Development and Evaluation of a Novel Method for Adult Hearing Screening: Towards a Dedicated Smartphone App

Edoardo Maria Polo[1,2], Marco Zanet[3] Marta Lenatti[2], Toon van Waterschoot[4][0000-0002-6323-7350], Riccardo Barbieri[2][0000-0001-9381-3833], and Alessia Paglialonga[3*][0000-0002-1341-2560]

[1] DIAG, Sapienza University of Rome, 00185 Rome, Italy
[2] Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), 20133 Milan, Italy
[3] National Research Council of Italy (CNR), Institute of Electronics, Information Engineering and Telecommunications (IEIIT), 20133 Milan, Italy
[4] KU Leuven, Department of Electrical Engineering (ESAT-STADIUS), 3001 Leuven, Belgium
* alessia.paglialonga@ieiit.cnr.it

Abstract. Towards implementation of adult hearing screening tests that can be delivered via a mobile app, we have recently designed a novel speech-in-noise test based on the following requirements: user-operated, fast, reliable, accurate, viable for use by listeners of unknown native language and viable for testing at a distance. This study addresses specific models to (i) investigate the ability of the test to identify ears with mild hearing loss using machine learning; and (ii) address the range of the output levels generated using different transducers. Our results demonstrate that the test classification performance using decision tree models is in line with the performance of validated, language-dependent speech-in-noise tests. We observed, on average, 0.75 accuracy, 0.64 sensitivity and 0.81 specificity. Regarding the analysis of output levels, we demonstrated substantial variability of transducers' characteristics and dynamic range, with headphones yielding higher output levels compared to earphones. These findings confirm the importance of a self-adjusted volume option. These results also suggest that earphones may not be suitable for test execution as the output levels may be relatively low, particularly for subjects with hearing loss or for those who skip the volume adjustment step. Further research is needed to fully address test performance, e.g. testing a larger sample of subjects, addressing different classification approaches, and characterizing test reliability in varying conditions using different devices and transducers.

Keywords: Classification, Decision Trees, Hearing Loss, Hearing Screening, Smartphone App, Speech-in-Noise testing.

1 Background

The digital health revolution, supported by ubiquitous connectivity, enables new ways of delivering decentralized healthcare services using eHealth and mHealth solutions, and hearing healthcare makes no exception [1]-[4]. In the 'new normal' brought by the COVID-19 pandemic, no-touch services are now critical for individuals with age-related hearing loss, who are typically at the highest risk for morbidity and mortality due to their age [5]. In this context, smartphone hearing health apps have grown popular but the availability of validated apps for hearing screening and assessment is still limited [4], [6]-[8]. Some validated hearing testing apps are currently available on the market, for example *SHOEBOX Audiometry* (an FDA Class II medical device for pure-tone audiometry requiring calibrated transducers) developed by SHOEBOX Ltd, *hearScreen* (a pure-tone audiometry screening app coupled with calibrated headphones), and *hearWHO* (a speech-in-noise testing app, based on the digits-in-noise test [9] in English, endorsed by the World Health Organization), both developed by hearX Group.

Hearing screening in adults is particularly important to identify early signs of hearing loss, and therefore trigger timely intervention, thus preventing or delaying the progression of hearing loss and its impact on communication and psychosocial functioning. In fact, hearing loss is typically neglected in adults and access to care is frequently delayed until major effects in health-related quality of life occur, leading to increased health care costs and utilization patterns [10]-[11].

Speech-in-noise tests can be helpful in adult hearing screening to identify the real-life communication problems and to promote awareness in individuals who would otherwise not seek help, or who would seek help very late. Speech-in-noise tests can overcome some of the limitations that make pure-tone audiometry unfeasible for widespread automated self-testing on remote (e.g., need for calibrated transducers, need for low-noise environment) [12]-[13]. Moreover, speech-in-noise tests can be easily implemented in an automatic way, for example using multiple-choice tasks on a user-operated interface (e.g., [14]-[16]). However, a potential limitation of speech-in-noise tests in the context of widespread screening (for example via smartphone apps) is related to the fact that they are typically language-dependent. In fact, these tests typically use sentences, words, or digits and therefore they need to undergo translation, adaptation for psychometric performance, and validation when a new language version has to be developed (e.g., [15]-[16]). The use of language-dependent tests may potentially lead to decreased access to screening, disparities, or inaccurate results for non-native listeners and minorities. This is particularly relevant for tests delivered at a distance via smartphone apps as the target population is scattered across native languages.

Recently, we have developed a new and automated speech-in-noise test that reduces possible issues related to language dependence. The test is aimed at adult hearing screening for future implementation in a smartphone app. The main requirements followed for test design are discussed below, along with an outline of the current stage of development:

- *Automated, user-operated execution.* The test is based on a multiple-choice recognition task via an easy-to-use graphical user interface that is optimized for delivery via

a touch-sensitive screen. A three-alternative forced-choice task is used with alternatives determined on a maximal opposition criterion (different in place, manner, and voicing) as a proven trade-off between test complexity and psychometric performance [14], [16].

- *Speech stimuli viable for use in individuals of unknown language*. The test is based on meaningless Vowel-Consonant-Vowel (VCV) stimuli spoken by a professional male native English speaker and a set of consonants common across some of the top spoken languages worldwide (e.g., English, Spanish, French, Portuguese, German, Italian), taking into consideration their viability for application in non-native listeners who are familiar with the Latin alphabet [18]-[19]. Preliminary results in native and non-native listeners indicated that the test performance was stable in listeners of varying language [20].
- *Short test duration*. To enable faster convergence of the adaptive algorithm, the test uses a newly developed staircase procedure that, based on the estimated psychometric curves of stimuli, determines optimized upward and downward steps as opposed to conventional staircase procedures that use pre-determined, equal upward and downward steps. Preliminary results in normal hearing adults showed that the test duration of the new staircase was, on average, two minutes shorter than that of a conventional staircase (i.e., about 3 min 30 s vs. about 5 min 30 s). Moreover, similar values of test duration were observed in subjects with normal hearing and in subjects with hearing loss [20]-[21].
- *Reliable in repeated measures*, to ensure intra-individual repeatability of test results. Preliminary results showed that the proposed test provides repeatable estimates of the speech reception threshold (SRT) and repeatable performance (number of stimuli presented, test duration, and percentage of correct responses) in individuals with normal hearing and with hearing loss. In addition, results showed that, thanks to the short test duration and the multiple-choice design, no perceptual learning effect was observed in the second execution of the test compared to the first one [20]-[21].
- *Accurate in identifying hearing loss*, to ensure accurate screening outcomes. Preliminary results obtained in a population of 98 adults (including normal hearing and unscreened adults) have shown that, in terms of SRT estimation, the test was as accurate as a conventional adaptive staircase [20]-[21]. Preliminary analysis of the test classification performance, based on the SRT only, showed that the accuracy (ACC) of the test for the identification of ears with pure-tone thresholds higher than 25 dB HL at 1, 2, or 4 kHz was equal to 0.82 and the area under the receiver operating characteristic (AUC) was equal to 0.84 [20]. For a full characterization of classification performance for the purpose of test validation, comprehensive analysis of classification performance based on the full set of test features (e.g., SRT, number of trials, test duration, percentage of correct responses, average reaction time, and so on) and on a larger sample of individuals is needed. This contribution presents the first results obtained from a multivariate classification approach on a population of 148 adults including subjects with normal hearing and subjects with varying degrees of hearing loss.

– *Viable for testing at a distance*, to ensure reliability of results in varying settings (i.e., in different environments and with different instrumentation). Preliminary results from a group of 26 normal hearing adults showed that the test provided consistent outcomes in terms of SRT estimation and test-retest repeatability in controlled environmental noise settings (audiometer-controlled output levels) and in uncontrolled environmental noise settings (self-adjusted test volume) [21]. However, to fully demonstrate the reliability of test results in varying settings, a deeper analysis of the possible influence of the environmental noise, instrumentation, and test settings is needed. This contribution presents the first results obtained in this direction, specifically a quantitative analysis of the influence of different transducers on the output levels of the test.

Within this context, the aim of this study was twofold. First, regarding the requirement of test accuracy in identifying hearing loss, we performed a preliminary analysis, using a machine learning approach, of the classification performance considering the full set of available features (Sections 2.2 and 3.1). Second, regarding the requirement of viability for testing at a distance in uncontrolled settings, we conducted an experiment to quantitatively address the influence of self-adjusted test volume settings using different consumer transducers on the output levels of the test (Sections 2.3 and 3.2). An investigation of these aspects is crucial to fully address the viability of the proposed test for adult hearing screening via a smartphone app.

2 Materials and Methods

2.1 Speech-in-Noise Test

The proposed speech-in-noise test is based on three-alternative forced-choice recognition of VCV stimuli via a graphical user interface. The set of stimuli includes 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 [14]-[20-21]. VCVs were combined with speech-shaped noise at varying signal-to-noise ratio (SNR). The noise is 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 [22] and a low-pass filter (cutoff =1.4 kHz, roll-off slope = 100 dB/octave) and then by adding a noise floor determined by the same filtered noise attenuated by 15 dB [23].

After initial collection of information about the subjects' age and gender through the graphical user interface, the test starts at an initial comfortable level of +8 dB SNR from a stimulus randomly selected from the set of VCVs. Then, it adapts the intelligibility based on a one-up/three-down (1U3D) rule, i.e. the intelligibility is decreased after three correct responses and increased after one incorrect response. The intelligibility is adjusted by changing, concurrently, the VCV and the SNR based on a newly developed staircase procedure [20-21]. Specifically, the upward and downward steps in SNR are determined adaptively at each trial, based on the psychometric curves of VCV stimuli and therefore the steps depend on the specific stimulus and SNR at each trial. The steps are set using the optimal recommended ratio between downward and upward step size

for 1U3D staircases (i.e., 0.74), therefore enabling rapid convergence of the tracking procedure as suggested by [24]. At each step, the VCV presented and the order of the alternatives displayed on the screen are randomized. The procedure is terminated after 12 reversals in SNR and the SRT is estimated as the average of the SNRs at the mid-points of the last four ascending runs [20]-[24].

2.2 Classification Performance

Participants. Participants were 148 adults (age = 52.1 ± 20.4 years; age range: 20-89 years; 46 male, 102 female) tested in uncontrolled environmental noise settings in the lab and at local health screening initiatives (i.e. at universities of senior citizens, health prevention and awareness events for the general public). The group of participants included individuals with normal hearing and individuals with varying degrees of hearing impairment. The experimental protocol was approved by the Politecnico di Milano Research Ethical Committee (Opinion n. 2/2019). All subjects were informed about the protocol and the study as a whole and they took part in the experiment on a voluntary basis. Due to the opportunistic nature of the local screening initiatives, participants were given the option to choose in which ear(s) to perform the test. As a result, 8 participants performed the test sequentially in both ears whereas 140 performed the test only in one ear, for a total of 156 ears tested.

Procedures. An outline of the experiment is shown in Fig. 1. Participants were tested with: (i) pure-tone audiometry at 0.5, 1, 2, and 4 kHz (Amplaid 177+, Amplifon with TDH49 headphones), and (ii) the proposed speech-in-noise test in uncontrolled environmental noise settings. The test was run on an Apple® Macbook Air® 13'' (OS X Yosemite version 10.10.5 and macOS High Sierra version 10.13.6) connected to Sony MDRZX110APW headphones. The speech-in-noise test was executed in self-adjusted volume settings therefore participants were given the option to adjust the volume at a comfortable level before the test via the graphical user interface.

Data Analysis. The pure-tone threshold average (PTA) was computed as the average of hearing thresholds at the four frequencies tested. Then, the tested ears were classified based on their PTA in two classes using the World Health Organization (WHO) criterion for slight/mild hearing impairment: $PTA_{>25\ dB\ HL}$ (slight/mild hearing loss) and $PTA_{\leq 25\ dB\ HL}$ (no hearing loss) [25]. For a multivariate analysis of the test classification performance, the following features were extracted from the speech-in-noise test software: SRT, total number of trials (#trials), number of correct responses (#correct), percentage of correct responses (%correct), average reaction time (i.e., the average of individual response time throughout the test), test duration, output volume, and age. The eight features extracted from the test were used as input variables of a decision tree (DT) model and the PTA class was used as the output variable (Fig. 1).

A DT approach was used in this first study as it is one of the most broadly used classification methods due to its ability to convert paths in the tree into intelligible decision rules. The Gini index was chosen as splitting rule in place of entropy as they lead to similar results but the Gini index has lower computational weight [26]. To limit the possible effects of overfitting and therefore model bias, the dataset was first split randomly into training (80% of the sample, 124 ears) and test (20% of the sample, 32 ears) datasets. Then, the DT model was optimized using 5-fold cross-validation on the training dataset and finally its predictions were tested on the test dataset. Classification performance was assessed by measuring: accuracy (ACC) on the training and test datasets, area under the curve (AUC), sensitivity (SEN), and specificity (SPE). Due to the relatively small size of the dataset, we also addressed the variability of the model performance. Specifically, we ran 1000 iterations of the model optimization process by randomly changing the initial splitting into training and testing datasets (keeping an 80%/20% splitting ratio) and the inner cross-validation subsets (keeping a 5-fold inner splitting) and then we measured the average and standard deviation of the performance parameters (ACC on the training and test datasets, AUC, SEN, and SPE).

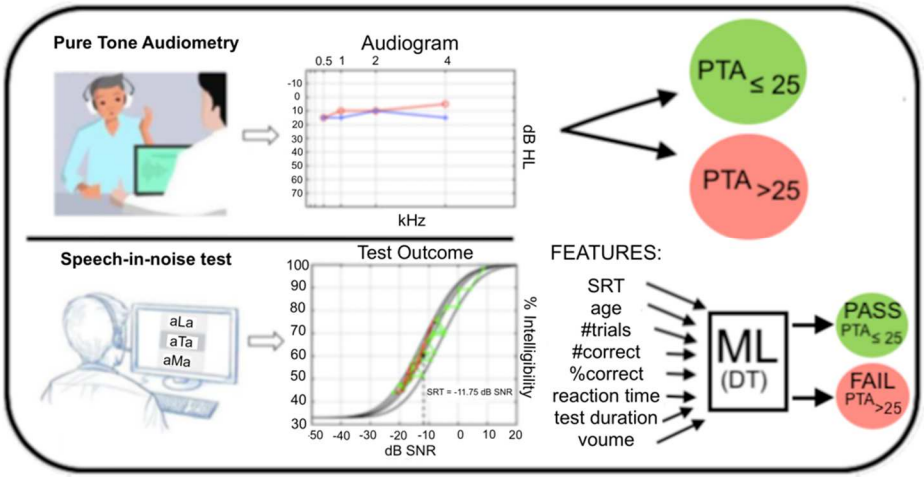

Fig. 1. Classification performance analysis methodology. Top panel: Based on the PTA measured with pure-tone audiometry, ears were classified using the WHO criterion for slight/mild hearing impairment as $PTA_{>25\ dB\ HL}$ and $PTA_{\leq 25\ dB\ HL}$. Bottom panel: Based on eight features extracted from the speech-in-noise test sotware, ears were classified into *pass* and *fail* using a DT approach based on the PTA class as the output variable. DT = decision tree; PTA = pure-tone threshold average; WHO = World Health Organization.

2.3 Characterization of Transducers

In order to assess the influence of self-adjusted volume settings on the actual output levels of the test, we characterized the performance of different consumer transducers, spefically we measured the actual output levels of the test obtained using a variety of transducers as a function of the self-adjusted test volume settings. In fact, when a

screening test is delivered via a smartphone app in uncontrolled environmental noise settings, a variety of transducers can be used therefore it is important to understand the actual output level for a given test volume selected by the user. In this study, we have characterized five different headphones models and two different earphones models widely available on the market, i.e.: Bose Quitecomfort II (in two versions: noise canceling mode ON and noise canceling mode OFF), Sony MDRZX110APW, Sony MDR-7506, Sennheiser PC 310, Akg Y45, Apple EarPods, and Mpow In-ear (price range: 9.99 to 299 €).

The experimental setup for the characterization of transducers is shown in Fig. 2. First, to take into account all the stimuli in the set, an audio file including the sequence of all VCVs, with no pauses, was created. The volume range of the same laptop computer used in the listening tests (described in 2.2) was discretized in 17 levels using a step equal to the size of the volume bars of the laptop. The VCV sequence reproduced via the different transducers, at each of the 17 volume levels, was recorded using a Neumann KU 100 dummy head powered from an external P48 phantom power supply. Each recording was then routed back to the laptop computer via a RME Babyface Pro sound card and the corresponding digital audio files were saved using GarageBand software (version 10.1.3). In order to record the VCV sequence under the same conditions for all the transducers, the dummy head was positioned and remained in the same location in a quiet room and the RME Babyface Pro input gain (dummy head/laptop) was maintained fixed and low throughout the experiment to avoid possible sound saturation.

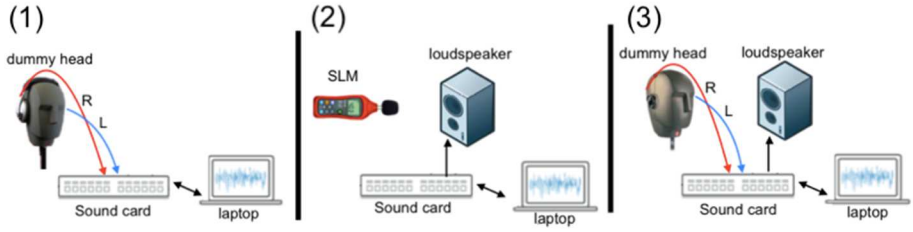

Fig. 2. Characterization of transducers. Panel (1): the VCV sequence for each transducer is recorded via a dummy head and a sound card. Panel (2) and (3) show the two steps of the calibration process. Panel (2): the sound card output gain was set to let the white noise from the loudspeaker reach 90 dB SPL on the SLM. Panel (3): the SLM was replaced by the dummy head which recorded the 90 dB SPL white noise and saved it into the laptop maintaining the same input gain of the sound card as that of the VCV sequence recordings. R = right ear; L = left ear; SLM = sound level meter.

A calibration step was needed to convert wave units into dB sound pressure level (SPL). In fact, the absolute amplitude of the recording sequences was in wave units, which are given by the combination of the actual SPL and any other digital gains of the transfer function of the laptop-transducer chain. The calibration process was performed with the dummy head and a loudspeaker and therefore a white noise was used as the recorded loudness of white noise is less influenced by the acoustical attenuation due to the wide frequency range. The white noise was sent via the laptop computer and sound card to

the loudspeaker and the loudness was measured using a Sound Level Meter (SLM; Brüel & Kjær Type 2250 Hand Held Analyzer with BZ-7222 Sound Level Meter Software) placed on a tripod in front of the loudspeaker at a distance of 1 meter. First, the output gain of the sound card was changed to let the noise loudness reach 90 dB SPL at the SLM. Then, the tripod with the SLM was removed and replaced by the dummy head. Finally, to obtain a calibration factor the white noise was played with the output gain corresponding to a known SLM loudness of 90 dB SPL maintaining the input gain used for the VCV sequence and recorded by the dummy head. The VCV sequences recorded with the different transducers at each of the 17 volume levels were therefore converted in dB SPL and then filtered using an A-weighted filter to approximate the SPL perceived by the average human ear. The experiment was conducted at the Department of Electrical Engineering (ESAT) at KU Leuven.

The characteristics of the transducers, in terms of dB SPL as a function of the volume level were obtained by fitting the 17 sampled measures using polynomial models of order 1-9 using the least-squares method and the optimal model order for each curve was determined using the Akaike information criterion.

3 Results

3.1 Classification Performance

Fig. 3 shows the optimal DT model obtained by using the full set of eight features. Here, it is important to note that, in classifying ears into *pass* or *fail* (as defined by the WHO criterion for slight/mild hearing loss) the top decision node takes into account the value of SRT estimated by the test and uses a cut-off equal to about -7.5 dB SNR to split the tree into branches. On the right-hand side branch, the SRT is also used again down the tree (fourth level split) to classify a subset of 27 ears from subjects older than 58 years (second level split) in which the number of correct responses was lower than or equal to 89 (third level split).

In general, in addition to SRT, the most relevant features for classification are: the subjects' age (with fail outcomes associated with older age), the test duration (totsec in the figure, with fail outcomes associated with longer test duration), and the average reaction time (avg_ans_time in the figure, with fail outcomes associated with longer duration). Splitting rules involving other features such as the total number of correct responses (#correct in the figure), the percentage of correct responses (perc_correct in the figure), or the total number of trials (#trials in the figure) seem to be less informative as rules based on these features are associated with leaf nodes with a very small number of ears (less than 3).

To address the possible relationships between the model features and their distribution in the two PTA classes, we analyzed the scatter plots for each combination of paired features as well as the distribution of each feature in the two classes. Results are shown in Fig. 4.

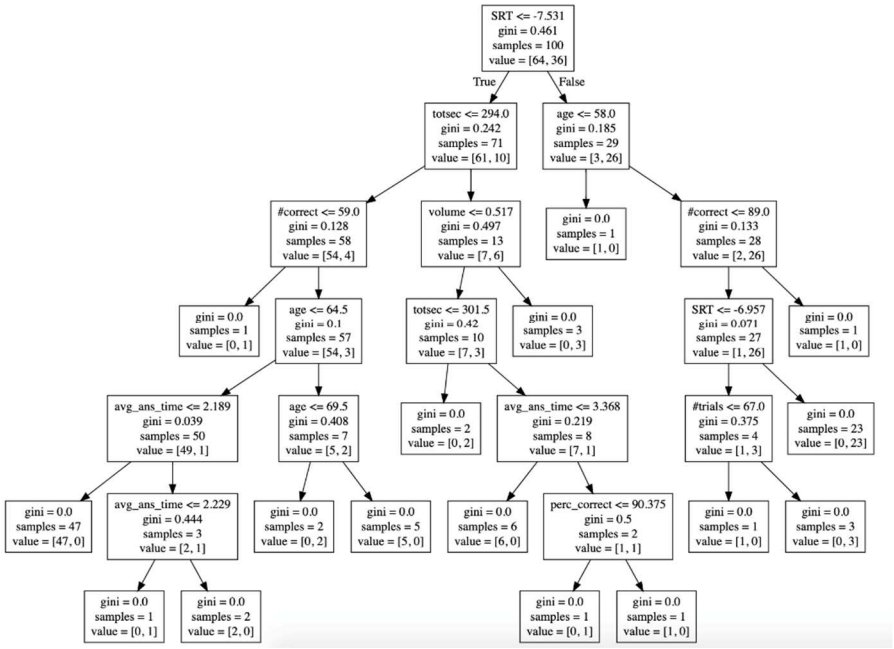

Fig. 3. Optimal model for classification of ears into *pass* and *fail* using the full set of features as input variables and the WHO definition of normalhearing /mild hearing loss as output variable.

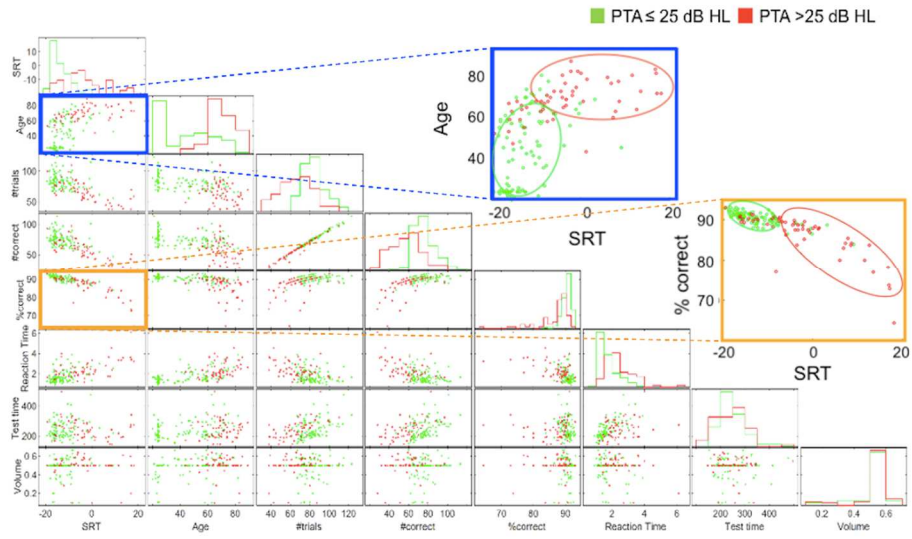

Fig. 4. Scatter plots of paired features and, on the diagonal, distributions of single features in the two classes. Magnified versions of two exemplary scatter plots (SRT vs. Age and SRT vs. %correct) are shown on the right-hand side. Green dots: ears with PTA ≤ 25 dB HL. Red dots: ears with PTA > 25 dB HL.

The distributions of each feature, reported on the diagonal of the matrix in Fig. 4, show that some features may be likely candidates for classifying ears in the two PTA classes. Specifically, the better candidates are SRT, age, number of trials and number of correct responses, as well as percentage of correct responses derived from the previous two, and average reaction time. Features such as test duration and volume level show comparable values between the two PTA classes. The scatter plots between pairs of features show that, for example, the SRT combined with features such as age, percentage of correct responses, and average reaction time (as shown in the leftmost column in Fig. 4) generate clusters of data points that are relatively grouped into the two PTA classes, suggesting that a selection of these features may be possibly used to generate simpler DT classification models that are based on a smaller set of rules and may be possibly more intelligible.

To address the performance of models based on a smaller set of rules, we considered DTs with four input features (SRT, age, percentage of correct responses, and average reaction time), as determined by the analysis of the results in Fig. 4. The maximum depth of the DT with four features was set to 4 to counterbalance possible overfitting effects caused by a reduced number of features. The optimal DT model obtained using this selection of features is shown in Fig. 5.

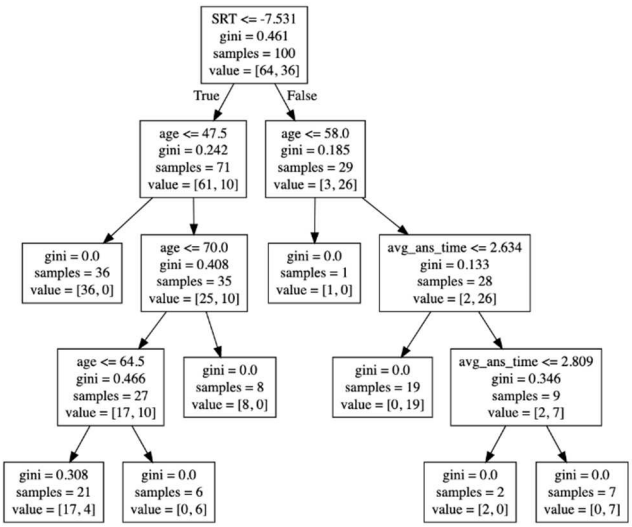

Fig. 5. Optimal model for classification of ears into pass and fail using four features as input variables and the WHO definition of normalhearing /mild hearing loss as output variable.

The root node of the DT with four selected features showed in Fig. 5 uses the same rule as the root node of the full model with eight features, i.e. a split on the SRT value using a cut-off equal to -7.5 dB SNR. Noticeably, this cut-off value is close to the cut-off determined in our earlier study using only the SRT on a smaller set of cases, i.e., -8 dB SNR [20]. Importantly, using a generalized linear model with age and SRT as input

variables on a dataset of 58 subjects we also showed that, in addition to SRT, the interaction between age and SRT was a significant predictor of PTA class (slight/mild hearing impairment vs no hearing impairment) [27]. This aspect is reflected in the observed DT model as the second most important feature is the subject's age which is used at the second, third, and fourth levels of the tree (Fig. 5). The average reaction time (avg_ans_time in the figure) is also relevant as shown in the right-hand branch of the model at the third and fourth levels of the tree, with fail outcomes associated with lower or alternatively with higher reaction times. In fact, it may be that individuals who have difficulty in speech recognition may just tend to select an alternative on the screen quickly and let the test proceed when they haven't heard the stimulus or, alternatively, it may be that they spend a relatively long time to choose among the alternatives because they are not sure about what they have heard.

Table 1 shows the classification performance and the variability of performance on 1000 iterations (mean ± standard deviation) obtained with the optimal DT models, the one with eight features and the one with four features. The first row shows the accuracy observed in the training dataset (ACC_{train}) whereas the remaining rows show the performance (ACC, AUC, SEN, and SPE) observed in the test dataset.

Table 1. Classification performance and variability of performance of the optimal DT models with eight input features and with four input features.

Parameter	Optimal model (8 features)	1000 iterations (8 features)	Optimal model (4 features)	1000 iterations (4 features)
ACC_{train}	0.73	0.77±0.0342	0.77	0.76±0.0339
ACC	0.78	0.76±0.0721	0.75	0.77±0.0707
AUC	0.77	0.74±0.0810	0.72	0.75±0.0803
SEN	0.73	0.67±0.1418	0.64	0.67±0.1482
SPE	0.81	0.81±0.0895	0.81	0.82±0.0868

It is important to notice that the two models have slightly different values of ACC on the training and test datasets. For the model with eight features, ACC on the test set is slightly higher than ACC_{train} whereas for the model with four features the opposite trend is observed. However, for each model the mean values of ACC observed on 1000 iterations are similar in the training and testing datasets, and the ACC values are similar between the two models indicating that, overall, the DT classifiers here shown have an average accuracy in the range 0.76 to 0.77, with standard deviation of about 0.07. In general, the observed performance is similar for the DT models with eight and four features in terms of ACC, AUC, SEN, and SPE values. The SEN observed with the optimal model with eight features is seemingly higher than that observed with the optimal model with four features (i.e., 0.73 vs 0.64). However, the analysis of the mean values obtained on 1000 iterations shows that the average performance of the two models is strikingly similar. Moreover, the observed standard deviations are relatively low for each of the observed performance parameters. The standard deviation is smaller than 0.1 for all the parameters except for sensitivity for which it is about 0.14 and 0.15

for the models with eight and four features, respectively. Therefore, the observed difference in SEN between the two models (0.73 vs 0.64) may be the result of the inherent variability of the models generated using different distributions of data into the traning and test datasets.

3.2 Characterization of Transducers

The characteristics of the tested transducers, in terms of dB SPL as a function of volume level (percent value), are shown in Fig. 6. Overall, two different sets of curves were observed for earphones (Apple EarPods and Mpow in-ear) and for headphones. Among the headphones here used, the highest output levels were observed with the Bose QuiteComfort II headphones with noise canceling mode OFF whereas the characteristics of the remaining headphones were, overall, similar, with differences lower than 6.1 dB SPL across the volume range.

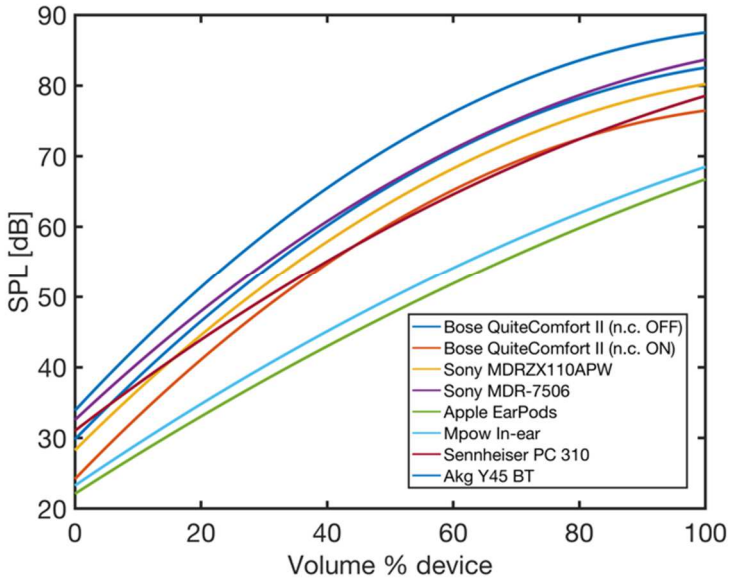

Fig. 6. Transducers characteristics: measured output levels (in dB SPL) as a function of laptop volume level (percent value). n.c. = noise canceling.

The two tested earphones had an overall linear dynamics and produced lower SPL for a given volume level across the whole volume range. For example, when the volume level is 50% (i.e., the default level on the test software) the output level with the Bose QuiteComfort II headphones with noise canceling mode OFF is about 71.5 dB SPL whereas with Apple ear-pods and Mpow in-ear the output level is about 47.5 and 50 dB SPL, respectively. Similarly, when the volume level is set at the top limit of the range, i.e. at 100% the output level with the Bose QuiteComfort II headphones with noise

canceling mode OFF is about 87.5 dB SPL whereas with Apple ear-pods and Mpow in-ear the output level is about 66.5 and 68.5 dB SPL, respectively. The largest differences between transducers were observed at a volume of about 70%, with an observed difference of about 25 dB SPL between Bose QuiteComfort II headphones with noise canceling mode OFF and Apple ear-pods.

4 Discussion

In this study, we begun our analysis by addressing classification performance of the proposed speech-in-noise test against the WHO criterion for mild hearing loss using a DT approach in two different configurations: (i) using the full set of eight features (SRT, number of trials, number of correct responses, percentage of correct responses, average reaction time, test duration, output volume, and age) and (ii) using a subset of four features (i.e., SRT, age, percentage of correct responses, and average reaction time) that were selected based on their distributions in the two classes of the output variable.

Overall, the performance of the DT models using eight and four features was similar as shown in Table 1. In addition, the average model performance, determined by running 1000 iterations of model optimization on different realizations of the training and test datasets, was strikingly similar for the DT models with eight and four features in terms of ACC, AUC, SEN, and SPE. The observed variability of performance was also similar between the two models (Table 1).

Compared to the DT model with the full set of eight features, the DT model with four features may have the advantage of being less demanding in terms of computational cost and simpler in terms of interpretability of rules. The DT model with SRT, age, percentage of correct responses, and average reaction time as input features had on the test dataset an ACC equal to 0.75, an AUC equal to 0.72, SEN equal to 0.64 and SPE equal to 0.81. In our earlier analyisis of test performance [20], using only the SRT to classify 106 ears from 98 subjects into pass and fail, we observed an ACC equal to 0.82, SEN equal to 0.7, and SPE equal to 0.9 using a cut-off SRT value of -8 dB SNR, i.e. a value close to the cut-off value found in the root nodes of the DT models in this study (Fig. 3, Fig. 5). The AUC measured across the different cut-off SRTs was equal to 0.84 [20]. The values here observed with the DT approach are slightly lower than the values previously observed on a smaller dataset but still in line with the average classification performance of speech-in-noise tests. In general, a moderate level of accuracy is expected due to the inherently different nature of the two hearing tests compared (i.e. pure-tone detection and speech-in-noise recognition), that involve different auditory functions. The performance of other speech-in-noise tests based on multiple-choice recognition of short words is similar or lower than the one here observed. For example, the Earcheck and the Occupational Earcheck (i.e. Internet-based adaptive speech-in-noise tests based on multiple-choice recognition of consonant-vowel-consonant words) had a sensitivity of 0.51 and 0.92 and a specificity of 0.90 and 0.49, respectively, for the detection of ears with noise-induced hearing loss [28]. Similarly, another study on the Occupational Earcheck in a noise-exposed population showed a sensitivity of 0.65 and specificity of 0.63 to detect high-frequency hearing loss above 25 dB HL. Another

example is the Speech Understanding in Noise (SUN) test, that uses a list of VCV stimuli in a three-alternatives multiple-choice task presented at predetermined SNRs. The test, when administered sequentially in both ears reached a sensitivity and specificity of about 0.85 for detecting disabling hearing impairment (i.e., PTA > 40 dB HL) [14]. Similarly, the original version of the digits-in-noise test delivered by telephone yielded a sensitivity of 0.75 and a specificity of 0.91 to identify ears with PTA higher than 20.6 dB HL and for the U.S. version of the digits-in-noise test a sensitivity of 0.8 and a specificity of 0.83 to identify ears with PTA higher than 20 dB HL were reported [9], [30]. Therefore, considering the performance of the DT model with four features and the clarity and coherence of the rules generated by this simpler model, so far the DT with SRT, age, percentage of correct responses, and average reaction time as input features is the best candidate model for implementation into the smartphone app for the sake of identifying ears with mild hearing loss.

In the second part of the study, we have analysed the quantitative relationship between the self-adjusted volume levels of the test and the actual output levels in terms of SPL values using a range of consumer transducers, including headphones and earphones. This is relevant for the sake of implementation of a smartphone app as the app may be used with unknown transducers and the actual output levels, i.e. the level at which the speech-in-noise stimuli are delivered to the users' ears, may vary significantly.

Results in Fig. 6 show that the characteristics of different transducers vary greatly across the range of volume levels enabled by the device. The headphones yielded overall higher output SPL compared to earphones, with no substantial differences between types of earphones and types of headphones, except for the Bose QuiteComfort II headphones with noise canceling mode OFF that showed the broadest dynamics and the highest SPL values across the entire volume range of the device.

In general, these results suggest that including the self-adjusted volume option in a future app may help compensate, at least in part, for the different characteristics of the chosen transducer as the subject has the option to set the volume at a comfortable level based on his/her actual loudness perception, which depends on the actual SPL of the stimuli all the other things being equal (hearing sensitivity, mobile device characteristics, transducer, and environmental noise levels). This provides further support to the choice of incluing a self-adjusted volume option to enable test delivery via a mobile app in uncontrolled environment and with unknown transducers.

The results in Fig. 6 also point out that the range of output SPLs that can be generated with earphones is, in general, narrower than the one that can be obtained with headphones, with earphones providing much lower SPL values than headphones – up to about 20-25 dB SPL lower when a volume in the range 50-70% is used. This means that, even by using the full range of available output levels, the maximum SPL that can be obtained by using the tested earphones will be between 65 and 70 dB. As a benchmark, conversational speech occurs at an average of 65 dB SPL with a typical dynamic range of 30 dB (12 dB above and 18 dB below the average) [31-32]. Therefore, the maximum SPL reached by the earphones models here tested might not be sufficient to provide clearly intelligible speech stimuli to subjects with reduced hearing sensitivity due hearing loss. For example, subjects with minimal hearing loss as defined by the

WHO criterion would hear the test sounds attenuated by an average of 25 dB and the perceived sounds will be even weaker in case of mild or moderate hearing losses with higher PTA. For example, in case of PTA equal to 25 dB HL (which is the lowest PTA in the mild hearing loss range), a sound at 70 dB SPL would be perceived, on average, in a similar way as a sound of 45 dB SPL, a level that corresponds to faint speech and that would be therefore barely perceivable. Moreover, it may also happen that some of the tested subjects may not wish to adjust the device volume, as we observed for example in our listening tests. In fact, the average volume measured in our experiments was 0.48 (standard deviation = 0.11) and 83 out of 48 participants left the volume unchanged to the default value of 50%. Therefore, these results indicate that it may be important to recommend that headphones are used rather than earphones in combination with the smartphone testing app.

5 Limitations and Future Work

This study provided promising results but it has some limitations. First, the values of performance observed following a single model optimization process are slightly different than the average performance estimated across several iterations of the process, as shown in Table 1. This discrepancy is mainly due to the limited size of our dataset therefore it happens that a single optimized model may not accurately reflect the potential performance of the method in this task. Future studies will be necessary, on a significantly larger sample of adults and older adults, to increase the sample size and therefore improve the quality of the classifier developed before the algorithm can be safely implemented into a mobile app.

Second, in this study we investigated only one type of classification algorithm, the DT, due to its widespread use and its ability to generate intelligible decision rules. However, to fully explore the potential of multivariate classification algorithms for the identification of hearing loss and the possible advantages of machine learning methods over univariate classification approaches further research is needed. For example, it will be important to test different DT models by varying the set of input features and different machine learning algorithms, including both explainable (rule-based) and black-box approaches. The final outcome of this future investigation will be the identification of an optimal classification model that is accurate and reliable in identifying ears with hearing loss and that can be implemented into a smartphone app as part of the testing software.

In addition, the test partially addresses the problem of language-dependence. On the one hand, the use of meaningless VCVs and a set of consonants that is common across some of the top spoken languages worldwide makes the test feasible for use in non-native listeners. On the other hand, listeners who are not familiar with the Latin alphabet or native listeners of character-based languages such as Mandarin or Japanese may be subjected to ambiguities arising from the phoneme-grapheme correspondence. Investigation of test performance in a larger sample of native languages would be important to understand the possible modifications needed to improve test validity on a broader population.

Although we are getting very close to our final goal, further research is still needed to fully define the optimal settings for test delivery via a smartphone app. For example, it will be important to measure the output levels obtained using mobile devices to understand the potential output levels range that can be obtained using a smartphone or a tablet device. Moreover, it will be useful to address the characteristics of additional transducers, including additional earphones models with broader dynamic range, to understand which models are more suited for test execution and define minimum requirements for transducers to be used in the test in uncontrolled environmental conditions.

6 Conclusions

In this study, we have shown that an approach based on explainable machine learning using a decision tree algorithm in combination with the proposed speech-in-noise test can classify ears with similar accuracy, sensitivity, and specificity as that of popular validated speech-in-noise tests. Moreover, we have shown that a simplified classification model with a reduced set of features can lead to simpler, more intelligible rules and maintain the same performance of more complex models based on a larger set of input features. Regarding the analysis of the output levels of the test as a function of different transducers, we have highlighted important differences in output levels between types of transducers (headphones vs earphones) as well as between different transducer models, also showing that some transducers may not be adequate to ensure an appropriate range of output levels for the sake of conducting the test in self-adjusted volume settings. Further research will be needed to fully address the accuracy of the test, specifically it will be important to collect data from a larger sample of subjects, to evaluate the performance of different machine learning algorithms, preferably those able to generate explainable (e.g., rule-based) models, and to characterize the reliability of the test in varying environmental noise conditions and using different devices.

7 Acknowledgement

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 or ERC Consolidator Grant: SONORA (773268). This article reflects only the authors' views, and the Union is not liable for any use that may be made of the contained information.

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 from Politecnico di Milano who helped with data collection at Associazione La Rotonda. The Authors would also like to thank Dr. Randall Ali from the Department of Electrical Engineering at KU Leuven for providing guidance during the transducers characterization experiment.

References

1. Swanepoel, D.W.: eHealth Technologies Enable more Accessible Hearing Care. Seminars in Hearing 41(02), 133-140 (2020).
2. Paglialonga, A.: eHealth and mHealth for Audiologic Rehabilitation. In: Montano, J.J., Spitzer, J.B. (eds) Adult Audiologic Rehabilitation. 3rd Edition, Plural Publishing, San Diego, CA, USA (2020).
3. Paglialonga, A., Cleveland Nielsen, A., Ingo, E., Barr, C., & Laplante-Lévesque, A.: eHealth and the hearing aid adult patient journey: A state-of-the-art review. BioMedical Engineering Online, 17, 101 (2018).
4. Paglialonga, A., Tognola, G., Pinciroli, F.: Apps for hearing science and care. American Journal of Audiology 24(3), 293-298 (2015).
5. Audiology Today: Tele-audiology in a Pandemic and Beyond: Flexibility and Suitability in Audiology Practice. https://www.audiology.org/audiology-today-julyaugust-2020/tele-audiology-pandemic-and-beyond-flexibility-and-suitability, last accessed 2020/07/24.
6. Bright, T., Pallawela, D.: Validated smartphone-based apps for ear and hearing assessments: A review. JMIR Rehabilitation and Assistive Technology 3(2), e13 (2016).
7. Colsman, A., Supp, G.G., Neumann, J., Schneider, T.R.: Evaluation of Accuracy and Reliability of a Mobile Screening Audiometer in Normal Hearing Adults. Frontiers in Psychology 11, 744 (2020).
8. De Sousa, K.C., Swanepoel, D.W., Moore, D.R., Smits, C.: A smartphone national hearing test: Performance and characteristics of users. American Journal of Audiology, 27(3S), 448-454 (2018).
9. Smits, A., Kapteyn, T., Houtgast, T.: Development and validation of an automatic speech-in-noise screening test by telephone. International Journal of Audiology, 43, 1-28 (2004).
10. Nachtegaal, J., Festen, J. M., Kramer, S. E.: Hearing Ability and its Relationship with Psychosocial Health, Work-Related Variables, and Health Care Use: The National Longitudinal Study on Hearing. Audiology research 1(1), e9. (2011).
11. Reed, N.S., Altan, A., Deal, J.A., Yeh, C., Kravetz, A.D., Wallhagen, M., Lin F.R.: Trends in Health Care Costs and Utilization Associated With Untreated Hearing Loss Over 10 Years. JAMA Otolaryngol Head and Neck Surgery, 145(1), 27-34 (2019).
12. Humes, L.E.: Understanding the speech-understanding problems of older adults. American Journal of Audiology, 22(2), 303-305 (2013).
13. Killion, M.C., Niquette, P.A., Gudmundsen, G.I., Revit, L.J., Banerjee, S. 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, 2395-2405 (2004).
14. Paglialonga, A., Tognola, G., Grandori, F.: 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 (2014).
15. Jansen, S., Luts, H., Wagener, K.C., Frachet, B., Wouters, J.: The French digit triplet test: A hearing screening tool for speech intelligibility in noise. International Journal of Audiology, 49(5), 378–387 (2010).
16. Potgieter, J.M., Swanepoel, W., Myburgh, H.C., Smits, C.: 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 (2019).
17. Leek, M.R.: Adaptive procedures in psychophysical research. Perception & Psychophysics, 63(8), 1279-1292 (2001).

18. Cooke, M., Lecumberri, M.L.G., Scharenborg, O., van Dommelen, W.A.: Language-independent processing in speech perception: Identification of English intervocalic consonants by speakers of eight European languages. Speech Communication, 52, 954-967 2010.
19. Rocco G.: Design, implementation, and pilot testing of a language-independent speech intelligibility test. M.Sc. Thesis Dissertation, Department of Electronics Information and Bioengineering, Politecnico di Milano, Milan, Italy (2018).
20. Paglialonga, A, Polo E.M, Zanet M, Rocco G, van Waterschoot T, Barbieri R.: An automated speech-in-noise test for remote testing: development and preliminary evaluation. American Journal of Audiology, 29(3S), 564-576 (2020).
21. Zanet, M., Polo, E.M., Rocco, G., Paglialonga, A., Barbieri, R.: 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) (2019).
22. Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R.,et al: An international comparison of long-term average speech spectra. The Journal of the Acoustical Society of America, 96(4), 2108–2120 (1994).
23. Leensen, M.C., de Laat, J.A., Snik, A.F., Dreschler, W.A.: 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 (2011).
24. García-Pérez, M.A.: Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties. Vision Research, 38(12), 1861-1881 (1998).
25. World Health Organization (WHO), Grades of hearing impairment, https://www.who.int/pbd/deafness/hearing_impairment_grades/en/#, last accessed 2020/08/02.
26. Raileanu L.E., Stoffel K.: Theoretical comparison between the Gini Index and Information Gain criteria. Annals of Mathematics and Artificial Intelligence, 41, 77-93 (2004).
27. Polo, E.M., Zanet, M., Paglialonga, A., Barbieri, R.: Preliminary evaluation of a novel language independent speech-in-noise test for adult hearing screening. Proceedings of the 8th IFMBE European Medical and Biological Engineering Conference (EMBEC) (2020).
28. Leensen, M.C., de Laat, J.A., Dreschler, W.A.: 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 (2011).
29. Sheikh Rashid, M., Dreschler, W.A.: Accuracy of an Internet-based speech-in-noise hearing screening test for high-frequency hearing loss: Incorporating automatic conditional re-screening. International Archives of Occupational and Environmental Health, 91(7), 877-885 (2018).
30. Watson, C., Kidd, G., Miller, J., Smits, C., Humes, L.: 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 (2012).
31. Accredited Standards Committee S3, Bioacoustics. ANSI S3.5-1997, Methods for Calculation of the Speech Intelligibility Index (American National Standards Institute, New York) (1997).
32. Pearsons, K.S., Bennett, R.L., Fidell, S.: Speech Levels in Various Noise Environments (US Environmental Agency, Office of Health and Ecological Effects, Office of Research and Development, Washington, DC) (1977).