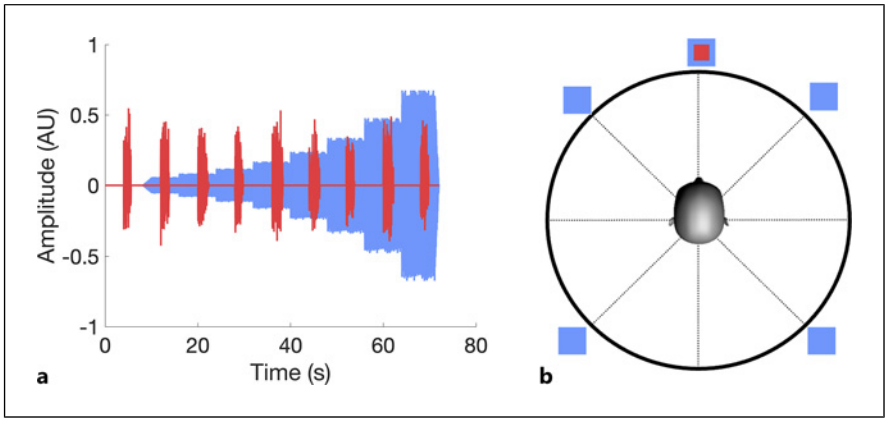

Fig. 1. Illustration of the VRB paradigm. **a** Illustration of a VRB list. A speech sentence (in red) is presented at a fixed level, first in quiet, and then in babble noise (in blue) increasing in 3-dB steps to cover SNRs from +18 dB to −3 dB. **b** Speech is presented at 65 dBA from a single loudspeaker located 1-m away in front of the participant. Noise is presented over five loudspeakers surrounding the participant.

technical error. Importantly, there was no exclusion based on the VRB test results. Even floor or ceiling outcomes were included in the analyses.

All analyses were thus based on N = 641 participants, henceforth termed "the cohort." The cohort comprised 314 females and 317 males (sex data unavailable for 10 participants). Age at the time of testing, in years, was: M = 58.0, SD = 22.9, maximum = 94, minimum = 19.

Pure Tone Audiometry

Pure tone air conduction thresholds were measured for each ear in each participant, using headphones, for audiometric frequencies 0.5, 1, 2, 4, 6, and 8 kHz, using standard clinical practice (International Bureau for Audiology, https://www.biap.org/en/). Participants were sorted into 4 different hearing status groups. The sorting criteria were identical to Bestel et al. [12] and followed the recommendations of the International Bureau for Audiology. First, the mean PTA (PTA_m) for right and left ear was computed for each participant, considering only frequencies of 0.5, 1, 2, and 4 kHz. Then, boundaries on PTA_m were applied to define hearing status groups: normal hearing (NH, $PTA_m \leq 20$ dB HL); mild hearing loss (MILD, $20 < PTA_m \leq 40$ dB HL); moderate hearing loss grade 1 (MOD1, $40 < PTA_m \leq 55$ dB); and moderate hearing loss grade 2 (MOD2, $55 < PTA_m < 65$ dB HL). The cutoff for the MOD2 group was lower than recommended by the International Bureau for Audiology (70 dB HL). As the VRB test operates at a fixed speech level, it was expected that losses above 65 dB HL would make the speech signal inaudible even in quiet.

In addition to the PTA_m summarizing the standard audiometric frequencies between 0.5 and 4 kHz, a high-frequency PTA, notated PTA_{m_hf}, was computed. For each participant, the PTA_{m_hf} was computed by averaging pure tone thresholds across the two ears for frequencies of 6 kHz and 8 kHz.

Speech-in-Noise Main Measure: VRB

The VRB design has been described in Leclercq et al. [14]. As the present study is intended as a future reference for VRB, we present the test method again.

Lists of everyday sentences were presented to participants at increasing levels of difficulty (Fig. 1a). The first sentence of a list was presented in quiet. Then, babble noise was added to each new sentence in the list, with the level of speech fixed and the level of noise progressively increased. Eight different SNRs were tested, from +18 dB to −3 dB in steps of 3 dB. Sounds were presented over loudspeakers (Fig. 1b). The sound level of the speech signal at the location of the participant was calibrated and set at 65 dB SPL. Participants provided their responses by repeating the speech sentence as accurately as possible. The clinician scored online the correctness of 3 predefined keywords for each sentence.

Sentences were selected from the MBAA corpus [15]. Sentence selection and characterization were detailed in Leclercq et al. [14]. Sentences were all recorded by the same female speaker in neutral French from France accent. The noise was a babble noise consisting of two male and two female talkers conversing in French and English [30]. A single 8-s loop of noise was used for all sentences, with the aim of reducing variability across sentences [14].

Speech and noise were delivered through loudspeakers, in a configuration designed to simulate a talker facing the participant while embedded in diffuse noise surrounding the participant (Fig. 1b). Specifically, the target speech was presented from a single loudspeaker located directly in front of the participant (0° azimuth). The babble noise was presented from 5 loudspeakers (0°, 45°, 135°, 225°, and 315° azimuth). The same noise without delay or attenuation was presented from each of the 5 loudspeakers. The participant was seated in the middle of the speaker array, at a distance of 1 m from each loudspeaker. Stimuli were presented from an audiometer connected to the 5 loudspeakers. Note that this loudspeaker setup is commonly available in audiology laboratories in France, where it is installed in audiometry sound insulating booths. All participants of the present study were tested in such sound insulating booths, which were all certified to the AFNOR ISO 8253-1 norm.

Raw results for the VRB test were the percent correct of reported keywords for each SNR. They were summarized by a single SRT for each individual, notated SRT_{vrb}. The SRT_{vrb} was estimated through the Spearman-Kärber formula, as prescribed by Leclercq et al. [14]:

$$SRT_{vrb} = i + \frac{d}{2} - d.r \Big/ n$$

with i the value of the easiest SNR (here, +18 dB); d the step difference between trials (here, +3 dB); n the number of items scored within a trial (here, 3); and r the total number of correct responses in a list (here, from 0 to 24, corresponding to all responses for 8 SNRs and 3 items per SNR). Note that this value has been termed "SNR-50" [14] or "SNR loss" [18] in previous publications on VRB.

Other Speech Measures: Lafon Lists

For each participant, unaided word recognition in quiet was also measured. Monosyllabic French words from the Lafon lists [31] were presented over the same target loudspeaker at 0° azimuth as used in VRB testing. Stimuli were presented from the audiometer connected to the loudspeaker. The soundfield measured at the location of the participant was again set at 65 dB SPL. Two lists of 17 words each were tested for each participant. During testing, a word from the list was presented to the participant, who was asked to repeat what they heard as accurately as possible. The clinician scored online each phoneme that was correctly identified, as well as each whole word correctly identified. Phoneme scores and word scores were averaged across the two lists.

Statistical Analyses

Because of the large sample size, standard parametric tests were used to assess statistical significance. An alpha level of $\alpha = 0.01$ was set a priori. Analyses were implemented in Matlab R2020a (The Mathworks) except for the ANOVA and the intra-class correlation coefficient that were implemented in JASP [32].

To test for differences between a distribution and a single value, or between two distributions, two-tailed t tests were applied. To test for the effect of hearing status on SRT_{vrb}, a one-way ANOVA was applied with hearing status group as the factor. Effect sizes were reported using the η^2 statistics. Because an effect of hearing status was observed, post hoc test comparisons between all hearing status groups were run using the conservative Bonferroni correction.

Psychometric functions were further fitted to the raw percent correct results obtained for the different SNRs, first averaged for each hearing status group. A logistic function was used to fit speech intelligibility:

$$SI\,(SNR) = \gamma + (1-\gamma)*1\Big/\left(1+e^{4s\left(SRT_{fit}-SNR\right)}\right)$$

with SI the proportion of correct responses at each SNR, SNR the different SNR values tested, γ the guess level, SRT_{fit} the SNR at 50% correct, and s the slope at SRT_{fit}. The logistic function was fit using nonlinear regression, with all parameters of the function left free to vary. The method was similar to that used in previous studies investigating French speech-in-noise tests, to facilitate comparison [10, 11, 14].

Individual fits for each participant were also performed, using the same method and formula. The participants ($N = 43$) who did not register a single correct response were excluded from this analysis, as for them the fit would be meaningless. Furthermore, because of the greater variability in individual data compared to the averaged data, we set limits on the parameter search range. In particular, to accommodate for nonmonotonicity in a small number of participants, we imposed a positive slope of at least 1%/dB. Other arbitrary bounds, determined by considering the average fits and visual inspection of the individual fits, were: $-0.2 < \gamma < 0.2$, $-10 < SRT_{fit} < 50$. As visually checked, fits were satisfactory. The average of the root mean square error was $RMSE = 0.05$.

Reliability analyses were performed using two different measures. First, the average intra-participant standard deviation divided by $\sqrt{2}$, notated SD_{intra} [33], was computed. We also computed the intra-class correlation coefficient $ICC1,1$ [34] for the whole cohort, with 641 participants and 5 measurements per participants. Correlations between measures were quantified using Pearson correlation coefficients or linear regression models and their significance assessed using standard t tests.

Acoustic Analyses

To illustrate the speech and noise material used by VRB, acoustic analyses were performed. Time-frequency analyses were obtained by passing the stimuli in a standard simulation of peripheral auditory filtering using "gammatone" filters, with filter widths adjusted to normal-hearing listeners [35]. The output of each filter was then half-wave rectified, square-root compressed, and low-pass filtered at 10 Hz. The resulting time frequency representations are termed "cochleagrams" [36]. Long-term spectral analyses were also obtained by applying third-octave filtering and computing the average power in each frequency band.

Results

Acoustic Analyses

We first present an acoustic analysis of the speech and noise material used in VRB, which had not been available up to now. Figure 2a–c shows the waveforms and "cochleagrams" for a typical sentence (Fig. 2a), for a noise excerpt (Fig. 2b), and for a mixture of the two at 0 dB SNR (Fig. 2c). Cochleagrams show the time-frequency output of an auditory model simulating peripheral auditory filtering [36].

For speech, typical temporal and spectrotemporal modulations can be seen in the waveform and cochleagram. The babble noise displays much fewer temporal modulations in the waveform, but the cochleagram reveals that a dynamic spectrotemporal structure is indeed present (unlike what would be observed for speech-shaped noise, for instance). When mixing speech and noise at 0 dB SNR, the sparse emergence of speech features can be observed (emergence defined as SNR >0 dB for each time-frequency bin). Note that emergence is an acoustical criterion, which may or may not represent the speech features actually used by listeners.

Figure 2d shows the average third-octave spectrum of all speech sentences used in the VRB test, together with the average third-octave spectrum of the full noise loop. Relative levels have been set to 0 dB SNR for this analysis. Overall, the speech and noise material of VRB are well matched in terms of spectral composition. Both have highly similar third-octave spectra between about

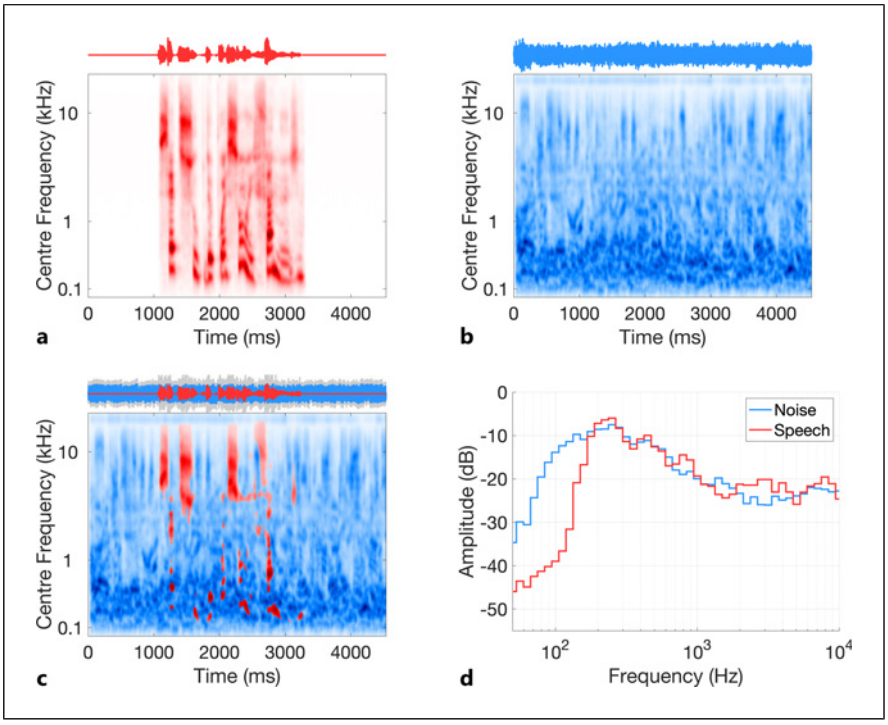

Fig. 2. Acoustic analyses of the VRB speech and noise material. **a** Waveform and cochleagram (see text for details) of a speech sentence. **b** Waveform and cochleagram of the babble noise. **c** Waveform and cochleagram of speech in noise at 0 dB SNR. In the cochleagram, speech is only shown for SNRs greater than 0 dB. **d** Average third-octave spectra for all speech and noise material, normalized at 0 dB SNR.

200 Hz and 10 kHz, covering the range of the female voice that was used to record the target speech sentences. The noise contains more low-frequency components, below 200 Hz, because it additionally includes male talkers with lower-pitched voices.

Audiometric Characterization of the Cohort

Figure 3 illustrates the hearing status of the cohort, as measured by tonal audiometry. Figure 3a displays the distribution of PTA_m split across subgroups. The distribution of participants across subgroups was as follows: NH: $N = 188$; MILD: $N = 148$; MOD1: $N = 212$; MOD2: $N = 93$. By construction, each subgroup occupied a different PTA_m range.

Figure 3b shows, for the same individuals sorted in the same subgroups, the high-frequency pure tone average PTA_{m_hf}. Note that PTA_{m_hf} and PTA_m were computed over nonoverlapping frequencies (6 kHz and 8 kHz vs. 0.5–4 kHz, respectively). There was a large spread in PTA_{m_hf}, even as participants were drawn from homogeneous groups relative to the lower frequency PTA_m.

Figure 3c depicts the asymmetry of hearing losses in the cohort, which was reduced a priori by enforcing an exclusion criterion of 20 dB HL asymmetry. Figure 3c shows that the level of residual asymmetry was focused around 0 dB HL and broadly consistent across hearing-impaired subgroups. Quantitatively, 93.3% of the cohort displayed an absolute difference between the PTAs of the left and right ears that was less than 10 dB HL. Per subgroup, the percentage of participants per subgroup with an absolute asymmetry less than 10 dB HL was: NH, 98.9%; MILD, 94.6%; MOD1, 88.7%; and MOD2, 90.3%. This reduces the influence of a "better ear advantage" on the speech-in-noise task.

Finally, Figure 3d illustrates the shape of the audiograms observed for the different subgroups in the cohort. Pure tone thresholds at audiometric frequencies were first averaged across the two ears for each participant, and then the mean and standard deviation across participants were computed for each subgroup. The audiograms had a shape typical of sensorineural hearing losses, as expected from the exclusion criteria.

Speech in Noise: VRB

Figure 4 illustrates the results of the VRB speech-in-noise task for the cohort, across the different subgroups. The summary outcome of the task, SRT_{vrb}, is shown. The SRT_{vrb} values were bounded to $SRT_{vrb} = -4.5$ dB for the best performers (all keywords correctly identified) to $SRT_{vrb} = +19.5$ dB for the worst performers (no keyword identified).

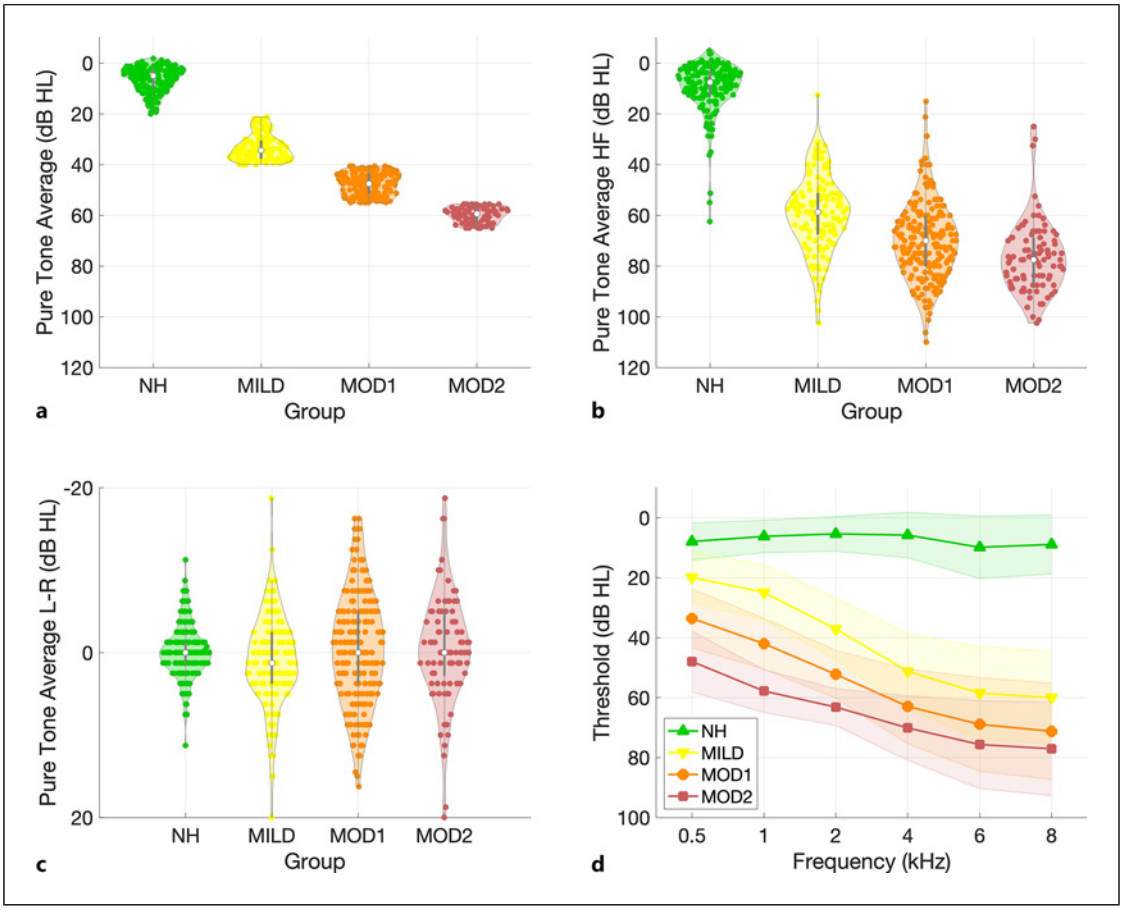

Fig. 3. Characterization of the cohort. **a** PTA audiometric thresholds over both ears and over frequencies of 0.5, 1, 2, and 4 kHz for all participants, split into hearing status groups (PTA_m, see text). Individual participants are shown as dots, with superimposed violin-plot distributions, medians, and interquartile ranges. **b** PTA audiometric thresholds over both ears and for frequencies 6 and 8 kHz for the same participants (PTA_{m_hf}, see text). **c** Asymmetry across ears in PTA thresholds for individual participants. **d** Average audiograms per hearing status group. The mean is shown together with the standard deviation about the mean as shaded areas.

Let us first describe the distribution of SRT_{vrb} for the NH subgroup. Individual points in Figure 4a correspond to the average SRT_{vrb} computed across all five sentence lists run by each participant. As intended by the initial calibration of the VRB task [14], SRT_{vrb} for normal-hearing individuals were distributed around 0 dB (NH SRT: M = 0.77 dB; SD = 1.51 dB, Min = −2.90 dB; Max = +4.90 dB). Interestingly, no individual reached perfect performance, so there was no ceiling effect in the NH subgroup. In the cohort, the mean NH SRT_{vrb} of 0.77 dB was significantly different from 0 dB ($t(187) = 6.98$, $p < 0.01$). When restricting the analysis to NH listeners that were less than 30 years old ($N = 119$), as in the VRB normative data for normal hearing [18], the mean NH SRT_{vrb} decreased to 0.51 dB but it was still significantly different from 0 dB ($t(118) = 3.95$, $p < 0.01$). A further analysis was performed to identify NH participants that displayed a high-frequency loss. In the NH subgroup, 16 participants with a $PTA_{m\_hf} > 20$ dB were observed. These participants were older than the rest of the NH subgroup: age in years M = 47 compared with M = 28 for NH participants with $PTA_{m\_hf} \leq 20$ dB. When computing the mean SRT_{vrb} for the NH subgroup with these 16 participants excluded, a value of 0.65 dB was found, still significantly different from 0 dB ($t(171) = 6.04$, $p < 0.01$).

For the other subgroups, as expected, SRT_{vrb} generally increased with the severity of the tonal hearing loss. The average SRT_{vrb} values per subgroup were as follows: MILD, M = 6.58 dB; MOD1, M = 11.50 dB; and MOD2, M = 17.15 dB. The effect of hearing status on SRT_{vrb} was highly significant, as shown by a one-way ANOVA with a subgroup as a factor: $F(3,637) = 631.68$, $p < 0.01$,

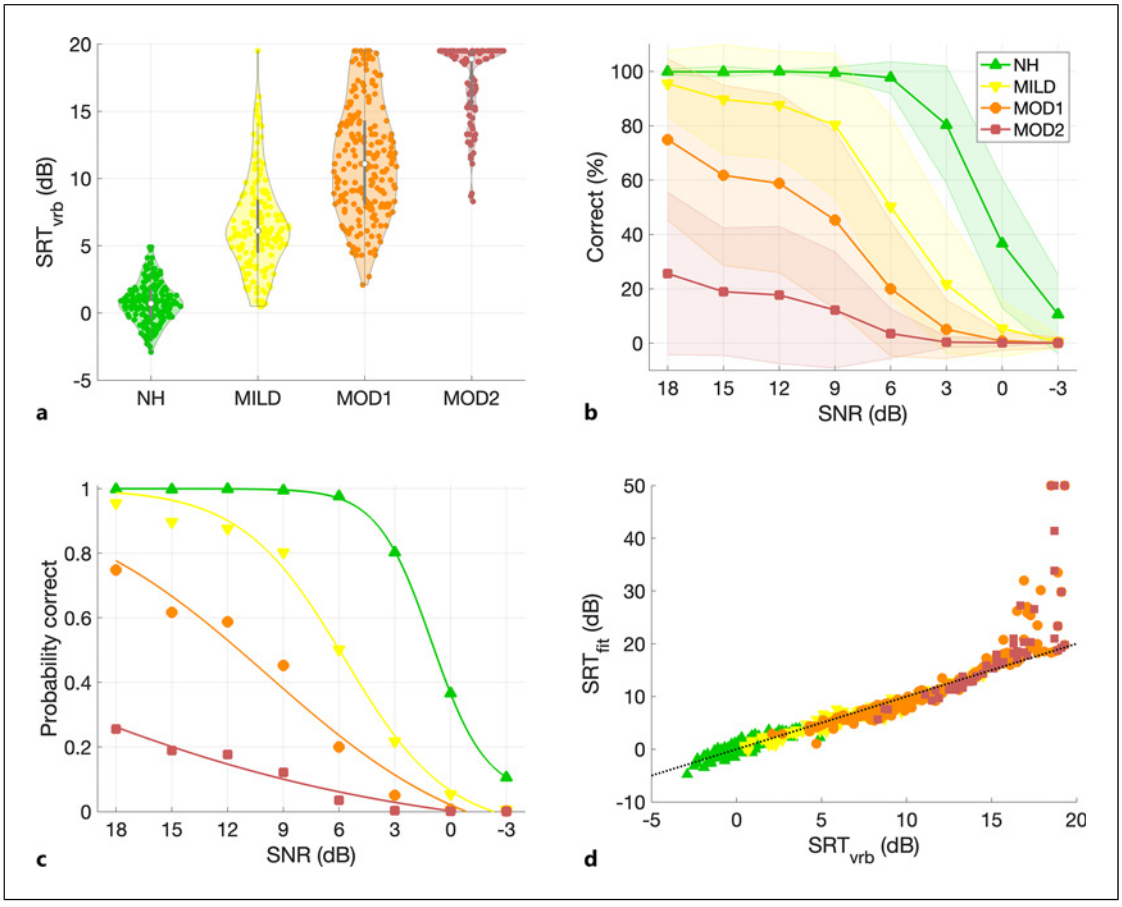

Fig. 4. Results of the speech-in-noise VRB task. **a** Speech reception threshold (SRT_{vrb}, see text) for individual participants split across hearing status group. **b** Average percent correct for each SNR tested in the VRB task, split across hearing status group. The mean is shows together with the standard deviation about the mean. Note that SNRs are presented in reverse order, to reflect the order of testing in the VRB task. **c** Psychometric functions fitted to the average percent correct data. **d** Comparison of the standard SRT_{vrb} measure and SRT_{fit} estimated from individual psychometric fits.

$\eta^2 = 0.75$. Post hoc tests comparing the different subgroups showed that they all differed in terms of SRT_{vrb}: all pairwise comparisons $p < 0.01$, Bonferroni corrected.

It is noticeable from Figure 4a that the spread of SRTs was very different across subgroups. For NH individuals, all SRTs were narrowly distributed around their mean value, close to 0 dB. However, there was a much larger spread of values for all other subgroups involving hearing impaired individuals. This spread was particularly striking for the MILD and MOD1 subgroups, where SRT_{vrb} values were observed covering the full range from normal values down to floor values. Note that this was in in spite of the subgroup's homogeneity in terms of hearing loss etiology and of the tonal loss severity, as estimated by PTA_m. For these two subgroups at least, this finding suggests that SRTs were partially dissociated from the tonal hearing loss as measured by the audiogram. Finally, for the last subgroup with the most severe tonal hearing losses, MOD2, there was also a large spread of SRT_{vrb} but also a noticeable floor effect. A number of individuals could not identify a single keyword throughout all five lists ($N = 43$ overall, $N = 36$ in the MOD-2 group). In this case, their actual SRT was essentially unknown.

To further characterize the VRB results, we also examined the percentage correct for the different SNRs tested. Figure 4b shows the average percent correct observed at different SNRs across subgroups. Percent correct generally decreased with decreasing SNR, for all subgroups of participants. However, the shape of the psychometric functions differed markedly across subgroups. On average, NH participants reached ceiling

performance at 6 dB SNR. For the MOD2 participants, floor effects were observed around the same 6 dB SNR value. For the MILD and MOD1 groups, the psychometric function was well centered over the range of tested SNRs.

Next, we fitted psychometric functions to the percent correct results, using a logistic function to fit the probability of correct responses at each SNR [10, 11, 14]. Figure 4c shows the results of the fitting procedure applied to the average data of Figure 4b. The fit was satisfactory for all subgroups, even for the MOD2 group where average performance did not reach the 50% correct defining the SRT.

The fits were first used to estimate the steepness of the psychometric function around SRT, a useful indicator of the sensitivity of the method as steeper curves produce more reliable estimates. For the NH subgroup, steepness was high: s = 17.15%/dB. This is in line with previous estimates of s = 19.3%/dB for VRB on a smaller cohort [14]. The slope of the average psychometric function decreased for the hearing-impaired subgroups: MILD, s = 8.99%/dB; MOD1, s = 4.38%/dB; and MOD2, s = 2.22%/dB. Note that the estimate for the MOD2 group may be biased by individuals with floor performance.

The fits were then used to compute another estimate of SRT, SRT_{fit}. Unlike SRT_{vrb}, SRT_{fit} is not hard-bounded and may produce results beyond the range of SNRs that were actually tested. From the average data of the different subgroups, SRT_{fit} values were as follows: NH, SRT_{fit} = 1.02 *dB*; MILD, SRT_{fit} = 5.72 *dB*; MOD1, SRT_{fit} = 9.78 *dB*; and MOD2, SRT_{fit} = 26.17 *dB*.

Finally, individual fits were produced for each participant. Figure 4d illustrates the relationship between the two SRT estimates. In general, the estimates were very consistent. Some discrepancies appeared for higher SRT values for the MOD2 group. This suggests that it may be possible to obtain estimates of SRTs beyond the range of measured SNRs for the more impaired participants of the cohort. However, the fits may also be more brittle for those participants, as indicated, e.g., by 3 fits reaching the boundary values arbitrarily imposed on the search space.

Other Speech Measures: Words and Phonemes in Quiet

Performance on words in quiet intelligibility was also assessed for the cohort, using the classic Lafon lists for the French language – the so-called "listes cochléaires monosyllabiques de Lafon" [31]. Mean performance and standard deviation by subgroup were: NH, M = 99.41%, SD = 1.52%; MILD, M = 82.19%, SD = 15.92%; MOD1, M = 59.57%, SD = 22.77%; and MOD2, M = 25.66%, SD = 23.25%. In the same task, phoneme recognition was also scored. Results were as follows: NH, M = 99.80%, SD = 0.61%; MILD, M = 92.56%, SD = 7.89%; MOD1, M = 77.51%, SD = 18.47%; and MOD2, M = 41.28%, SD = 29.32%. Overall, for these additional measures, the NH subgroup was close to ceiling, as expected. For all other groups, results follow the degree of hearing impairment, with sizeable variability.

Because the first item of each list in the VRB task is a speech in quiet measure, another independent word-in-quiet measure may be computed by averaging performance for the first item of each VRB list. For this alternate measure, results were as follows: NH, M = 99.96%, SD = 0.49%; MILD, M = 97.70%, SD = 10.07%; MOD1, M = 84.25%, SD = 27.02%; and MOD2, M = 35.48%, SD = 37.22%. The VRB and Lafon estimate of words in quiet were correlated ($r(637) = 0.77$, $p < 0.01$, two missing values for the Lafon lists), but performance was overall higher for the VRB estimate ($t(638) = 16.50$, $p < 0.01$). As the two estimates were derived for the same acoustic level of the speech material, such a difference must stem from other features of the tasks (e.g., choice of words, structure of the task).

Reliability Analyses

For the VRB test, five different lists were run by each participant, with a time cost of about 1 min per list. We investigated whether there were learning effects over lists. In particular, procedural learning may have led to a sharp improvement in performance between the first and second lists. Figure 5a shows the results pooled over the whole cohort, split across lists. The average SRT_{vrb} for successive lists were as follows: List 1: 8.38 dB, List 2: 8.00, List 3: 7.95, List 4: 7.86, and List 5: 8.00. A significant change was observed between the first and second lists ($t(639) = 3.98$, $p < 0.01$). All other comparisons between successive lists were not significant ($p > 0.05$). To be conservative, for this particular analysis we did not include any correction for multiple comparisons in the t-tests. In addition, effect size as estimated by Cohen's d was small ($d = 0.16$). Therefore, there was no sign of a sharp improvement across the first two lists indicating procedural learning, or of improvements between subsequent lists.

An important characteristic of any clinical test is its test-retest reliability, which intuitively quantifies the expected variability in outcome if the test is repeated several times for the same individual (assuming that the underlying "true" SRT does not vary). One measure to quantify reliability is the average intra-participant standard deviation divided by $\sqrt{2}$, notated SD_{intra} [33]. The

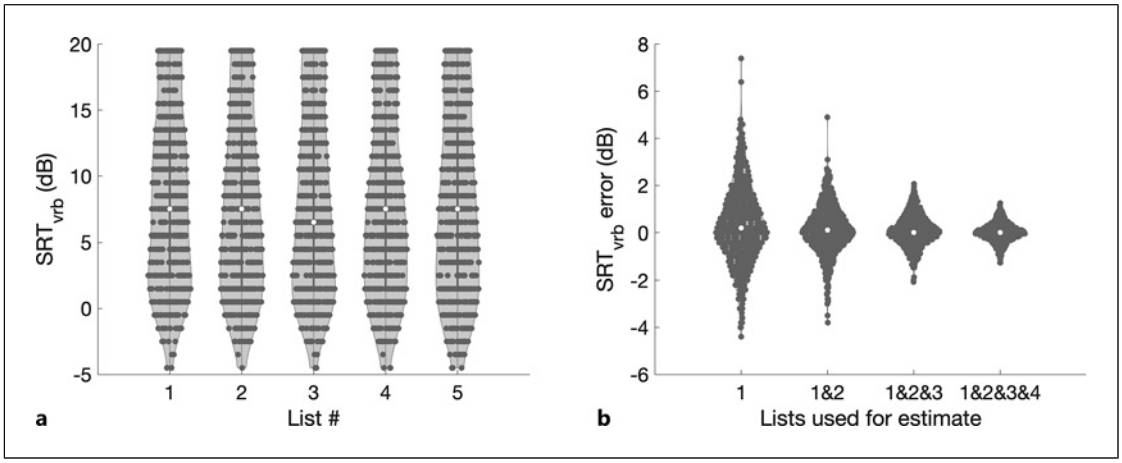

Fig. 5. Reliability analyses. **a** Individual results for all participants for successive lists of the VRB task. **b** Moving average of the SRT_{vrb} estimate. The best estimate is assumed to be the one obtained by averaging results for the 5 lists tested. Plotted is the difference, for individual participants, between an estimate averaged over 1, 2, 3, or 4 lists and the best estimate over 5 lists.

overall reliability of VRB was in line with other speech-in-noise tasks, with SD_{intra} = 1.07 dB. Across subgroup, reliability was consistent: NH, SD_{intra} = 0.94 dB; MILD, SD_{intra} = 1.13 dB; MOD1, SD_{intra} = 1.29 dB; and MOD2, SD_{intra} = 0.76 dB. Note that the variability for the MOD2 group is underestimated because of the floor effect for some participants. Another measure of test-retest reliability is the intra-class correlation, ICC, which estimates the ratio between the variance of interest and the total variance including measurement error [34]. The observed ICC for the whole cohort and 5 different measurements was in the "excellent" range: *ICC1,1* = 0.934, 95% confidence interval = (0.926; 0.941). Furthermore, we checked that there were no biases in the estimate by using other models of ICC, which provided identical estimates [34].

We finally used the guidelines from the French Society for Audiology [6] to set a cutoff point of 3 dB for SRT_{vrb}, in order to decide whether an individual should be categorized as having a clinically impaired SRT or not [18]. In the present cohort, and using the average of all five lists as ground truth, N = 445 individuals were classified as having clinically impaired SRTs. We then estimated how many of these individuals would have been correctly classified as impaired based on an SRT_{vrb} estimate from the *first* list only. This percentage was 98.6%. Adding additional lists increased the correct detection rate to more than 99.4%. We also estimated the false alarm rate, that is, the percentage of individuals wrongly classified as clinically impaired. This percentage was 3.1% after a single list and it steadily decreased to 1.6% as more lists were included, showing a high selectivity for the screening measure.

Correlation between Measures

Figure 6 shows scatterplots illustrating the relationships between the different measures collected in the cohort. Figure 6a shows the relationship between age and pure tone audiometry, summarized by PTA_m. As commonly observed, older participants tended to exhibit larger PTA_m values. However, it is noticeable that the cohort included a broad spread of age for all PTA_m. In particular, the MILD and MOD1 subgroups contained individuals covering most of the age span included in the study.

Figure 6b and c shows the relationship between speech-in-noise performance and pure tone audiometry, by displaying SRT_{vrb} as a function of PTA_m. It is clear from Figure 6b that there was an overall relationship between the two variables: larger (worse) PTA_m values were generally associated with larger (worse) SRT_{vrb} values. A Pearson correlation coefficient of $r(639) = 0.86$ was observed, suggesting a strong correlation explaining about 75% of the variance. The correlation was significant ($p < 0.01$), but this is not overly surprising given the large sample size. Even though age and PTA_m covaried (Fig. 6a), we further tested whether adding age as a factor in a multiple regression would markedly improve the correlation. It did not: a model with age and PTA_m resulted in R^2 = 0.752, compared to R^2 = 0.748 with PTA_m only.

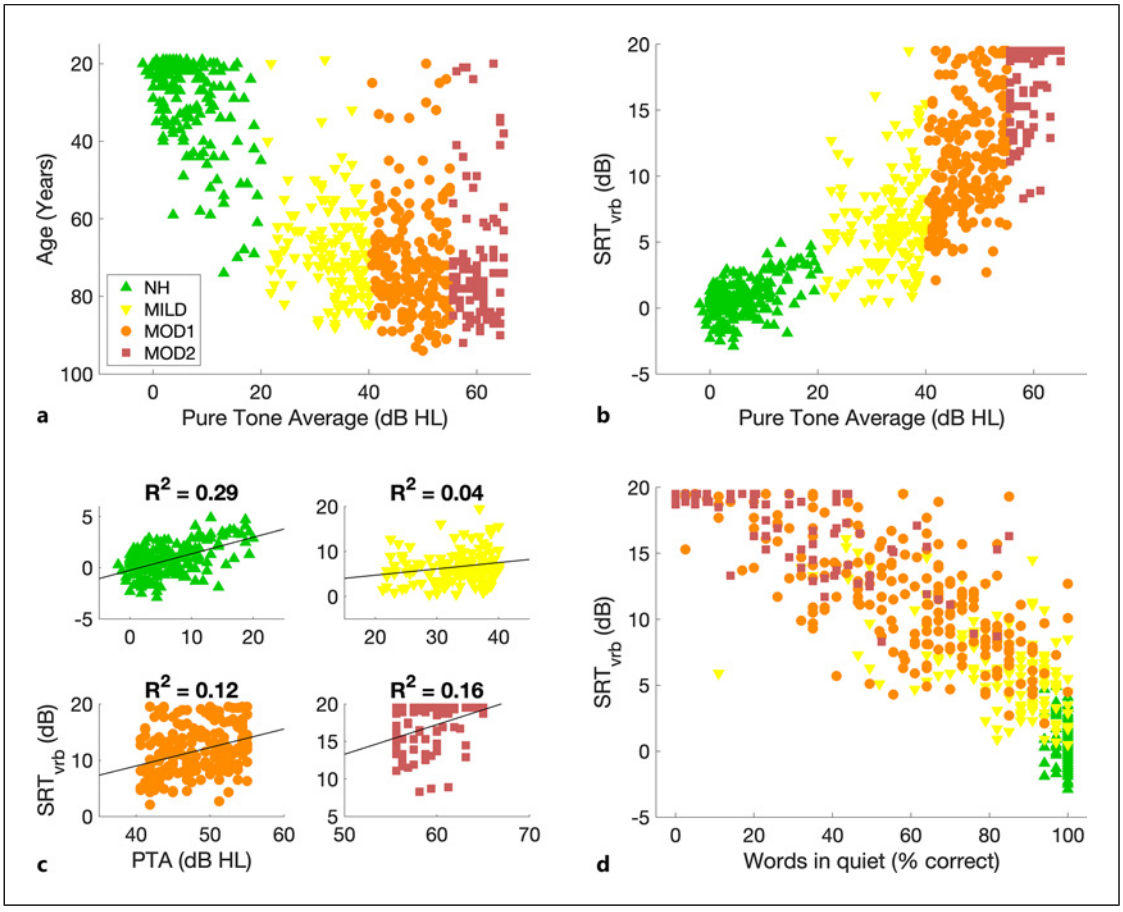

Fig. 6. Scatter plots. **a** The age of each participant is shown as a function of the summary audiometric measure PTA_m. **b** Speech-in-noise performance estimated by SRT_{vrb} is shown as a function of audiometric status estimated by PTA_m. **c** As for **b**, but split across hearing status group. The estimated variance R^2 explained by a linear fit is indicated on top of each subpanel. **d** Speech-in-noise performance estimated by SRT_{vrb} as a function of the score for words-in-quiet on the Lafon test.

However, two caveats must be mentioned. First, the relationship between PTA_m and SRT_{vrb} did not look linear, so a data transformation may yet increase the strength of the correlation. Second, the relationship was driven by the large average differences in SRT_{vrb} across subgroups. Specifically, for the MILD and MOD1 groups, the relationship did not appear particularly tight. Figure 6c replots the same data, but now focusing on each subgroup. A linear regression model was fitted to the data for each subgroup and the best fitting line is shown, together with the R^2 measure of explained variance. Interestingly, for the NH subgroup, there was a sizeable proportion of the SRT_{vrb} variance (about 30%) that was captured by PTA_m. For the other subgroups, the relation was weak. In particular, for the MILD subgroup, only 4% of the variance in SRT_{vrb} was explained by tonal audiometry. This means that, for the MILD group of participants with borderline audiometric results, a very large spread of speech-in-noise performance was observed.

A similar correlation analysis per subgroup was performed using the high-frequency PTA_{m_hf}. The R^2 measures were actually worse than for the correlation with the standard PTA_m: NH, $R^2 = 0.13$; MILD, $R^2 = 0.03$; MOD1, $R^2 = 0.00$; and MOD2, $R^2 = 0.00$. This shows that, in the present dataset, the unexplained variability in SRT_{vrb} cannot be accounted for by high-frequency audiometry.

Finally, Figure 6d shows the relationship between a standard measure of speech in quiet for the French language (Lafon lists of word) and SRT_{vrb}. Again, the relationship was in the expected direction: better intelligibility for words in quiet led to smaller SRT_{vrb}. The

correlation was sizeable, as measured by the Pearson correlation coefficient ($r(639) = -0.88$, $p < 0.01$). However, here again, there was a lot of spread, especially for the MILD and MOD1 subgroups. Not shown are the relationship between SRT_{vrb} and the phonemes identification scores ($r(639) = -0.80$, $p < 0.01$) and between SRT_{vrb} and the words in quiet measure from VRB ($r(639) = -0.74$, $p < 0.01$), which led to a similar conclusion: speech-in-quiet and speech-in-noise are correlated, but there remains substantial unexplained variance in SRT_{vrb}.

Eligibility for Hearing Aid Prescription

As a last data analysis, we revisited one of the original motivations of the present study: the new guidelines for hearing aid prescriptions in France [6]. Previously, hearing aid prescription was recommended for individuals with a tonal hearing loss corresponding to PTA_m >30 dB HL. Now, the new guidelines additionally recommend hearing aid prescription for individuals with a speech-in-noise impairment corresponding to an elevation in SRT of 3 dB or more compared to the norm, irrespective of tonal audiometry. An analysis was thus performed to evaluate how many individuals of our cohort would be eligible for hearing aid prescription if results of the VRB testing were taken into account.

Individuals with $PTA_m \leq 30$ dB HL, who would not be eligible for hearing aid prescription based on tonal audiometry alone, were first identified. There were $N = 233$ of them, comprising all participants in the NH subgroup ($N = 188$) and some participants in the MILD subgroup ($N = 35$). Among them, individuals with a 3-dB speech-in-noise impairment were further identified. This was done by selecting all participants with $SRT_{vrb} \geq 3$ dB. There were $N = 41$ of them (NH: $N = 16$, MILD: $N = 25$). Converted to percentages, these results show that 18.4% of all individuals who would not have qualified for hearing aid prescription based on their audiogram alone were now eligible based on speech-in-noise impairment revealed by the VRB test. Moreover, when focusing on "borderline" audiograms, with 20 dB HL < $PTA_m \leq 30$ dB HL, the prevalence newly eligible individuals increased to 71.4% after VRB testing.

Discussion

Summary of Findings

A dataset comprising audiometric measures and speech-in-noise measures has been collected on a large cohort ($N = 641$) of French speakers. The main aim of the study was methodological, to characterize the VRB speech-in-noise test [14, 18] over a broad range of individuals displaying varying degrees of sensorineural hearing loss. The secondary aim was fundamental, related to the ongoing question of whether speech-in-noise performance can be predicted by tonal audiometry or not in a clinical setting [1, 27].

About the methodological aim, the present dataset represents the largest and most diverse cohort in the French language for speech-in-noise testing. Thus, the results should be representative of the population encountered by clinicians and could serve as future reference for the VRB method. To further characterize VRB, acoustical analyses of its speech and noise material were also provided, as well as analyses not available in the commercial version of the test (Hubsound, Biotone), such as the fitting of psychometric functions, reliability analyses, or the investigation of the screening power of a single list of 1-mn duration. Overall, VRB was found to be fast, sensitive, and reliable.

About the fundamental aim, we found a large amount of variability in speech-in-noise performance that could not be accounted for by audiometric thresholds. Even though there was a clear correlation between tonal audiometry outcomes and speech-in-noise outcomes over the whole cohort, the relationship was much weaker within hearing status subgroups, with for instance 71% of unexplained variance in the normal-hearing subgroup and 96% of unexplained variance in the mild loss subgroup. As a result, using a 3-dB SRT elevation as a criterion, a sizeable proportion of individuals (18.4%) were revealed as suffering from a speech-in-noise impairment even though they had clinically normal tonal audiograms (under French guidelines, $PTA_m \leq 30$ dB HL). Such individuals would now be eligible for hearing aid prescriptions [6]. Even though it remains to be seen whether current hearing aids can improve SRTs for this newly eligible population, such a result adds further evidence to the usefulness of including speech in noise in routine clinical evaluations. The remainder of the discussion develops each of these findings in details, starting with the methodological findings before reexamining the fundamental findings.

Useful Characteristics of VRB Testing

The VRB test uses meaningful everyday sentences, combined with naturalistic babble noise and a spatialized audio presentation (Fig. 1, 2). The intention was to simulate situations that would be familiar to listeners, such as for instance trying to follow a conversation in a noisy restaurant. The use of a complex babble noise is also

suited to future tests using hearing aids, as more simple forms of noise such as stationary speech-shaped noise would be too easily canceled by modern denoising techniques and thus unrepresentative of realistic listening conditions.

The VRB task initiates each test list with a first trial using speech in quiet. This is the easiest possible condition of the test and should help familiarize participants to the procedure and material. Also, this first trial provides a speech in quiet measure that should be comparable to the Lafon lists commonly used by French audiologists [31], thus offering the opportunity to collect both speech-in-quiet and speech-in-noise performance in a single coherent testing procedure.

In VRB, the SRT measure is achieved by varying the level of the noise and not of the target speech, in a fixed stimulus paradigm. Therefore, unlike for adaptive procedure, a measure of SRT should be collected in a predictable amount of time, even for participants with poor performance. Moreover, again unlike adaptive procedures, the use of a fixed stimulus paradigm does not dwell disproportionately close to the threshold SNR of each participant, which is associated with a high subjective difficulty and failure rate. Rather, with the calibration retained [14], we found that the VRB psychometric functions were well centered around threshold for a range of hearing losses, and especially so for the light and moderate losses for which speech-in-noise performance is the most informative (Fig. 4).

In addition to these characteristics, the present results show that the VRB task exhibits high sensitivity. Sensitivity was estimated by means of the slope of the psychometric function at threshold. A steep slope indicates that the percent correct changes rapidly around threshold, thus making the SRT estimate more robust to measurement noise. The slope of VRB was found to be on par with other French speech-in-noise tests. For normal-hearing listeners, we observed a slope of s = 17.15%/dB, compared for instance to s = 14.0%/dB for FraMatrix [11] or 20.2%/dB for FIST [13].

Moreover, the test-retest reliability of VRB was high. Even though some evidence for an improvement between the very first list and the following ones was observed, such an improvement was small. The overall test-retest reliability across lists, as quantified by the ICC, was 93.4%, which is described as "excellent reliability" [34]. The intra-participant standard deviation, another measure of reliability, was of 1.07 dB, on par with other French speech-in-noise tests, with 0.4 dB for FraMatrix [11] and 1.02 dB for FIST [13].

Finally, in the cohort, there was a very large spread of SRT_{vrb} within hearing status groups with matched pure tone audiometry, especially so for the MILD and MOD1 subgroups. If, as suggested by the accuracy and reliability measures of the VRB task, such variability is not simply measurement noise, then it brings useful information to characterize the hearing status of an individual beyond the audiogram.

Comparison with Previous VRB Studies

Previous studies of VRB have focused on the calibration of the material [14] and normative normal-hearing data for various age groups [18]. We provided several further analyses that could be useful for future clinical uses of the test.

The commercial software that is available to clinicians (Hubsound, Biotone) runs 5 different lists and uses the Spearman-Kärber to provide an estimate of SRT_{vrb}. Here, we have used the same formula on each of the 5 successive lists to estimate training effects and reliability. As a result, we could show that useful screening results would be available from a single list. The number of lists to be run could be adjusted by the clinician according to the desired trade-off between speed and accuracy.

We also provide a new set of data for normal-hearing listeners. We did not find an average SRT_{vrb} of 0 dB, but rather close to 0.8 dB for the whole normal-hearing subgroup and 0.5 dB for the young (<30 years old) normal-hearing subgroup. This represents a small deviation from the normative value of 0 dB suggested by previous studies and used in the commercial software [14, 18]. The clinical relevance of this 0.5 dB difference in outcome across studies is unclear, especially as the exclusion criteria between the normative study and the present one were different. Here, we did not exclude participants on the basis of occupational hazards such as exposure to noise. This could have led to some participants classified as NH exhibiting some form of hidden hearing loss. Thus, we do not recommend altering the 0 dB value put forward by the normative study [18].

Finally, the use of psychometric fits to the raw data may allow a finer description of the results that the summary SRT_{vrb} measure alone. In the present dataset, SRT_{vrb} and SRT_{fit} were highly correlated, but SRT_{fit} could provide estimates outside of the range of SNRs that were effectively tested. In future studies evaluating for instance the individual benefit from hearing aid use, yet other descriptors could be derived from the psychometric fits, such as the area under the curve. It may be fruitful to compare different characterizations of the full psychometric function to establish the most sensitive measure of hearing aid benefit.

Limitations

A first limitation of the present dataset, and of the VRB test itself in general, stems from the choice of a fixed stimulus paradigm. Because the range of test SNRs is constant for all listeners, floor or ceiling effects are inevitable. Using psychometric functions to estimate SRTs beyond the testing range may help, but the fits will likely become brittle as performance nears floor or ceiling. A related point can be made about the choice of keeping the speech level fixed throughout the procedure. For severe losses and in the unaided ear, this fixed level will lead to inaudible speech and thus poor performance even in quiet. However, it should be noted that such floor effects will mostly appear for severe forms of hearing losses, for which pure tone audiometry already demonstrates the need for a clinical intervention. For normal listeners and light losses, where the diagnosis based on audiometry may be insufficient, our results show that the VRB task is well calibrated, mostly away from floor and ceiling performance.

A second issue concerns the choice of varying the noise level, and not the speech level. The opposite choice is made in the FraMatrix test, another popular French speech-in-noise test [11, 12]. When tested in the same individuals, in a relatively modest sample, both methods have been found to converge on the same SRT, so this design choice may not be critical [37]. However, using a fixed speech level is arguably more representative of ecological situations: conversational speech only covers a restricted range of levels, so not all levels are physiologically plausible. Moreover, the realistic production of speech at different levels is accompanied by timbre changes, which are not captured by a simple level adjustment. This issue would still be worth revisiting in a large-scale comparison of the two kinds of tasks in the same individuals.

Relationship between Tonal Audiometry and Speech in Noise

In line with recent large-cohort investigations using the FraMatrix speech-in-noise test [12] or retrospective analyses of words-in-noise [27], our dataset confirmed a partial disconnect between pure tone audiometry and speech-in-noise performance. While there was a sizeable proportion of the variance in speech-in-noise performance that could be explained by pure tone audiometry at the cohort level, this proportion decreased dramatically within each subgroup. In particular, the proportion of explained SRT variance by PTA_m was very low for listeners who had borderline audiograms or light hearing losses. Intriguingly, we found no evidence that average pure tone thresholds at higher frequencies, PTA_{m_hf}, were correlated to speech-in-noise performance, except for the NH group (R^2 = 0.13). In apparent contrast, extended high-frequency audiometry has been put forward as a possible diagnosis of hidden hearing losses [38]. However, our high-frequency measures only included 6 kHz and 8 kHz, which is less than the range of 10–16 kHz investigated by Zadeh et al. [38] Thus, it remains a possibility that some of the speech-in-noise variance in our data could be explained by tonal thresholds beyond 8 kHz, which were not collected as part of the present study.

Correlation analyses were consistent with previous studies [12] and suggested that SRTs cannot be fully predicted by PTAs. However, a limitation is that, as in previous studies, only linear correlations between PTAs and SRTs were investigated. Nonlinear transformations of the data may provide better correlations. Moreover, there could be better ways to summarize tonal audiometry outcomes, such as data-driven frequency-weighting of tonal thresholds [39]. We would argue that a full investigation of the links between tonal audiometry and speech in noise requires a comprehensive approach, leveraging data analysis and even machine learning techniques to formally estimate the predictive power of pure tone audiometry about speech-in-noise performance. Such an undertaking is beyond the scope of the present study, but, thanks to the increasing availability of large datasets such as the one made publicly available with the present study, it becomes a realistic goal for future investigations.

Conclusion

Speech-in-noise performance was reported over a large cohort using the VRB method. Results showed that such a method was appropriate for fast and reliable estimates of SRTs, which provided additional information above and beyond tonal audiometry. This suggests that practical speech-in-noise tests such as VRB can contribute to redefining what is clinically accepted as normal hearing [6].

Acknowledgments

We would like to thank all clinicians from the Audilab centers and the Laboratoire Renard for collecting the data presented in this study.

Statement of Ethics

The study obtained ethical agreement from the "Comité de Protection des Personnes Sud-Est IV" (approval number: 2018-A02948-47 [ID-RCB]). Under French law (loi Jardé, "Recherche Impliquant la Personne Humaine"), and as confirmed by the ethical committee, the study qualified as category 3: non-interventional research in which all procedures and products are within clinical standard of care, without additional or unusual procedures of diagnosis, treatment, or supervision. Therefore, written informed consent was not required or collected. Potential participants were provided with a study information sheet and were informed that they could refuse to participate, without any consequence on their clinical care. Moreover, participants were informed that they could withdraw from the study at any time, without providing any reason.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

The funder provided support for Daniel Pressnitzer, but did not have any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Julie Bestel initiated the study and supervised the data collection. Julie Bestel and Daniel Pressnitzer analyzed the data and wrote the first draft of the manuscript. Julie Bestel, Daniel Pressnitzer, Mathieu Robier, Frédéric Rembaud, Christian Renard, François Leclercq, and Christophe Vincent contributed to the final writing of the manuscript.

Data Availability Statement

The full dataset is provided as online supplemental information (for all online suppl. material, see https://doi.org/10.1159/000537768). Further inquiries can be directed to the corresponding author.

References

1 Wilson RH. Adding speech-in-noise testing to your clinical protocol. Hear J. 2004;57(2):10. doi: 10.1097/01.hj.0000292386.54654.5d.
2 Carhart R, Tillman TW. Interaction of competing speech signals with hearing losses. Arch Otolaryngol. 1970;91(3):273–9. doi: 10.1001/archotol.1970.00770040379010.
3 Wilson RH, McArdle R. Speech signals used to evaluate functional status of the auditory system. J Rehabil Res Dev. 2005;42(4 Suppl 2):79–94. doi: 10.1682/jrrd.2005.06.0096.
4 Kochkin S. Consumers rate improvements sought in hearing instruments. Hearing Rev. 2002;9(11):18–22.
5 Davidson A, Marrone N, Wong B, Musiek F. Predicting hearing aid satisfaction in adults: a systematic review of speech-in-noise tests and other behavioral measures. Ear Hear. 2021;42(6):1485–98. doi: 10.1097/AUD.0000000000001051.
6 Joly C-A, Reynard P, Mezzi K, Bakhos D, Bergeron F, Bonnard D, et al. Guidelines of the French society of otorhinolaryngology-head and neck surgery (SFORL) and the French society of audiology (SFA) for speech-in-noise testing in adults. Eur Ann Otorhinolaryngol Head Neck Dis. 2022;139(1):21–7. doi: 10.1016/j.anorl.2021.05.005.
7 Reynard P, Lagacé J, Joly CA, Dodelé L, Veuillet E, Thai-Van H. Speech-in-Noise audiometry in adults: a review of the available tests for French speakers. Audiol Neurootol. 2022;27(3):185–99. doi: 10.1159/000518968.
8 Theunissen M, Swanepoel DW, Hanekom J. Sentence recognition in noise: variables in compilation and interpretation of tests. Int J Audiol. 2009;48(11):743–57. doi: 10.3109/14992020903082088.
9 Wilson RH, McArdle RA, Smith SL. An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss. J Speech Lang Hear Res. 2007;50(4):844–56. doi: 10.1044/1092-4388(2007/059.
10 Jansen S, Luts H, Wagener KC, Frachet B, Wouters J. The French digit triplet test: a hearing screening tool for speech intelligibility in noise. Int J Audiol. 2010;49(5):378–87. doi: 10.3109/14992020903431272.
11 Jansen S, Luts H, Wagener KC, Kollmeier B, Del Rio M, Dauman R, et al. Comparison of three types of French speech-in-noise tests: a multi-center study. Int J Audiol. 2012;51(3):164–73. doi: 10.3109/14992027.2011.633568.
12 Bestel J, Legris E, Rembaud F, Mom T, Galvin JJ. Speech understanding in diffuse steady noise in typically hearing and hard of hearing listeners. PLoS One. 2022;17(9):e0274435. doi: 10.1371/journal.pone.0274435.
13 Luts H, Boon E, Wable J, Wouters J. FIST: a French sentence test for speech intelligibility in noise. Int J Audiol. 2008;47(6):373–4. doi: 10.1080/14992020801887786.
14 Leclercq F, Renard C, Vincent C. Speech audiometry in noise: development of the French-language VRB (vocale rapide dans le bruit) test. Eur Ann Otorhinolaryngol Head Neck Dis. 2018;135(5):315–9. doi: 10.1016/j.anorl.2018.07.002.
15 James CJ, Laborde M-L, Algans C, Tartayre M, Cochard N, Fraysse B, et al. The French MBAA2 sentence recognition in noise test for cochlear implant users. Int J Audiol. 2023;62(4):304–11. doi: 10.1080/14992027.2022.2045368.
16 Buisson-Savin J, Reynard P, Bailly-Masson E, Joseph C, Joly CA, Boiteux C, et al. Adult normative data for the adaptation of the hearing in noise test in European French (HINT-5 Min). Healthc. 2022;10(7):1306. doi: 10.3390/healthcare10071306.
17 Killion MC, Niquette PA, Gudmundsen GI, Revit LJ, Banerjee S. Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. J Acoust Soc Am. 2004;116(4 Pt 1):2395–405. doi: 10.1121/1.1784440.
18 Decambron M, Leclercq F, Renard C, Vincent C. Speech audiometry in noise: SNR Loss per age-group in normal hearing subjects. Eur Ann Otorhinolaryngol Head Neck Dis. 2022;139(2):61–4. doi: 10.1016/j.anorl.2021.05.001.
19 Dubno JR, Dirks DD, Morgan DE. Effects of age and mild hearing loss on speech recognition in noise. J Acoust Soc Am. 1984;76(1):87–96. doi: 10.1121/1.391011.
20 Moore BCJ. Frequency selectivity and temporal resolution in normal and hearing-impaired listeners. Br J Audiol. 1985;19(3):189–201. doi: 10.3109/03005368509078973.
21 Lorenzi C, Gilbert G, Carn H, Garnier S, Moore BCJ. Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. Proc Natl Acad Sci USA. 2006;103(49):18866–9. doi: 10.1073/pnas.0607364103.
22 Moore BCJ. Cochlear hearing loss. John Wiley & Sons; 2007. Available from: http://books.google.com/books/content?id=II726We9fSgC&printsec=frontcover&img=1&zoom=1&edge=curl&source=gbs_api.

23 Oxenham AJ. Predicting the perceptual consequences of hidden hearing loss. Trends Hear. 2016;20:2331216516686768. doi: 10.1177/2331216516686768.
24 Plack CJ, Barker D, Prendergast G. Perceptual consequences of "hidden" hearing loss. Trends Hear. 2014;18(0):2331216514550621. doi: 10.1177/2331216514550621.
25 Liberman MC, Epstein MJ, Cleveland SS, Wang H, Maison SF. Toward a differential diagnosis of hidden hearing loss in humans. PLoS One. 2016;11(9):e0162726. doi: 10.1371/journal.pone.0162726.
26 Bharadwaj HM, Masud S, Mehraei G, Verhulst S, Shinn-Cunningham BG. Individual differences reveal correlates of hidden hearing deficits. J Neurosci. 2015;35(5):2161–72. doi: 10.1523/JNEUROSCI.3915-14.2015.
27 Grant KJ, Parthasarathy A, Vasilkov V, Caswell-Midwinter B, Freitas ME, de Gruttola V, et al. Predicting neural deficits in sensorineural hearing loss from word recognition scores. Sci Rep. 2022;12(1):8929. doi: 10.1038/s41598-022-13023-5.
28 Killion MC, Niquette PA. What can the pure-tone audiogram tell us about a patient's SNR loss? Hear J. 2000;53(3):46–8. doi: 10.1097/00025572-200003000-00006.
29 Cannizzaro E, Cannizzaro C, Plescia F, Martines F, Soleo L, Pira E, et al. Exposure to ototoxic agents and hearing loss: a review of current knowledge. Balance Commun. 2014;12(4):166–75. doi: 10.3109/21695717.2014.964939.
30 Dodelé L, Dodelé D. Le test d'Audiométrie Vocale en présence de Bruit de Dodelé. Audio Infos. 2007;110:70–4.
31 Lafon J. Le Test phonétique et la mesure de l'audition. Paris, Eindhoven: Dunod; 1964.
32 JASPTeam. JASP (Version 0.16.3). 2022. Available from: https://jasp-stats.org/.
33 Leensen MCJ, de Laat JAPM, Dreschler WA. Speech-in-noise screening tests by internet, Part 1: test evaluation for noise-induced hearing loss identification. Int J Audiol. 2011;50(11):823–34. doi: 10.3109/14992027.2011.595016.
34 Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation – a discussion and demonstration of basic features. PLoS One. 2019;14(7):e0219854. doi: 10.1371/journal.pone.0219854.
35 Hohmann V. Frequency analysis and synthesis using a Gammatone filterbank: ingenta Connect. Acta Acustica united with Acustica. 2002;88(3):433–42.
36 Agus TR, Suied C, Thorpe SJ, Pressnitzer D. Fast recognition of musical sounds based on timbre. J Acoust Soc Am. 2012;131(5):4124–33. doi: 10.1121/1.3701865.
37 Wilson RH, McArdle R. Speech-in-noise measures: variable versus fixed speech and noise levels. Int J Audiol. 2012;51(9):708–12. doi: 10.3109/14992027.2012.684407.
38 Motlagh Zadeh L, Silbert NH, Sternasty K, Swanepoel DW, Hunter LL, Moore DR. Extended high-frequency hearing enhances speech perception in noise. Proc Natl Acad Sci. 2019;116(47):23753–9. doi: 10.1073/pnas.1903315116.
39 Lopez-Poveda EA, Johannesen PT, Pérez-González P, Blanco JL, Kalluri S, Edwards B. Predictors of hearing-aid outcomes. Trends Hear. 2017;21:2331216517730526. doi: 10.1177/2331216517730526.