An automated speech-in-noise test for remote testing: development and preliminary evaluation

Alessia Paglialonga[a], Edoardo Maria Polo[b,c], Marco Zanet[b], Giulia Rocco[b], Toon van Waterschoot[d], Riccardo Barbieri[b]

[a] National Research Council of Italy (CNR), Institute of Electronics, Computer and Telecommunication Engineering (IEIIT), Milan, Italy

[b] Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Milan, Italy

[c] DIAG, Sapienza University of Rome, Italy

[d] KU Leuven, Department of Electrical Engineering (ESAT-STADIUS/ETC), Leuven, Belgium

Manuscript type: Research Article

Short Title: A novel automated speech-in-noise remote test

Journal: American Journal of Audiology (https://pubs.asha.org/journal/aja)

Abbreviations:

1U3D = 1 Up 3 Down

3AFC = 3 Alternative Forced-Choice

DF = Degrees of Freedom

HHIE-S = Hearing Handicap for the Elderly - Screening version

ROC = Receiver Operating Characteristic

SIN = Speech-in-noise

SRT = Speech Reception Threshold

VCV = Vowel-Consonant-Vowel

UA = Unscreened Adults

YA = Young Adults

Corresponding author:
Alessia Paglialonga
CNR, Consiglio Nazionale delle Ricerche
Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni (IEIIT)
Piazza Leonardo da Vinci, 32
I-20133 Milan, Italy
Phone: +39 02 2399 3343
Email: alessia.paglialonga@ieiit.cnr.it;

Abstract

Purpose: To develop and evaluate a novel, automated speech-in-noise test viable for widespread in situ and remote screening.

Method: Vowel-consonant-vowel sounds in a multiple-choice consonant discrimination task were used. Recordings from a professional male native English speaker were used. A novel adaptive staircase procedure was developed, based on the estimated intelligibility of stimuli rather than on theoretical binomial models. Test performance was assessed in a population of 26 normal hearing young adults (YA) and in 72 unscreened adults (UA), including native and non-native English listeners.

Results: The proposed test provided accurate estimates of the speech reception threshold (SRT) compared to a conventional adaptive procedure. Consistent outcomes were observed in YA in test/retest and in controlled/uncontrolled conditions and in UA in native and non-native listeners. The SRT increased with increasing age, hearing loss, and self-reported hearing handicap in UA. Test duration was similar in YA and UA irrespective of age and hearing loss. The test-retest repeatability of SRTs was high (Pearson correlation coefficient = 0.84) and the pass/fail outcomes of the test were reliable in repeated measures (Cohen's kappa = 0.8). The test was accurate in identifying ears with pure-tone thresholds >25 dB HL (accuracy = 0.82).

Conclusions: This study demonstrated the viability of the proposed test in subjects of varying language in terms of accuracy, reliability, and short test time. Further research is needed to validate the test in a larger population across a wider range of languages and hearing loss, and to identify optimal classification criteria for screening purposes.

Key words: hearing screening, non-native listeners, speech recognition.

Introduction

Hearing impairment has been ranked as the fourth leading contributor to years lived with disability worldwide and among the top leading causes of moderate-to-severe disability in older adults (World Health Organization, 2011; Wilson, Tucci, Merson & O'Donoghue, 2017). About half a billion people are affected by disabling hearing impairment, projected to rise to over 900 million by 2050 (World Health Organization, 2018). Overall, hearing impairment has an estimated global cost of $750 billion, including costs related to healthcare and support as well as costs related to loss of productivity, increased risk of cognitive impairment and dementia, and decrease in quality of life (reduced social participation, depression, loneliness, anger, and a lack of self-confidence) (Dalton, Cruickshanks, Klein, Klein, Wiley, & Nondahl 2003; Graydon, Waterworth, Miller, & Gunasekera, 2018; Olusanya, Neumann, & Saunders, 2014). Despite the significant burden at individual and societal level, hearing impairment is still frequently under-diagnosed, particularly in adults (Mick & Pichora-Fuller, 2016; Davis & Smith, 2013).

Early identification and management of hearing impairment are key to limit the effects of untreated hearing loss. Secondary prevention strategies based on periodic hearing screening and prompt treatment for disabling hearing loss are recommended in older adults to mitigate hearing problems and the related consequences (Wilson et al., 2017). Hearing screening can help identify individuals with hearing problems early. In fact, individuals with hearing impairment typically get used to the slow progression of hearing loss and tend to seek help very late or may even fail to seek help (Davis & Smith, 2013). This reluctance to seek help early is also related to the fact that hearing tests are typically not included in adults' routine health care examinations and that a gradual decrease in hearing ability is commonly considered an inevitable part of aging. Hearing screening can help fulfill this unmet need (Pronk, Kramer, Davis, Stephens, Smith, Thodi, et al., 2011; Nash, Cruickshanks, Huang, Klein, Klein, Nieto, & Tweed, 2013; Davis & Smith, 2013).

The value of speech-in-noise (SIN) tests for adult hearing screening is well known. SIN tests can support implementation of widespread hearing screening in adults and can be helpful to identify the real-life communication problems and to promote awareness (Humes, 2013; Killion & Niquette, 2000; Smits, Kapteyn, & Houtgast, 2004). Moreover, SIN tests can overcome some limitations of pure tone audiometry (e.g., need for experienced operator, high cost of audiometers, and need for low-noise environment) and can be implemented in an automated way on user interfaces and can be self-administered either locally, via hand-held devices or personal computers, or at a distance via web applications or smartphone apps (De Sousa, Swanepoel, Moore, & Smits, 2018; Paglialonga, Tognola, & Grandori, 2014; Paglialonga, Tognola, & Pinciroli, 2015). Remote testing, in particular, has gained increasing attention recently as a possible means to expand access to hearing testing, for example in underserved populations, also thanks to ubiquitous use of personal mobile devices and the related increase in popularity of smartphone apps (Bright & Pallawela, 2016; Yousuf Hussein, Swanepoel, de Jager, Myburgh, Eikelboom, & Hugo, 2016; Paglialonga, 2020).

Recently, different self-administered SIN tests have been successfully introduced for remote testing. For example, the online Speech Perception Test uses speech features recognition for consonant-vowel-consonant words to predict the audiogram and the expected outcomes for aided speech perception (Blamey, Blamey, & Saunders, 2015). The Earcheck and the Occupational Earcheck online tests measure the speech reception threshold (SRT), i.e. the signal-to-noise ratio (SNR) that corresponds to 50% intelligibility, for consonant-vowel-consonant words in stationary masking noise (Leensen, de Laat, Snik, & Dreschler, 2011). The digits in noise test in its various language versions and formats (telephone, online, and mobile) estimates the SRT by using sequences of three random digits in speech-shaped noise (e.g.: Smits et al., 2004; Smits, Goverts, & Festen, 2013; Watson, Kidd, Miller, Smits, & Humes, 2012; De Sousa et al., 2018). A common feature of these tests is that, due to the use of words or digits, they are language-dependent and need to be translated, adapted for psychometric performance, and validated whenever a new language version

has to be developed. For example, various language versions of the digits in noise test have been introduced, to be used in listeners who have basic knowledge of the language used in the test (Potgieter, Swanepoel, Myburgh, & Smits, 2019). In screening settings, the native language of subjects who take a self-administered test is typically unknown, especially in today's multicultural societies. Moreover, screening tests delivered via web applications or smartphone apps can potentially reach a large population that is probably scattered across countries and across native languages. Therefore, when language-dependent tests are used in screening settings in situ or at a distance in a population of unknown native language, disparities may occur and a portion of the population, including minorities, may be penalized due to decreased access to screening or due to biased/inaccurate results.

The long-term goal of this research is to develop and validate a novel automated SIN test for widespread hearing screening, i.e. viable for remote testing, accurate, reliable in repeated measures, and viable for use in listeners of unknown language,. The aim of this article is to present preliminary results on the validity and reliability of the proposed test in a group of young adults with normal hearing sensitivity and in an unscreened population of adults of various native languages.

Methods

Test stimuli

To limit the possible influence of native language on test performance, the test was developed using meaningless vowel-consonant-vowel (VCV) stimuli (e.g., ama, ata, asa). Stimuli were administered in a multiple-choice task in a way that effort to encode the meaning of stimuli was not required, and a consonant discrimination test was developed. Moreover, a multiple-choice recognition task is helpful as it enables automated, user operated test execution via an easy-to-use graphical user

interface and a pointing device or a touch-sensitive screen, for example for remote delivery (e.g., De Sousa et al., 2018; Paglialonga et al., 2014; Leensen, de Laat, Snik, & Dreschler, 2011). Moreover, as long as the number of alternatives is small, this kind of task can help limit possible anxiety, the perceived difficulty, and higher-level effort (short-term memory, reading speed), i.e. factors that are known to influence speech recognition performance, particularly in older adults. Specifically, a three-alternative forced-choice (3AFC) task was used as in previously developed VCV-based tests (e.g., Paglialonga et al., 2014) because it represents a trade-off between test complexity and psychometric performance (Leek, 2001). Moreover, to limit consonant discrimination difficulty (especially for older adults and for non-native listeners) the three alternatives are displayed on the screen based on a maximal opposition criterion, i.e. the two wrong consonants differ from the spoken one in manner, voicing and place of articulation (e.g., ata, aka, ama) (Gierut, 1989; Paglialonga, Grandori, & Tognola, 2013; Paglialonga et al., 2014; Vaez, Desgualdo-Pereira, & Paglialonga, 2014). During the test, subjects are asked to select their response among three alternatives by using a graphical user interface. The written transcriptions of VCVs (i.e., the target stimulus and the two 'wrong' alternatives) are displayed on the screen with a predetermined size of 9 cm (width) x 4 cm (height). The position of the target stimulus within the three alternatives is randomized at each stimulus presentation. The graphical user interface and the test were implemented in MATLAB (R2017a, MathWorks™).

Stimuli in the form of VCV (intervocalic consonants) can be helpful in adult screening because decreased consonant recognition performance is among the first clues of age-related hearing loss (Killion & Niquette, 2000). Moreover, VCV recognition is largely independent on semantics and, also, effort to encode the meaning of stimuli is not required, especially in a multiple-choice task that can be executed by individuals with limited knowledge of the spoken language, as far as they are familiar with the written transcription of stimuli. The combination of meaningless stimuli and a multiple-choice task can be helpful to reduce the involvement of higher-level processing centers

and, also, to limit the possible influence of the subjects' education, literacy, and native language on test outcomes (Mattys, Brooks, & Cooke, 2009; Cooke, Lecumberri, Scharenborg, & van Dommelen, 2010).

Speech stimuli were 12 spoken consonants (/b, d, f, g, k, l, m, n, p, r, s, t/) in the context of the vowel /a/ (e.g., aba, ada) recorded from a male professional native English speaker (Paglialonga et al., 2013; Paglialonga et al., 2014; Vaez et al., 2014). The use of English speech materials to develop a test for widespread use is supported by the fact that English is the top language by total number of speakers worldwide. Specifically, it is the third top language by number of native speakers, the top second language worldwide, and the most widely used language used in the Web (Eberhard, Simons, & Fennig, 2019; Internet World Stats, n.d.). Therefore, individuals who undergo a hearing screening test are likely to have had some previous experience with English speech and the written transcriptions of consonants. The spoken VCVs were single exemplars spoken in a sound-treated room with no prosodic accent and with constant pitch across the list. Stimuli were recorded in a professional recording studio by using a Neumann TLM 103 microphone, a SSL S4000 64 channels mixer, Motu HD 192 A/D converters (44,1 kHz, 16 bit), and a GENELEC 1025A control room monitor. The level of recordings was equalized within and across the sets to meet the equal speech level requirement as in the ISO 8253-3:2012 standard (International Organization for Standardization, 2012) and to guarantee equal average levels of the sets of recordings. The speech-shaped noise added to VCVs was generated by filtering a Gaussian white noise of amplitude equal to the average level of VCV recordings with the international long term average speech spectrum (Byrne, Dillon, Tran, Arlinger, Wilbraham, Cox, et al., 1994) and a low pass filter (cut-off 1.4 kHz, roll-off slope 100 dB/octave) and then by adding a noise floor (i.e., the same filtered noise attenuated by 15 dB), as suggested by Leensen, de Laat, Snik, & Dreschler (2011).

Development of a novel staircase procedure

The proposed test was implemented by using an up/down adaptive staircase (Leek, 2001; Levitt, 1971), a popular approach in automated multiple-choice tests. Specifically, a one-up/three-down (1U3D) staircase was used (i.e., the SNR is increased after one incorrect response and decreased after three correct responses) because it maximizes efficiency, convergence, and precision of 3AFC tasks (Schlauch & Rose, 1990, Shelton & Scarrow, 1984).

Conventional staircase procedures use pre-determined, equal upward steps (Δ_{up}) and downward steps (Δ_{down}) (e.g., ± 2 dB SNR) and converge after a certain number of reversals at the point on the psychometric curve in which the probability of a decrease in the presentation level equals the probability of an increase in the presentation level (Levitt, 1971; Treutwein, 1995). This is based on an underlying binomial distribution model for correct/incorrect responses and on the assumption that the probability of a correct response at any given presentation level (i.e., at any given SNR in SIN tests) is the same for all stimuli in the set (i.e., the assumption of homogenous intelligibility of stimuli across the set). Under this assumption, a balanced 1U3D staircase is assumed to target the point at 79.4% intelligibility and, if Δ_{up} and Δ_{down} are equal, it can be terminated after 20 reversals (Schlauch & Rose, 1990, Shelton & Scarrow, 1984). The SRT can be estimated as the average of the SNRs at the midpoints of the last 8 ascending runs (Garcia-Pérez, 1998; Paglialonga, Fiocchi, Parazzini, Ravazzani, & Tognola, 2011). However, stimuli in a given set of speech material (including VCV stimuli) are typically not homogeneous, including VCV stimuli. A typical approach to ensure homogeneity of stimuli across the set in a conventional staircase procedure is to equalize their intelligibility at the target point, i.e. where the staircase procedure is expected to converge, for example the point at 79.4% intelligibility in a 1U3D task.

We have developed a novel staircase procedure that, instead of using pre-determined, equal upward and downward steps on a set of stimuli equalized at the target point, determines Δ_{up} and Δ_{down} adaptively based on the estimated psychometric curves of stimuli (preliminary results were

presented in: Zanet, Polo, Rocco, Paglialonga, & Barbieri, 2019). To optimize convergence towards the SRT and limit the number of stimuli presented, Δ_{up} and Δ_{down} are determined by using an optimal ratio $\Delta_{down}/\Delta_{up} = 0.74$ so that, as suggested by Garcia-Pérez (1998), the procedure can be terminated after 12 reversals and the SRT can be estimated as the average of the SNRs at the midpoints of the last 4 ascending runs (Zanet et al., 2019). To implement the novel procedure, we estimated the theoretical psychometric curves of VCV recordings in the range from -50 to +20 dB SNR in 2 dB steps by computing the Short-Time Objective Intelligibility (STOI) measure (Taal, Hendriks, Heusdens, & Jensen, 2011) and then by fitting the average STOI values obtained over 100 simulated realizations of stimuli plus noise with a cumulative normal model (sigmoid function) (Lyregaard, 1997). The proposed SIN procedure starts at a comfortable level of +8 dB SNR from a VCV randomly selected from the set. Then, it adapts the intelligibility based on a 1U3D rule (i.e., decreased intelligibility after three correct responses, increased intelligibility after one incorrect response) by changing, concurrently, the VCV and the SNR using Δ_{up} and Δ_{down} that are based on the estimated intelligibility of the specific stimulus at the specific presentation SNR. At each step, the VCV presented and the order of the alternatives displayed on the screen are randomized.

Test evaluation in normal hearing young adults

The performance of the proposed SIN test (variable SNR step size with a ratio $\Delta_{down}/\Delta_{up} = 0.74$) was assessed and compared with a conventional staircase (fixed SNR step size with $\Delta_{down} = \Delta_{up} =$ 2 dB SNR, stimuli equalized at the target point) in a group of normal hearing, non-native young adults (YA) (N = 26 subjects; 11 males, 15 females; age range 23-26 years; mean 24.2 years, s.d. 0.59 years; native language: Italian; pure-tone thresholds < 20 dB HL in the range 500-8000 Hz, otologically normal as in the ISO 7029:2017 standard (International Organization for Standardization, 2017)). Participants in the YA group underwent SIN testing in one ear (the better ear was chosen in case of asymmetric pure-tone thresholds), first with the conventional staircase and then with the proposed one. Each SIN test was run twice (test and retest) and under two

different conditions (audiometer-controlled output levels and self-adjusted output levels), for a total of eight tests per subject. Testing was conducted in two separate sessions conditions (first in audiometer-controlled output levels and then in self-adjusted output levels) that were at least five days apart.

The tests were run on an Apple® Macbook Air® 13'' (OS X Yosemite version 10.10.5). In the audiometer-controlled condition (output levels = 60 dB HL) the laptop was connected to a clinical audiometer (Amplaid 177+, AmplifonTM) with TDH49 headphones and in the self-adjusted output levels condition the laptop was connected to Sony MDRZX110APW headphones and the output levels set by the tested subject at a comfortable level by using a volume control interface. Participants took part in experiments on a voluntary basis after reading and signing an informed consent form. The experimental protocol was approved by the Politecnico di Milano Research Ethical Committee (Opinion n. 2/2019).

Test evaluation in unscreened adults

To collect preliminary evidence on the validity and reliability of the proposed test in the target population, experiments were conducted on an unscreened population of adults (UA) with varying degrees of hearing sensitivity and varying native language. Participants were recruited and tested in the framework of local health screening initiatives organized by local not-for-profit citizens associations in various settings: at a university for senior citizens, at cultural and recreational places, and within health prevention and awareness events for the general public.

This study presents preliminary results from a group of 72 UA (25 males, 47 females; age range 24-89 years; mean 63.2 years, s.d.14.27 years) of varying native language (Italian: 55 subjects; English: 10 subjects; French: 2 subjects; German, Spanish, Filipino, Efik, and Igbo: 1 subject). Participants underwent pure-tone air-conduction hearing thresholds measurement at 0.5, 1, 2, and 4 kHz, the Hearing Handicap for the Elderly - Screening version (HHIE-S) questionnaire (Ventry &

Weinstein, 1983), and the proposed test using the same equipment as in the self-adjusted condition described above. As the aim was to assess the performance of the proposed SIN as a screening tool, no diagnostic assessment of the type of hearing loss was performed. The pure-tone average (PTA) was defined as the average of pure-tone thresholds measured at the tested frequencies. Participants were given the option to choose in which ear(s) to perform the test: 64 subjects performed the test in one ear and 8 in both ears, for a total of 80 ears. Moreover, for a preliminary evaluation of test reliability in UAs, a subgroup of participants (N = 21; 5 male, 16 female; age range 42-89 years; mean 69.2 years, s.d. 12.46 years; native language: Italian) performed the test twice: 21 in one ear and 1 in both ears, for a total of 22 ears.

Data analysis

In the first experiment (on YA), we analyzed the accuracy and reliability of SRTs and the test duration measured with the proposed SIN and with the conventional procedure in audiometer-controlled and self-adjusted output levels conditions. In the second experiment (on UA), we analyzed the distributions of age, HHIE-S scores, PTA, and test performance (SRT, number of stimuli, test duration, and percentage of correct responses) in two subgroups of ears classified based on pure-tone thresholds ($UA_{\leq 25}$: thresholds better than or equal to 25 dB HL at 1, 2, and 4 kHz; $UA_{>25}$: thresholds higher than 25 dB HL at one or more of these frequencies (American Speech-Language-Hearing Association, n.d.)). In the subgroup of participants who performed the test twice, we analyzed the absolute test-retest differences in SRT, number of stimuli, test duration, and percentage of correct responses, and the Pearson correlation coefficient between SRTs measured in test and retest trials. For a preliminary evaluation of possible differences between native and non-native listeners, we compared the SRTs measured in native English listeners (UA_{EN}) and in a subgroup of non-native English listeners ($UA_{non\text{-}EN}$) matched for PTA (difference $\leq$ 5 dB HL) and age (difference $\leq$ 5 years).

Possible differences in the measured variables were assessed by parametric statistical analysis (t-tests with Bonferroni correction) if the variables were continuous and the distributions were normal (as assessed by the Shapiro-Wilk test) and by nonparametric statistical analysis (Wilcoxon rank sum test/signed rank test with Bonferroni correction) otherwise. A generalized linear model with age and PTA as predictor variables was used to investigate the possible factors influencing the SRT. The significance level α was set at 0.05.

For a preliminary evaluation of the proposed test as a possible hearing screening test, we analyzed the receiver operating characteristic (ROC). We systematically varied the cut-off SRT from -19 to 10 dB SNR, in 2 dB steps, and for each cut-off SRT we computed the accuracy, sensitivity, and specificity of the SIN test (*pass*: SRT < cut-off SRT; *fail*: SRT ≥ cut-off SRT) against the above criterion for classification of ears into $UA_{\leq 25}$ and $UA_{>25}$. We measured the area under the curve and the 95% confidence interval. We identified three candidate cut-off SRTs by selecting the points closer to the corner (0,1). For each of the candidate cut-off SRTs, we estimated test reliability of test outcomes (pass/fail) from the test-retest data available by computing the Cohen's kappa (*k*), a measure of repeatability for binary outcomes (Cohen, 1960).

Data analysis was implemented using MATLAB (R2017a, MathWorks™).

Results

Test performance in normal hearing young adults

Table 1 shows that the SRTs measured by the conventional and the proposed procedure in the test trials in audiometer-controlled condition were similar. There were no statistically significant differences in mean SRT between the two procedures (t-test, DF = 25: p = 0.89). The observed test-retest differences were slightly higher for the conventional procedure than for the proposed SIN

(i.e., -1.58 vs 0.19 dB SNR in audiometer-controlled conditions and -1.31 vs -0.43 dB SNR in self-adjusted output levels conditions). The observed test-retest differences were significant for the conventional procedure in audiometer controlled conditions (paired samples t-test, DF = 25: $p = 0.0015$) and in self-adjusted output levels conditions (Wilcoxon signed rank test: $p = 0.011$). No statistically significant test-retest differences were observed for the proposed procedure (audiometer controlled conditions, paired samples t-test, DF = 25: $p = 0.7$; self-adjusted output levels conditions: paired samples t-test, DF = 25: $p = 0.38$). The difference in SRTs measured by the proposed procedure in audiometer-controlled and in self-adjusted output levels conditions was small and not significant (mean difference < 0.5 dB SNR; paired samples t-test, DF = 25: $p = 0.83$).

The mean test duration of the proposed SIN was lower than the conventional procedure (i.e., about 3 minutes and 50 seconds vs about 5 minutes and 40 seconds) and this difference was statistically significant (t-test, DF = 25: $p << 0.05$). For both procedures, the test duration was similar in the test and retest trials (on average, the test-retest difference in test duration was 12 s with the conventional procedure and 3 s with the proposed SIN in YA).

Preliminary evaluation in an unscreened population of adults

As shown in Table 2, subjects in the $UA_{>25}$ group were older (t-test, DF = 78: $p << 0.05$) and reported higher hearing handicap (Wilcoxon rank sum test: $p = 0.02$) than those in the $UA_{\leq 25}$ group. Specifically, in the $UA_{\leq 25}$ group, 24 subjects reported no hearing handicap (score ≤ 8), 8 reported mild-to-moderate handicap ($8<score<24$), and only 2 reported severe hearing handicap (score ≥ 24). In the $UA_{>25}$ group, 20 subjects reported no hearing handicap, 13 reported mild-to-moderate handicap, and 5 reported severe hearing handicap.

The SRTs measured in the $UA_{\leq 25}$ group were, on average, higher than in the YA group (about 3.7 dB higher) and SRTs measured in the $UA_{>25}$ group were higher than in the $UA_{\leq 25}$ group (about 6.4 dB higher). All the observed differences in mean SRTs were statistically significant (t-tests: $UA_{\leq 25}$

vs YA, DF = 61: $p << 0.05$; $UA_{>25}$ vs $UA_{\leq 25}$, DF = 78: $p << 0.05$; $UA_{>25}$ vs YA, DF = 67: $p << 0.05$). The Table also shows that the standard deviation of SRTs tended to increase from the YA to the $UA_{\leq 25}$ and $UA_{>25}$ groups, suggesting an increasing variability of SRT estimates across the groups. In addition, the total number of stimuli presented in the test and the percentage of correct responses decreased across the groups (from YA to $UA_{\leq 25}$ to $UA_{>25}$). Therefore, the higher (i.e., worse) the SRT, the lower the percent recognition performance, and the lower the number of stimuli required by the adaptive procedure to reach the SRT. The observed across-group decrease in total number of stimuli and the decrease in percentage of correct responses were statistically significant (t-tests for #stimuli, $UA_{\leq 25}$ vs YA, DF = 61: $p << 0.05$; $UA_{>25}$ vs $UA_{\leq 25}$, DF = 78: $p << 0.05$; $UA_{>25}$ vs YA, DF = 67: $p << 0.05$; Wilcoxon rank sum test for the percentage of correct responses, $UA_{\leq 25}$ vs YA: $p << 0.05$; $UA_{>25}$ vs $UA_{\leq 25}$: $p << 0.05$; $UA_{>25}$ vs YA: $p << 0.05$). The mean test duration was similar across the three groups (about 3'50'') indicating that, despite the fact that less stimuli were presented in the UA groups, these subjects required nearly the same amount of time to complete the test as subjects in the YA group. In fact, differences in test duration across the three groups were not statistically significant (Wilcoxon rank sum test, $UA_{\leq 25}$ vs YA: $p = 0.95$; $UA_{>25}$ vs $UA_{\leq 25}$: $p = 0.38$; $UA_{>25}$ vs YA: $p = 0.49$).

Figure 1 shows the scatterplot of SRT and PTA measured across the three groups (YA, $UA_{\leq 25}$, $UA_{>25}$) and the resulting linear regression fit. Within each group, the individual SRTs tended to increase (on average, an increase in SRT of about 2.6 dB per 10 dB increase in PTA) and to be more scattered with increasing PTA, in line with the increase in inter-individual variability across groups shown in Table 2. The figure also show that the relationship between SRT and PTA in native listeners (UA_{EN}) was in line with the overall trend in SRTs as a function of PTA observed in the whole sample. The mean SRT measured in UA_{EN} was -9.1 dB SNR (s.d. 5.22; range: -17.25÷4.50) and was similar to the mean SRT measured in the age- and PTA-matched ($UA_{non\text{-}EN}$) subgroup (mean: -9.2 dB SNR (s.d. 5.84; range: -17.75÷1.75). The observed differences in median

SRT between the two groups were not significant (Wilcoxon signed rank test: $p = 0.81$). The mean difference in age between UA_{EN} and $UA_{non\text{-}EN}$ subjects was 2 years (s.d. 1.6; range 0÷4) and the mean difference in PTA was 1.67 dB HL (s.d. 1.44; range 0÷3.75 dB HL).

Figure 2 shows the scatterplot of SRT and age across the three groups (YA, $UA_{\leq 25}$, $UA_{>25}$) and the resulting linear regression fit. There was a clear overlap in age between the $UA_{\leq 25}$ and $UA_{>25}$ groups and participants in the $UA_{>25}$ group tended to have higher age and higher SRT. Overall, the SRT increased as a function of age (on average, an increase in SRT of about 2 dB per decade), especially within the $UA_{>25}$ group. As the distribution of SRTs was not normal (Shapiro-Wilk test: $p << 0.05$), the generalized linear model was computed following transformation of SRTs values using a logarithmic mapping function. The generalized linear model ($F(102) = 37.9$; $p << 0.05$) showed that neither PTA nor age alone were significant predictors of SRTs ($p = 0.36$ and $p = 0.25$, respectively) whereas the interaction between age and PTA was a significant predictor of SRTs ($p = 0.01$).

Table 3 shows preliminary test-retest results obtained from the 21 UA participants who underwent the proposed SIN test twice ($UA_{t/r}$: N = 22 ears). Of these 22 ears, 7 were in the $UA_{>25}$ class and 15 were in the $UA_{>25}$ class. Overall, participants in the $UA_{t/r}$ subgroup showed varying degrees of hearing sensitivity (from normal hearing to moderate hearing loss) and a large range of self-reported hearing handicap (from no handicap to severe handicap). The Pearson correlation coefficient for SRTs measured in test and retest trials was high (i.e., 0.84). There was no statistically significant differences between test and retest SRTs (Wilcoxon signed rank test: $p = 0.70$). The observed absolute variation in $UA_{t/r}$ subjects shown in Table 3 was higher than the absolute variation observed in the YA group (mean: 1.70 dB, s.d. 1.5 dB, range 0.25÷7.00 dB). On average, the absolute variations in #stimuli and percentage of correct responses were limited, suggesting that the proposed algorithm for the 1U3D adaptive procedure produced consistent patterns in repeated measurements in UA participants. Test duration showed a mean absolute variation of less than one minute in the $UA_{t/r}$ group with individual variations up to about two minutes and a half with the

retest trial being, on average, longer than the test trial (mean increase in test duration equal to about half a minute).

Figure 3 shows the ROC obtained from the whole study sample (N = 106 ears from 98 subjects). The area under the curve, representing the probability that the SRT of a randomly chosen ear in the $UA_{>25}$ group will be ranked higher than the SRT of a randomly chosen ear in the $UA_{\leq 25}$ group, was 0.84 (95% confidence interval: 0.75-0.94), suggesting good overall classification performance. The three candidate cut-off SRTs, representing a trade-off between sensitivity and specificity, were: -11.75 dB SNR (accuracy: 0.73; sensitivity = 0.79; specificity = 0.68; k = 0.70), -10 dB SNR (accuracy: 0.79; sensitivity = 0.77; specificity = 0.81; k = 0.80), and -8 dB SNR (accuracy: 0.82; sensitivity = 0.70; specificity = 0.90; k = 0.72).

Discussion

As a first step to develop an accurate and reliable screening test for use in listeners of unknown language, a novel SIN test has been designed. The proposed test is based on recognition of VCV stimuli in 3AFC format and uses a novel staircase procedure that introduces adaptive upward and downward steps that are based on estimated VCV psychometric curves (Zanet et al., 2019).

Test performance

Evaluation of test performance in a group of normal hearing young adults (YA) showed that the proposed test was accurate (i.e., as accurate as conventional staircase procedures) and reliable in repeated measures (i.e., it provided consistent results in test and retest trials) and, also, that it required a shorter test time (i.e., about 2 minutes shorter) than a conventional staircase (Zanet et al., 2019). Thus, results from YA participants suggested that the newly developed staircase procedure may be suitable for applications in which fast and reliable methods are needed, for example in adult hearing screening. Moreover, the novel test was able to provide reliable estimates of SRTs in

audiometer-controlled and in self-adjusted output levels conditions, yet requiring a shorter time compared to a conventional procedure. Accordingly, the proposed test may be viable for use in uncontrolled environments (either locally or for remote testing) provided that subjects are instructed to adjust the output volume at a comfortable level (Zanet et al., 2019). This is further supported by results from a recent study that showed, by using laboratory measurements of sound pressure levels with a range of consumer transducers, that giving adjusting the output level through the laptop volume can help compensate for the different transducer characteristics, thus further supporting the viability of the test to be used for remote testing.

To further evaluate the feasibility of the proposed test as a hearing screening test for adults, we performed a preliminary study in an unscreened population of adults (UA) of varying native language. Results showed that the SRTs measured in $UA_{>25}$ ears were significantly higher (i.e., worse) than those measured in $UA_{\leq 25}$ ears and that, in turn, SRTs in $UA_{\leq 25}$ ears were significantly higher than in YA ears (Table 2). Considering the group characteristics and the SRT distributions reported in Table 2 for the YA and UA groups, these results suggested that the higher the pure-tone thresholds, age, and self-reported hearing handicap, the higher the SRT and its inter-individual standard deviation. Interestingly, despite a decrease in speech recognition performance and an increase in age from YA to $UA_{\leq 25}$ to $UA_{>25}$, test duration in UA was similar as in YA. This is due to the fact that adaptive procedures such as the one here developed are able to compensate for the possibly longer response time and the possibly higher effort incurred by older hearing impaired subjects because the total number of stimuli required to estimate the SRT decreases with decreasing performance, thus keeping the test duration approximately constant (Table 2). The observed decrease in the number of stimuli can be explained mainly by a shorter initial phase of the 1U3D adaptive procedure in subjects with poorer speech recognition performance. Specifically, in adaptive procedures that start from well above threshold levels (such as the +8 dB SNR level used in this study), subjects with better speech recognition typically go through several correct responses

(i.e., several 3D steps) before incurring into the first reversal (i.e., an error leading to a 1U step), thus leading to an overall higher number of stimuli presented. Vice versa, subjects with poorer speech recognition typically go through fewer 3D steps before incurring in the first error and they tend to incur in the first reversal more rapidly than subjects with better speech recognition.

Influence of hearing sensitivity and age on SRTs

The decrease in SRT and the related increase in variability with increasing hearing thresholds, as shown in Figure 1 and Table 2, are in line with data from the literature. In fact, although threshold detection of pure-tones and suprathreshold speech recognition relate to two inherently different mechanisms, a significant relationship between pure-tone audiograms and SRTs can be observed across a range of pure-tone thresholds in adults (Bosman & Smoorenburg, 1995; Smoorenburg, 1992; Beattie, 1989). The dominant source of reduced speech recognition in adults with mild-to-moderate hearing loss, such as the majority of UA participants in this study, is known to be related to the combined effect of hearing loss and masking noise that reduce the audibility of speech, although suprathreshold deficits also play a role (Zurek & Delhorne, 1987). In fact, suprathreshold deficits in addition to audibility can play a considerable role in speech recognition in noise in adults, even at levels well above threshold such as the ones used in this study in the self-adjusted output levels condition. The observed increase in SRT across the groups, from YA to $UA_{\leq 25}$ to $UA_{>25}$, has been shown to be mainly related to the concurrent increase of PTA and age (Figure 1 and 2). This is in line with the well-known effects of hearing loss, auditory and cognitive processing, and age on speech intelligibility in noise (e.g., Ching & Dillon, 2013; Humes, Kidd, & Lentz, 2013; Summers, Makashay, Theodoroff, & Leek, 2010; Nuesse, Steenken, Neher, & Holube, 2018). Moreover, our results clearly indicated that age alone, as well as PTA alone, were not significant predictor of SRTs but that their interaction was a significant factor. This is in line with various studies that demonstrated a decline in auditory processing and cognitive abilities with age as well as an interaction between age, cognition, and hearing loss in a way that speech recognition performance is

the result of complex interactions between these factors (Füllgrabe, Moore, & Stone, 2015; Füllgrabe, 2013; Smith & Pichora-Fuller, 2015).

Inter-individual and intra-individual variability

Increased SRT variability was observed in the study sample with increasing pure-tone thresholds and increasing age (Figure 1 and Figure 2). All ears in the YA group were in the normal hearing range. Ears in the $UA_{\leq 25}$ group were in the normal, slight-, and mild-hearing loss range and subjects in this group were older than in the YA group. Ears in the $UA_{>25}$ group were in the mild-, moderate-, and moderately severe-range (following criteria by Clark, 1981) and subjects in this group were, on average, older than subjects in the other two groups. This is in line with previous studies that reported increased inter-individual variability of SRTs for various types of speech material in subjects with hearing loss compared to subjects with normal hearing (Leensen, de Laat, & Dreschler, 2011; Nielsen & Dau, 2011; Summers, Makashay, Theodoroff, & Leek, 2010) as well as increased inter-individual variability of SRTs in older adults with hearing impairment compared to those with age appropriate hearing (Nuesse et al., 2018).

Regarding intra-individual variability, as measured in test-retest experiments, it is typically influenced by learning in subsequent trials. In general, test results in retest trials are likely to improve compared to the first trial as the participant is better able to deal with the test stimuli and presentation mode, is more familiar with the interface and task, and may become more able to separate the auditory characteristics of speech from noise – in fact, some auditory training programs use repeated SIN exercises to improve listening in noise (Song, Skoe, Banai, & Kraus, 2012). However, in this study no significant learning effect was observed, the mean SRT measured by the proposed SIN was stable across conditions, and the average change in SRT in the retest trial compared to the test trial was small (-0.35 dB) and not statistically significant. The test-retest variability in SRTs in the $UA_{t/r}$ group was higher than in the YA group (on average, the absolute SRT variation was 1.7 and 3.47 dB in the YA and $UA_{t/r}$ groups with standard deviations of 1.50 and

2.87, respectively, Table 3). Previous studies have reported intra-individual standard deviation of test-retest SRTs in the range 0.4÷1.2 dB in normal hearing subjects and in the range 0.8÷2.0 dB in hearing impaired subjects for sentence-, words-, and digits-based SIN tests (Jansen, Luts, Wagener, Kollmeier, Del Rio, Dauman, et al., 2012; Jansen, Luts, Wagener, Frachet, & Wouters, 2010; Killion, Niquette, Gudmundsen, Revit, & Banerjee, 2004; Nielsen & Dau 2011; Semeraro, Rowan, van Besouw, & Allsopp, 2017; Sheikh Rashid & Dreschler, 2018) although a benchmark for multiple-choice recognition of VCV is not available. It may be that the higher intra-individual variability here observed was due, at least in part, to the initial design process as VCV recordings were selected based on a combination of criteria that included, among others, the requirement of lower slope for the psychometric function that is, in turn, related to higher variability (Strasburger, 2001). In addition, the use of a 3AFC task may be another possible source of performance variability as the slope is lower in multiple-choice recognition tasks compared to open set recognition (Klein, 2001). However, in this study the correlation between SRTs obtained in test and retest trials in the $UA_{t/r}$ group was high (i.e., 0.84) and the test-retest reliability for classification of ears into pass and fail, as measured by Cohen's kappa (*k*), ranged from 0.70 to 0.80 for the three cut-off SRTs here investigated, suggesting substantial agreement of test outcomes were reliable repeated measures (Cohen, 1960).

Classification performance

Despite the statistically significant increase in mean SRTs observed in the $UA_{>25}$ group compared to the $UA_{\leq 25}$ group, results in Table 2 and Figure 1 show that the distributions overlapped, also due to the related increase in inter-individual variability. Therefore, for the sake of classifying subjects into *pass* and *fail* based on the outcomes from the proposed test, a trade-off is necessary as shown by the ROC analysis in Figure 3. The test reached an overall accuracy of up to 0.82 and an area under the curve equal to 0.84, suggesting that the test is moderately accurate (Fischer, Bachmann, & Jaeschke, 2003). This moderate level of accuracy is related to the overlap between the UA

subgroups and, also, to the inherently different nature of the measures compared (tone detection and speech recognition) so that a true gold standard for SIN test evaluation is not available within this study. It is to note that, in general, the sensitivity and specificity of SIN screening tests can vary greatly depending on the type of stimuli, the testing procedure, and the cut-off criteria. The performance of the SIN test here presented is in line with the performance documented for speech recognition tests using similar approaches, i.e. multiple choice recognition of short words. For example, Leensen, de Laat, & Dreschler (2011) investigated the accuracy of the Earcheck and the Occupational Earcheck, i.e. internet-based adaptive SIN tests based on a closed set of eight equally intelligible Dutch consonant-vowel-consonant words in a nine-alternative multiple-choice task. Both tests used a cut-off SRT equal to -10 dB SNR and yielded sensitivity of 0.51 and 0.92 and a specificity of 0.90 and 0.49, respectively to detect ears with noise induced hearing loss, with test-retest reliability, as measured by the intra-class correlation coefficient, equal to 0.75 and 0.68, respectively, i.e. lower than the ones here observed for the same cut-off SRT (k = 0.80). Recently, Sheikh Rashid & Dreschler (2018) investigated the accuracy of the Occupational Earcheck in an occupationally noise-exposed population. They found that, by using a cut-off SRT of -14.9 dB SNR, the test had 0.65 sensitivity and 0.63 specificity to detect high-frequency hearing loss above 25 dB HL and that in the older age group the sensitivity was 0.69 and sensitivity was 0.46. When a second round of conditional rescreening was added, the sensitivity and specificity of the Occupational Earcheck were 0.65 and 0.92 in the older age group, i.e. similar to the ones we observed with a cut-off SRT of -8 dB SNR. A fixed-levels screening test based on 3AFC recognition of a list of 12 VCVs presented at predetermined SNRs (the SUN, Speech Understanding in Noise test) administered sequentially in both ears reached about 0.85 sensitivity and specificity for detecting disabling hearing impairment (i.e. PTA > 40 dB) when a cut-off score of 6 out of 12 correctly identified VCVs was set and results from both ears were combined (Paglialonga et al., 2014). The original version of the digits in noise test by telephone used a cut-off SRT equal to -4.1 dB SNR (Smits et al., 2004), i.e., higher than the candidate cut-off SRTs identified in Figure 3, due

to the inherently different task (open choice vs forced choice) and speech material (digit triplets vs VCVs). The test yielded a sensitivity of 0.75 and specificity of 0.91 to identify ears with PTA (computed at 0.5, 1, and 2 kHz) higher than 20.6 dB HL, i.e. values comparable to the ones we observed with a cut-off SRT of -8 dB SNR (sensitivity 0.70, specificity = 0.90). Similarly, the US version of the digits in noise test obtained, using a cut-off SRT of -5.7 dB SNR, sensitivity and specificity of 0.80 and 0.83 to identify ears with PTA higher than 20 dB HL (Watson et al., 2012), i.e. values close to the ones we obtained in this preliminary study with a cut-off SRT of -10 dB SNR (sensitivity 0.77, specificity 0.81).

Limitations and future research

It is acknowledged that this study has some limitations. Due to the preliminary nature of data collected from UA, a comprehensive assessment of the possible influence of native language on test performance was not possible. In this preliminary study, we observed no noticeable differences in test performance between 12 native listeners and 12 non-native listeners who were matched for hearing sensitivity and age. It is important that this finding be further investigated in future studies by recruiting a larger population across a range of native languages. In addition, the proposed study did not follow a comprehensive protocol for hearing assessment and SIN testing in all the tested subjects due to the specific settings in which the measurements were conducted in the UA group (i.e., at opportunistic hearing screening initiatives and health prevention and awareness events). Similarly, test-retest reliability was addressed in a relatively small sample (26 YA and 21 UA) and only in one ear, mainly on the basis of the measured SRTs and test performance. It would be important to evaluate the test-retest reliability in a larger sample and by analyzing the patterns of responses in more detail (e.g. the confusion matrices, the individual responses and reaction times) for a better understanding of the possible factors influencing test performance and the magnitude of possible learning effects in repeated measures. Further research is needed, by using a comprehensive protocol for hearing assessment and SIN testing to fully characterize the

performance of the proposed test in adults, and to analyze its accuracy and reliability compared to well-established screening and diagnostic speech-based tests. It will also be important to recruit a bigger population of unscreened adults to cover a wider range of hearing sensitivity, types of hearing loss (including sensorineural, conductive, and mixed), and native languages in order to investigate the effectiveness of the proposed test as a screening tool and to measure the influence of native language on test performance. Moreover, it will be important to address the performance of the test in listeners of characters-based and non-roman alphabet-based languages to evaluate the influence of the alphabet on test outcomes. In addition, it will be useful to investigate new, more complex classifiers that, in addition to the estimated SRT, take into account further parameters (e.g., percentage of correct responses, age, or detailed pattern of responses including the confusion matrix) to build more accurate and robust models of the population for the sake of hearing screening. Finally, more experiments with web-based or mobile-based versions of the test on adults with normal hearing and hearing impairment are needed to get a deeper insight into the feasibility of the test for remote delivery.

Conclusions

The main outcome of this study is to demonstrate, for the first time, the viability of a novel SIN test that can be administered to subjects of unknown language, that is reliable in repeated measures, and that can be available for use in uncontrolled settings and for remote testing. In general, SIN tests such as the one proposed here can be useful as a preventive measure by identifying individuals with reduced speech recognition abilities and help them gain awareness on the importance of having a full hearing assessment and consult an audiologist, thus contributing to early identification and management of hearing impairment. Further research is needed to fully validate the test, to determine its classification performance for screening purposes, and to assess the feasibility of the test for remote delivery.

Acknowledgments

Portions of this paper were presented at the 4rd International Internet & Audiology Meeting, Southampton, UK.

The Authors are grateful to the Lions Clubs International and to Associazione La Rotonda, Baranzate (MI) for their support in the organization and management of experiments in the unscreened population of adults. The Authors wish to thank Anna Bersani, Carola Butera, and Antonio Carrella who helped with data collection at Associazione La Rotonda.

The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

References

American Speech-Language-Hearing Association (n.d.). Adult Hearing Screening (Practice Portal). Retrieved from https://www.asha.org/PRPSpecificTopic.aspx?folderid=8589942721&section=Overview

Beattie, R.C. (1989). Word recognition functions for the CID W-22 test in multitalker noise for normally hearing and hearing-impaired subjects. *Journal of Speech and Hearing Disorders, 54(1)*, 20-32.

Blamey, P., Blamey, J., & Saunders E. (2015). Effectiveness of a Teleaudiology approach to hearing aid fitting. *Journal of Telemedicine and Telecare, 2(8)*, 474-478. https://doi.org/10.1177/1357633X15611568

Bright, T., & Pallawela, D. (2016). Validated smartphone-based apps for ear and hearing assessments: a review. JMIR *Rehabilitation and Assistive Technology, 3(2),* e13. doi: 10.2196/rehab.6074

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., et al. (1994). An international comparison of long-term average speech spectra. *Journal of the Acoustical Society of America, 96,* 2108–2120. https://doi.org/10.1121/1.410152

Bosman, A.J., & Smoorenburg, G.F. (1995). Intelligibility of Dutch CVC syllables and sentences for listeners with normal hearing and with three types of hearing impairment. *Audiology, 34(5),* 260-284.

Ching, T.Y. & Dillon, H. (2013). A brief overview of factors affecting speech intelligibility of people with hearing loss: Implications for amplification. *American Journal of Audiology, 22,* 306-309. https://doi.org/10.1044/1059-0889(2013/12-0075)

Clark, J. G. (1981). Uses and abuses of hearing loss classification. *Asha, 23,* 493-500.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement,* 20, 37-46. https://doi.org/10.1177/001316446002000104

Cooke, M., Lecumberri, M.L.G., Scharenborg, O., & van Dommelen, W.A. (2010). Language-independent processing in speech perception: Identification of English intervocalic consonants by speakers of eight European languages. *Speech Communication, 52,* 954–967. https://doi.org/10.1016/j.specom.2010.04.004

Dalton, D.S., Cruickshanks, K.J., Klein, B.E., Klein, R., Wiley, T.L., & Nondahl, D.M. (2003). The impact of hearing loss on quality of life in older adults. *Gerontologist, 43(5),* 661-668. https://doi.org/10.1093/geront/43.5.661

Davis, A., & Smith, P. (2013). Adult hearing screening: Health policy issues-What happens next? *American Journal of Audiology, 22,* 122–125. https://doi.org/10.1044/1059-0889(2013/12-0062)

De Sousa, K.C., Swanepoel, D.W., Moore, D.R., & Smits, C. (2018). A Smartphone National Hearing Test: Performance and Characteristics of Users. *American Journal of Audiology, 27(3S),* 448-454. https://doi.org/10.1044/2018_AJA-IMIA3-18-0016

Eberhard, D.M., Simons, G.F., & Fennig, C.D. (2019). (eds.). Ethnologue: Languages of the World. Twenty-second edition. Dallas, Texas: SIL International. Retrieved from http://www.ethnologue.com

Fischer, J.E., Bachmann, L.M. & Jaeschke, R. (2003). A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Medicine, 29,* 1043. https://doi.org/10.1007/s00134-003-1761-8

Füllgrabe, C. (2013). Age-dependent changes in temporal-fine-structure processing in the absence of peripheral hearing loss. *American Journal of Audiology, 22,* 313-315. https://doi.org/10.1044/1059-0889(2013/12-0070)

Füllgrabe, C., Moore, B.C.J., & Stone, M.A. (2015). Age-group differences in speech identification despite matched audiometrically normal hearing: contributions from auditory temporal processing and cognition. *Frontiers in Aging Neuroscience, 6,* 347. https://doi.org/10.3389/fnagi.2014.00347

García-Pérez, M.A. (1998). Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Research, 38(12),* 1861-1881.

Gierut, J.A. (1989). Maximal opposition approach to phonological treatment. *Journal of Speech and Hearing Disorders, 54(1),* 9-19. https://doi.org/10.1044/jshd.5401.09

Graydon, K., Waterworth, C., Miller, H., & Gunasekera H. (2018). Global burden of hearing impairment and ear disease. *Journal of Laryngology and Otology, 133(Special Issue 1),* 18–25. https://doi.org/10.1017/S0022215118001275

Humes, L.E. (2013). Understanding the speech-understanding problems of older adults. *American Journal of Audiology, 22,* 303–305. https://doi.org/10.1044/1059-0889(2013/12-0066).

Humes, L.E., Kidd, G.R., & Lentz, J.J. (2013). Auditory and cognitive factors underlying individual differences in aided speech-understanding among older adults. *Frontiers in Systems Neuroscience, 7*, 55. https://doi.org/10.3389/fnsys.2013.00055

International Organization for Standardization (2017). ISO 7029:2017. Acoustics Statistical distribution of hearing thresholds related to age and gender. Standard.

International Organization for Standardization (2012). ISO 8253-3:2012. Acoustics audiometric test methods part 3: Speech audiometry. Standard.

Internet World Stats. (n.d.) Internet world users by language: top 10 languages. Retrieved from https://www.internetworldstats.com/stats7.htm

Jansen, S., Luts, H., Wagener, K.C., Frachet, B., & Wouters, J. (2010). The French digit triplet test: a hearing screening tool for speech intelligibility in noise. *International Journal of Audiology, 49(5),* 378-387. https://doi.org/10.3109/14992020903431272

Jansen, S., Luts, H., Wagener, K.C., Kollmeier, B., Del Rio, M., Dauman, R., James, C., Fraysse, B., Vormès, E., Frachet, B., Wouters, J., & van Wieringen A. (2012). Comparison of three types of French speech-in-noise tests: a multi-center study. *International Journal of Audiology, 51(3),* 164-173. https://doi.org/10.3109/14992027.2011.633568

Killion, M.C., Niquette, P.A., Gudmundsen, G.I., Revit, L.J., & Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America, 116(4 Pt 1),* 2395-2405. https://doi.org/10.1121/1.1784440

Killion, M.C., & Niquette, P.A. (2000). What can the pure tone audiogram tell us about a patients SNR loss. *The Hearing Journal, 53,* 46–53. https://doi.org/10.1097/00025572-200003000-00006

Klein, S. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics, 63(8),* 1421-1455. https://doi.org/10.3758/BF03194552

Leek M. R. (2001)**,** Adaptive procedures in psychophysical research. *Perception and Psychophysics, 63(8),* 1279-1292. https://doi.org/10.3758/BF03194543

Leensen M.C, de Laat J.A , & Dreschler W.A. (2011). Speech-in-noise screening tests by internet, part 1: Test evaluation for noise-induced hearing loss identification. *International Journal of Audiology, 50(11),* 823-834. https://doi.org/10.3109/14992027.2011.595016

Leensen M.C, de Laat J.A , Snik A.F, & Dreschler W.A. (2011). Speech-in-noise screening tests by internet, part 2: improving test sensitivity for noise-induced hearing loss. *International Journal of Audiology, 50(11),* 835-848. https://doi.org/10.3109/14992027.2011.595017

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America, 49(2 Suppl 2),* 467-477. https://doi.org/10.1121/1.1912375

Lyregaard, P. (1997). Towards a theory of speech audiometry tests. In: M. Martin (Ed.), *Speech Audiometry, 2nd Edition*. Chichester, UK: John Wiley & Sons, 34-62.

Mattys, S.L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: dissociating energetic from informational factors. *Cognitive Psychology, 59(3),* 203-243. https://doi.org/10.1016/j.cogpsych.2009.04.001

Mick, P., & Pichora-Fuller, M.K. (2016). Is hearing loss associated with poorer health in older adults who might benefit from hearing screening? *Ear and Hearing,* 37(3), e194-201. https://doi.org/10.1097/AUD.0000000000000267

Nash, S.D., Cruickshanks, K.J., Huang, G.H., Klein, B.E.K., Klein, R., Nieto, F.J., & Tweed, T.S. (2013). Unmet Hearing Health Care Needs: The Beaver Dam Offspring Study. *American Journal of Public Health, 103(6),* 1134-1139. https://doi.org/10.2105/AJPH.2012.301031

Nielsen, J.B., & Dau, T. (2011). The Danish hearing in noise test. International Journal of Audiology, 50(3), 202-208. https://doi.org/10.3109/14992027.2010.524254

Nuesse, T., Steenken, R., Neher, T., & Holube, I. (2018). Exploring the link between cognitive abilities and speech recognition in the elderly under different listening conditions. *Frontiers in psychology, 9,* 678. https://doi.org/10.3389/fpsyg.2018.00678

Olusanya, B., Neumann, K., & Saunders, J. (2014). The global burden of disabling hearing impairment: a call to action. Bulletin Of The World Health Organization, 92(5), 367-373

Paglialonga, A., Fiocchi, S., Parazzini, M., Ravazzani, P., & Tognola, G. (2011). Influence of tinnitus sound therapy signals on the intelligibility of speech. *Journal of Laryngology and Otology, 125*, 795-801. https://doi.org/10.1017/S0022215111000867

Paglialonga, A., Grandori, F., & Tognola, G. (2013). Using the Speech Understanding in Noise (SUN) Test for Adult Hearing Screening. *American Journal of Audiology, 22*, 171-174. https://doi.org/10.1044/1059-0889(2012/12-0055)

Paglialonga, A., Tognola, G., & Grandori, F. (2014). A user-operated test of suprathreshold acuity in noise for adult hearing screening: The SUN (Speech Understanding in Noise) test. *Computers in Biology and Medicine, 52*, 66-72. https://doi.org/10.1016/j.compbiomed.2014.06.012

Paglialonga, A., Tognola, G., & Pinciroli, F. (2015) Apps for Hearing Science and Care. *American Journal of Audiology, 24(3)*, 293-298. https://doi.org/10.1044/2015_AJA-14-0093

Paglialonga, A. (2020) eHealth in Adult Audiologic Rehabilitation. In: J.J. Montano & J.B. Spitzer (Eds.), *Adult Audiologic Rehabilitation 3rd Edition*. San Diego, CA, USA: Plural Publishing.

Potgieter, J.M., Swanepoel, W., Myburgh, H.C., & Smits, C. (2019). The South African English Smartphone Digits-in-Noise Hearing Test: Effect of Age, Hearing Loss, and Speaking Competence. *Ear and Hearing, 39(4)*, 656-663. https://doi.org/10.1097/AUD.0000000000000522

Pronk, M., Kramer, S.E., Davis, A.C., Stephens, D., Smith, P.A., Thodi, C., Anteunis, L.J., Parazzini, M., & Grandori, F. (2011). Interventions following hearing screening in adults: a systematic descriptive review. *International Journal of Audiology, 50(9)*, 594-609. https://doi.org/10.3109/14992027.2011.582165

Schlauch, R.S., & Rose, R.M. (1990). Two-, three-, and four-interval forced-choice staircase procedures: estimator bias and efficiency. *The Journal of the Acoustical Society of America, 88(2),* 732-740. https://doi.org/10.1121/1.399776

Semeraro, H.D., Rowan, D., van Besouw, R.M., & Allsopp, A.A. (2017). Development and evaluation of the British English coordinate response measure speech-in-noise test as an occupational hearing assessment tool. *International Journal of Audiology, 56(10),* 749-758. https://doi.org/10.1080/14992027.2017.131737

Sheikh Rashid, M., & Dreschler, W.A. (2018). Accuracy of an internet-based speech-in-noise hearing screening test for high-frequency hearing loss: incorporating automatic conditional rescreening. International Archives of Occupational and Environmental Health, 91(7), 877–885. https://doi.org/10.1007/s00420-018-1332-5

Shelton, B.R., & Scarrow, I. (1984). Two-alternative versus three alternative procedures for threshold estimation. *Perception and Psychophysics, 35(4),* 385-392. https://doi.org/10.3758/BF03206343

Smith, S.L., & Pichora-Fuller, M.K. (2015). Associations between speech understanding and auditory and visual tests of verbal working memory: effects of linguistic complexity, task, age, and hearing loss. Frontiers in Psychology, 6, 1394. https://doi.org/10.3389/fpsyg.2015.01394

Smits, A., Kapteyn, T., & Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *International Journal of Audiology, 43,* 1-28. https://doi.org/10.1080/14992020400050004

Smits, C., Goverts, T.S., & Festen, J.M. (2013). The digits-in-noise test: assessing auditory speech recognition abilities in noise. *The Journal of the Acoustical Society of America, 133(3),* 1693-1706. https://doi.org/10.1121/1.4789933

Smoorenburg, G.F. (1992). Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. *The Journal of the Acoustical Society of America, 91(1),* 421-437.

Song, J.H., Skoe, E., Banai, K., & Kraus, N. (2012). Training to improve hearing speech in noise: biological mechanisms. *Cerebral cortex, 22(5),* 1180-1190. https://doi.org/10.1093/cercor/bhr196

Strasburger, H. (2001). Converting between measures of slope of the psychometric function. *Perception & Psychophysics, 63,* 1348-1355. https://doi.org/10.3758/BF03194547

Summers, V., Makashay, M.J., Theodoroff, S.M., & Leek, M.R. (2010). Suprathreshold auditory processing and speech perception in noise: hearing-impaired and normal-hearing listeners. *Journal of the American Academy of Audiology, 24*, 274-92. https://doi.org/10.3766/jaaa.24.4.4

Taal, C.H., Hendriks, R.C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing, 19(7),* 2125–2136. https://doi.org/10.1109/TASL.2011.2114881

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research, 35*, 2503-2522. https://doi.org/10.1121/1.4789933

Vaez, N., Desgualdo-Pereira, L., & Paglialonga, A. (2014). Development of a test of suprathreshold acuity in noise in Brazilian Portuguese: a new method for hearing screening and surveillance. *Biomed Research International, 2014,* 652838. https://doi.org/10.1155/2014/652838

Ventry, I.M., & Weinstein, B.E. (1983). Identification of elderly people with hearing problems. *ASHA, 25,* 37-42.

Watson, C., Kidd, G., Miller, J., Smits, C., & Humes, L. (2012). Telephone screening tests for functionally impaired hearing: Current use in seven countries and development of a US version. *Journal of the American Academy of Audiology, 23(10),* 757-767. https://doi.org/10.3766/jaaa.23.10.2

Wilson, B., Tucci, D., Merson, M., & O'Donoghue, G. (2017). Global hearing health care: new findings and perspectives. *The Lancet, 390(10111),* 2503-2515. https://doi.org/10.1016/S0140-6736(17)31073-5

World Health Organization (WHO). (2011). World report on disability. WHO press, Geneve, Switzerland. Retrieved from: http://whqlibdoc.who.int/publications/2011/9789240685215_eng.pdf

World Health Organization (WHO) (2018). Deafness and hearing loss. Fact Sheet n. 300. 15 March 2018. Retrieved from http://www.who.int/mediacentre/factsheets/fs300/en/.

Yousuf Hussein, S., Swanepoel, D.W., de Jager, B.L., Myburgh, H.C., Eikelboom, R.H., & Hugo, J. (2016). Smartphone hearing screening in mHealth assisted community-based primary care. *Journal of Telemedicine and Telecare, 22(7),* 405-412. https://doi.org/10.1177/1357633X15610721

Zanet, M., Polo, E.M., Rocco, G., Paglialonga, A., & Barbieri, R. (2019). Development and preliminary evaluation of a novel adaptive staircase procedure for automated speech-in-noise testing. *Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* July 23-27 2019, Berlin, Germany, pp. 6991-6994. https://doi.org/10.1109/EMBC.2019.8857492.

Zurek, P.M., & Delhorne, L.A. (1987). Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment. *The Journal of the Acoustical Society of America, 82(5),* 1548-1559.

Table 1. SRTs (mean, standard deviation, and range) measured with the conventional staircase and with the proposed test in test and retest trials in audiometer-controlled and self-adjusted output levels conditions in the YA group (N = 26 ears from 26 subjects).

		audiometer-controlled		self-adjusted	
		test	retest	test	retest
Conventional	mean (s.d.)	-15.45 (1.38)	-17.03 (2.14)	-17.8 (1.87)	-19.11 (2.37)
	range	-18.43 ÷ -13.14	-21.14 ÷ -11.71	-21.29 ÷ -14.31	-24.20 ÷ -10.95
Proposed	mean (s.d.)	-15.39 (1.84)	-15.20 (1.71)	-15.28 (1.87)	-15.71 (1.61)
	range	-18.25 ÷ -10.25	-18.50 ÷ -10.75	-18.75 ÷ -11.75	-18.75 ÷ -12.00

Table 2. Participants characteristics (number and gender, age, HHIE-S, and PTA) and SIN test outcomes (SRT, number of stimuli, test duration, and percentage of correct responses) in normal hearing young adults (YA, N = 26 ears from 26 subjects), in $UA_{\leq 25}$ ears (N = 37 ears from 34 subjects) and in $UA_{>25}$ ears (N = 43 ears from 38 subjects).

		YA	**$UA_{\leq 25}$**	$UA_{>25}$
Subjects	number (m/f)	26 (11/15)	34 (14/20)	38 (11/27)
Age (years)	mean (s.d.) range	24.2 (0.59) 23÷26	55.2 (14.46) 24÷79	70.4 (9.60) 48÷89
HHIE-S (score)	mean (s.d.) range	0.0 (0.00) 0÷0	6.5 (6.95) 0÷30	11.1 (9.63) 0÷36
PTA (dB HL)	mean (s.d.) range	-0.5 (6.46) -10.0÷20.0	21.6 (2.91) 16.25÷26.25	39.4 (8.94) 30.0÷73.75
SRT (dB SNR)	mean (s.d.) range	-15.3 (1.87) -18.7÷-11.8	-11.6 (4.02) -17.8÷-1.3	-5.2 (6.82) -18.2÷9.7
#Stimuli (number)	mean (s.d.) range	91.9 (12.01) 68÷114	80.1 (9.55) 56÷101	67.9 (15.80) 39÷110
Test duration (s)	mean (s.d.) range	233 (39.07) 165÷309	233 (48.90) 153÷358	251 (70.78) 147÷497
Correct responses (%)	mean (s.d.) range	91.1 (1.00) 88.9÷92.9	89.8 (2.58) 76.8÷93.1	87.1 (5.02) 64.4÷93.0

Table 3. Group characteristics (subject numbers and gender, age, HHIE-S, and PTA) and test-retest absolute variations in SRT, number of stimuli, test duration, and percentage of correct responses in the subgroup of unscreened adults who underwent the SIN test twice in the same ear ($UA_{t/r}$, N = 22 ears from 21 subjects).

		$UA_{t/r}$
Subjects	number (m/f)	21 (5/16)
Age (years)	mean (s.d.) range	69.1 (12.46) 42÷89
HHIE-S (score)	mean (s.d.) range	8.9 (8.06) 0÷28
PTA (dB HL)	mean (s.d.) range	31.8 (7.92) 20.0÷73.75
SRT (dB SNR) absolute variation	mean (s.d.) range	3.5 (2.86) 0.2÷9.7
#Stimuli (number) absolute variation	mean (s.d.) range	11.5 (9.58) 0÷40
Test duration (s) absolute variation	mean (s.d.) range	44.8 (38.53) 4÷153
Correct responses (%) absolute variation	mean (s.d.) range	2.0 (2.28) 0÷9.5

Figure legends

Figure 1. Scatterplot of SRT and PTA measured in the study sample (N = 106 ears from 98 subjects) and linear regression analysis. Cross markers: YA group (N=26 ears from 26 subjects); triangle markers: $UA_{\leq 25}$ (N=37 ears from 34 subjects); circle markers: $UA_{>25}$ (N=43 ears from 37 subjects). Filled markers indicate ears from subjects who were native speakers of English in the UA groups.

Figure 2. Scatterplot of SRT and age measured in the study sample (N = 106 ears from 98 subjects) and linear regression analysis. Cross markers: YA group (N=26 ears from 26 subjects); triangle markers: $UA_{\leq 25}$ (N=37 ears from 34 subjects); circle markers: $UA_{>25}$ (N=43 ears from 37 subjects).

Figure 3. Receiver operating characteristic (ROC) obtained with the proposed SIN test in the UA population (N = 106 ears from 98 subjects). Dot marks indicate the measured points obtained by varying the cut-off SRT from 9.75 to -18.75 dB SNR in 0.25 dB steps. Cross markers indicate the three points closer to the point at (0,1), i.e. the better trade-offs between sensitivity and specificity.

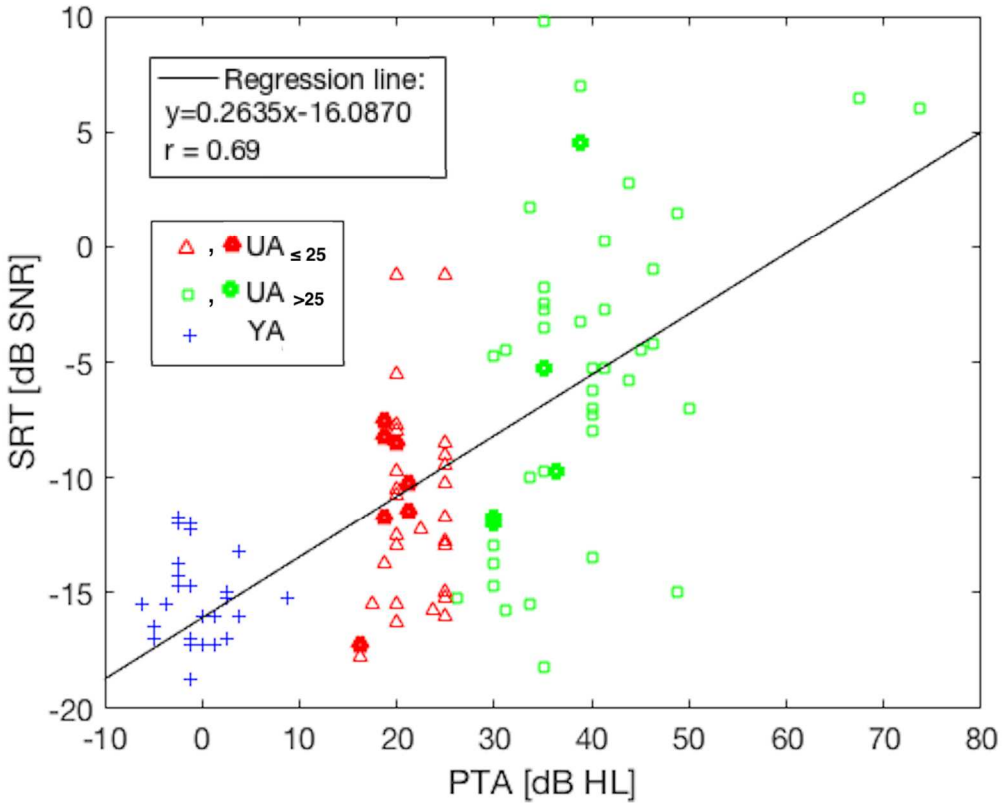
Regression line:
y=0.2635x-16.0870
r = 0.69
UA ≤ 25
UA >25
YA
SRT [dB SNR]
PTA [dB HL]
10
5
0
-5
-10
-15
-20
-10
0
10
20
30
40
50
60
70
80

Figure 1.

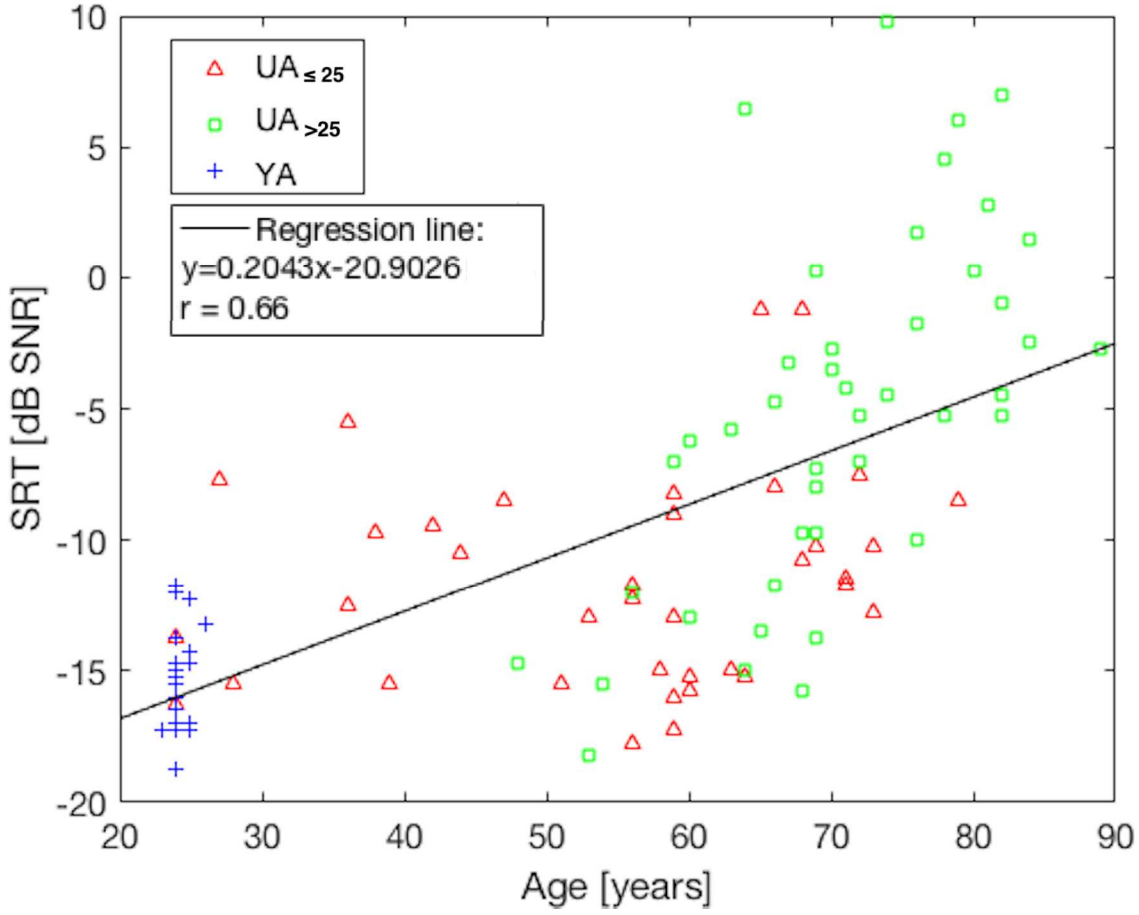
UA ≤ 25
UA >25
YA
Regression line:
y=0.2043x-20.9026
r = 0.66
SRT [dB SNR]
Age [years]
10
5
0
-5
-10
-15
-20
20
30
40
50
60
70
80
90

Figure 2.

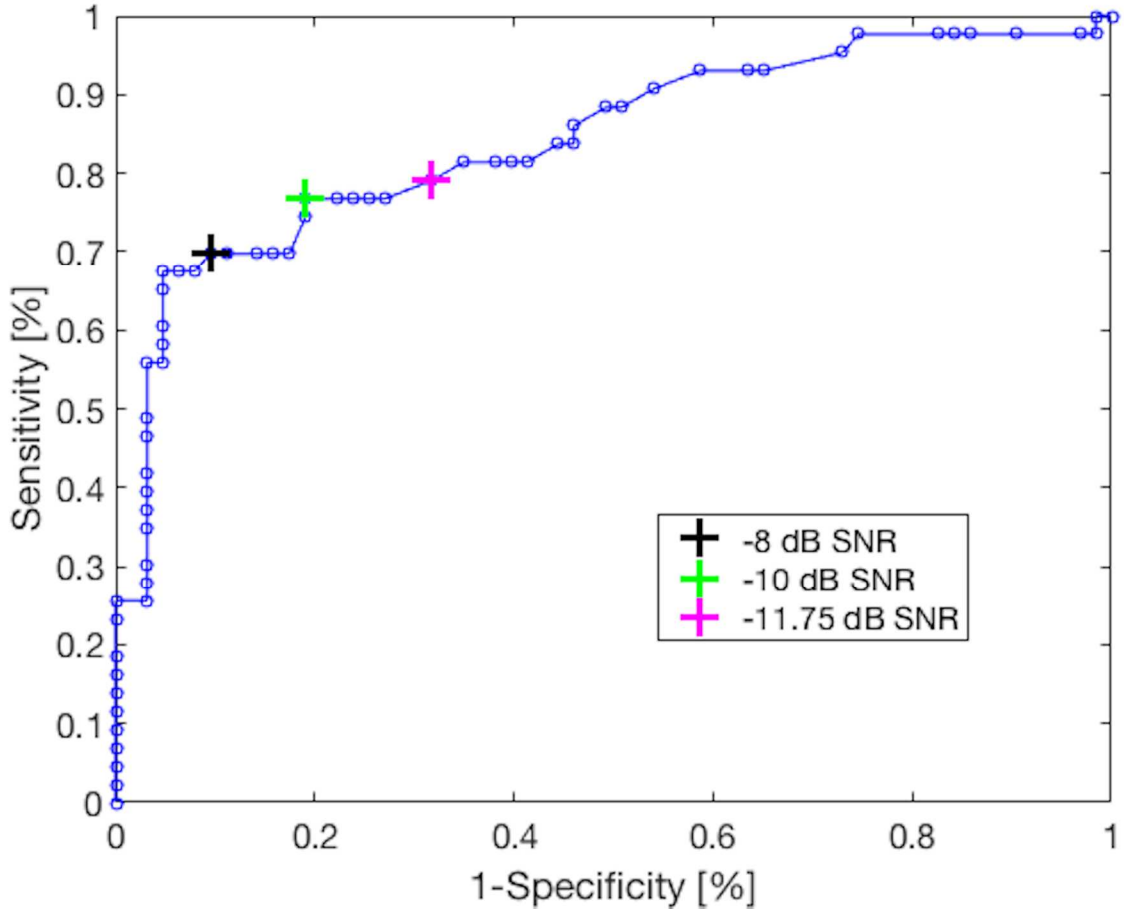
Sensitivity [%]
1-Specificity [%]
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
-8 dB SNR
-10 dB SNR
-11.75 dB SNR

Figure 3.