Evaluation of a Novel Speech-in-Noise Test for Hearing Screening at a Distance: Classification Performance and Transducers' Characteristics

Marco Zanet*, Edoardo M. Polo*, Marta Lenatti, Toon van Waterschoot, *Member, IEEE*, Maurizio Mongelli, Riccardo Barbieri, *Senior Member, IEEE*, and Alessia Paglialonga.

***Abstract*—One of the current gaps in teleaudiology is the lack of methods for adult hearing screening that are accurate and reliable in individuals of unknown language and in varying environments. Recently, we have developed a novel automated speech-in-noise test for future implementation via web and mobile platforms. The test uses speech material viable for use in nonnative listeners, is fast, and is reliable in laboratory settings and in uncontrolled environmental noise settings. The aim of this study was: (i) to evaluate the ability of the test to identify slight/mild hearing loss using a multivariate classifier and (ii) to evaluate the influence of transducers' characteristics on the ear-level sound pressure levels. The measures provided by the test had an accuracy of 0.79, sensitivity 0.79, specificity 0.79, and area under the curve of about 0.9, as measured in a population of 148 adults. The analysis of the ear-level sound pressure levels using several consumer transducers as a function of the test volume showed substantial variability, with earphones yielding up to 22 dB lower levels than headphones. Overall, these results suggest that the proposed approach may be a viable method for hearing screening at a distance if an option to self-adjust the volume is included and if headphones are used in uncontrolled environmental noise settings. Future research is needed to fully demonstrate the viability of the test for screening at a distance, for example by addressing test performance, including the influence of the user interface, device, and settings, on a large sample of participants with varying degrees of hearing loss.**

***Index Terms*—Hearing Screening, Mobile Applications, Speech-in-Noise Test, Teleaudiology, Telemedicine.**

Manuscript received Dec 31, 2020. This work was supported in part by 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.

M. Zanet, M. Mongelli, and A. Paglialonga are with the National Research Council of Italy (CNR), Institute of Electronics, Information Engineering and Telecommunications (IEIIT), Italy (e-mail: marco.zanet@ieiit.cnr.it; maurizio.mongelli@ieiit.cnr.it; corresponding author A. Paglialonga: phone: +39-02-23993343; e-mail: alessia.paglialonga@ieiit.cnr.it).

E. M. Polo is with DIAG, Sapienza University of Rome, Italy and with Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), 20133 Milan, Italy (e-mail: polo@diag.uniroma1.it).

M. Lenatti and R. Barbieri are with Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), 20133 Milan, Italy (e-mail: marta.lenatti@mail.polimi.it; riccardo.barbieri@polimi.it).

T. van Waterschoot is with KU Leuven, Department of Electrical Engineering (ESAT-STADIUS/ETC), Leuven, Belgium (e-mail: toon.vanwaterschoot@esat.kuleuven.be).

*Co-first authors (equal contribution)

I. Introduction

THE ongoing telehealthcare revolution is opening new opportunities to deliver audiological services at a distance, including hearing screening, hearing aid fitting, and adult audiologic rehabilitation [1]-[3]. The concept of teleaudiology was introduced as early as in 1994 by Cherry & Rubinstein [4] who suggested remote follow-up by telephone following face-to-face hearing aid fitting. Since then, a variety of services were created for delivery at a distance, via web or mobile app technology [2], [5]. Today, during the current pandemic emergency, teleaudiology services are even more necessary as patients with hearing loss are typically at the highest risk for COVID-19 due to their age. Today, key health authorities such as the CDC and the WHO lobby for ways to minimize physical contact between patients and healthcare providers [6]-[7].

From a more general perspective, the value of teleaudiology is widely recognized not only in contexts of reduced access to care (e.g., during and after a pandemic, in underserved areas, and in individuals with low socio-economic status) but also in usual care contexts. Teleaudiology has the potential to increase cost-efficiency, improve patient outcomes and satisfaction, and support widespread access to care [2], [8].

Increasing access to hearing health care is therefore a key challenge as there is substantial unmet need in adults and older adults. Hearing loss is one of the most important health burdens globally (about 466 million people with disabling hearing loss today, and over 900 million estimated by 2050 [9]) and is ranked by the World Health Organization as one of the top leading causes of number of years lived with disability globally [10]. Nevertheless, hearing loss is frequently considered as an inevitable component of aging and, typically, individuals with age-related hearing loss tend to seek help when it is too late or do not seek help at all. This leads to decreased quality of life and increased healthcare costs [11] as untreated hearing loss may trigger a cascade of effects that include isolation, depression, cognitive decline and dementia [12]-[13].

Early identification and timely intervention are key to limit the possible impact of hearing loss in older adults. Hearing screening can help increase awareness about hearing loss and its impact on communication, it can help identify individuals with

hearing loss early, when the first difficulties in communication occur, thus enabling timely intervention [14]-[15]. To identify the earlier effects of age-related hearing loss, speech-in-noise tests are particularly appropriate as one of the earliest complaints of older adults with hearing loss is just a decreased ability to understand speech in noisy environments (e.g., in crowded places, at the restaurant). Noticeably, many adults may experience difficulties in speech communication even when the outcomes of the standard clinical hearing test (i.e., pure tone audiogram) are within the normal limits [16]-[17].

Examples of self-administered speech-in-noise tests viable for use at a distance include: the digits-in-noise test, based on sequences of three random digits in speech-shaped noise and delivered in various formats (telephone, online, and mobile) [18]-[19]; the Earcheck and Occupational Earcheck online tests, based on consonant-vowel-consonant words in stationary masking noise [20]; the Speech Perception Test, an online test to address aided speech perception that uses speech features recognition for consonant-vowel-consonant words [21]; and the Speech Understanding in Noise (SUN) test, that uses a list of VCV stimuli in a three-alternatives multiple-choice task presented at predetermined signal-to-noise ratios (SNRs) [22]. However, none of the abovementioned speech-in-noise tests is readily applicable to widespread screening as these tests make use of speech material (e.g., words, digits) in specific languages and their accuracy in testing nonnative listeners is unknown.

This article presents original research towards the development of a novel methodology for widespread hearing screening. Specifically, in this study:

(i) we evaluate the performance of a newly developed speech-in-noise test in terms of ability to identify hearing loss in an unscreened population of adults, using an original multivariate classifier based on machine learning; and

(ii) we evaluate the influence of transducers' characteristics on the ear-level sound pressure levels of test stimuli to address the viability of the test for screening at a distance.

The novel test is an automated procedure for hearing screening at a distance (e.g., via web or mobile devices) that makes use of speech material viable for use in nonnative listeners [23]-[25]. The test is based on a user-operated speech-in-noise recognition task delivered in a three-alternatives multiple-choice format via an easy-to-use graphical interface. The recognition task is based on meaningless words, specifically vowel-consonant-vowel (VCV) stimuli (e.g., *aba*, *ada*, *afa*) in stationary speech-shaped noise. The set of stimuli includes 12 consonants common across some of the top spoken languages worldwide (i.e., English, Spanish, French, Portuguese, German, and Italian) [25]. The test is based on a novel one-up/three-down staircase [26]-[27] that uses optimized upward and downward steps and that is about two minutes shorter than conventional staircases both in individuals with normal hearing and with hearing loss [22], [28].

In an earlier study, we assessed the ability of the test to identify ears with pure-tone thresholds higher than 25 dB HL in the range from 1 to 4 kHz in a population of 98 unscreened adults [23]. Specifically, we used a univariate classifier based on the speech reception threshold (SRT), in line with the typical approach followed by previous studies in the literature (e.g., [18]-[21]). The results were promising as the accuracy was equal to 0.82, the area under the receiver operating characteristic (AUC) was equal to 0.84, and the test reliability was high, with no observable perceptual learning effects in test and retest trials [23], [28]. However, there is evidence that other features, in combination with the SRT, may be significant predictors of hearing loss - for example, the subject's age and the average reaction time [24], [29]. In this study, we addressed the ability of the test to identify hearing loss in a population of unscreened adults using an original approach, based on machine learning algorithms using a set of features in addition to the SRT (e.g., average reaction time, age, test duration, number and percentage of correct responses, and so on).

Regarding the viability of the test for screening at a distance, in a preliminary study we showed that the test provided consistent results in controlled laboratory settings and in uncontrolled environmental noise settings. Specifically, repeatable SRT estimates and similar test-retest repeatability were observed in normal hearing individuals who were given the option to self-adjust the output volume before the test in uncontrolled environmental noise settings [28]. However, for a full demonstration of the viability of the test for screening at a distance, a comprehensive analysis of the possible influence of the hardware and the environment is necessary. In fact, individuals performing the test at a distance, e.g. via a web or mobile app, may use different transducers and therefore the actual ear-level sound pressure levels of test stimuli will depend on the transducer's characteristic. Considering the complex nature of the speech signal, the actual sound pressure levels cannot be accurately estimated using the transducer's technical specifications (e.g., sensitivity, calibration table). To better understand the influence of the transducers' characteristics on the actual sound pressure levels of the test and identify possible minimum requirements for transducers to be used in the test, in this study we measured the characteristics of several consumer transducers, including commercially available headphones and earphones, as a function of the volume level set by the user.

II. Materials and Methods

A. Evaluation of Classification Performance

An outline of the experiment is shown in Fig. 1. Participants underwent pure-tone audiometry at 0.5, 1, 2, and 4 kHz in their left and right ears using a clinical audiometer (Amplaid 177+, Amplifon with TDH49 headphones). The pure-tone thresholds average (PTA) was computed as the average of hearing thresholds measured at the tested frequencies. Ears were classified into two classes using the World Health Organization (WHO) criterion for slight/mild hearing loss: PTA ≤ 25 dB HL (no hearing loss) and PTA > 25 dB HL (slight/mild hearing loss) [30]. Participants performed the newly developed speech-in-noise test in uncontrolled environmental noise

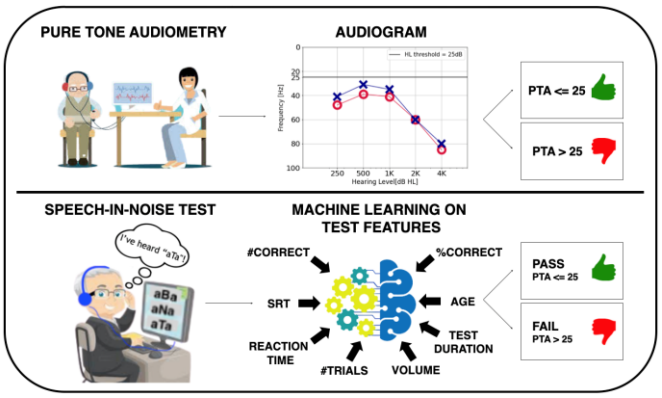

Fig. 1. Outline of the experiment to evaluate classification performance. Top panel: Pure-tone audiometry was performed at 0.5, 1, 2, and 4 kHz and the PTA was computed as the average of hearing thresholds measured at the tested frequencies. Ears were classified using the WHO criterion for slight/mild hearing impairment: PTA ≤ 25 dB HL (no hearing loss) and PTA > 25 dB HL (mild hearing loss). Bottom panel: The speech-in-noise test was executed in a self-administered way and eight features were extracted. Ears were classified by a machine learning approach into 'pass' and 'fail' using the PTA class as the target variable.
PTA = pure-tone threshold average; WHO = World Health Organization.

settings using a graphical user interface to adjust the volume at a comfortable level before the test. The test and graphical user interface were implemented in Matlab (MathWorks, version R2019b) and 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. A set of eight features were extracted from the 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, self-adjusted volume, and age. A machine learning approach was then used to classify ears into 'pass' and 'fail' considering the PTA class (no hearing loss vs slight/mild hearing loss) as the target variable.

The experiment was run on 148 unscreened adults (age = 52.1 ± 20.4 years; age range: 20-89 years; 46 male, 102 female) of varying native language (Italian, English, French, German, Spanish, Filipino, Efik, and Igbo). Participants were recruited and tested in opportunistic health screening initiatives (i.e. at universities of senior citizens, health prevention and awareness events for the public) to reflect the potential target group of typical screening initiatives and a realistic proportion of subjects with and without hearing loss . Eight out of 148 participants performed the test in both ears and 140 performed the test only in one ear, resulting in 156 ears tested. The experimental protocol was approved by the Politecnico di Milano Research Ethical Committee (Opinion n. 2/2019, Feb 19 2019). Participants received detailed information about the protocol and took part in the experiment on a voluntary basis.

To evaluate the classification performance of the test, a logistic regression algorithm was used in this study following preliminary evaluations that showed improved classification performance of this algorithm compared to other widely used ones (e.g., decision tree, support vector machine, k-nearest neighbor, random forest) [31]. The dataset was split randomly into training (80% of the sample, 124 ears) and test (20% of the sample, 32 ears) datasets. Stratification was applied to maintain the same percentage of records in the two PTA classes in the original dataset and in the training and test partitions. Considering the relatively small size of the dataset, the classification model was optimized using 5-fold cross-validation on the training dataset and its predictions were tested on the test dataset. The performance of the classification model was assessed by measuring accuracy on the training dataset (i.e., the average accuracy obtained following 5-fold cross-validation), accuracy on the test dataset, AUC, sensitivity, specificity, and F1-score. Given the relatively small sample size, we addressed the variability of classification performance by changing the underlying data. Specifically, we run 1000 iterations of the model optimization process on 1000 random partitions of the training and testing datasets and we computed the average and standard deviation of the abovementioned performance measures.

B. Evaluation of the Sound Pressure Levels

An outline of the experiment to evaluate the actual sound pressure levels of the test with different consumer transducers is shown in Fig. 2. The ear-level sound pressure levels of the test obtained with different consumer transducers across the full range of test volume were measured in the lab using a dummy head. First, an audio file was created by joining the 12 VCV recordings with no pauses. The file was recorded via the dummy head (Neumann KU 100 dummy head powered by an external P48 phantom power supply) and a sound card (RME Babyface Pro) with low input gain. The same laptop computer described in Section II.A was used, coupled with eight different transducers, across the full range of output volumes (0 to 100%, in 6.25% steps) (Fig. 2(A)).

The following consumer transducers were evaluated, covering a price range from €9.99 to €299: Bose Quitecomfort II headphones with noise canceling mode ON and OFF, Sony MDRZX110APW headphones, Sony MDR-7506 headphones, Sennheiser PC 310 headphones, Akg Y45 headphones, Apple

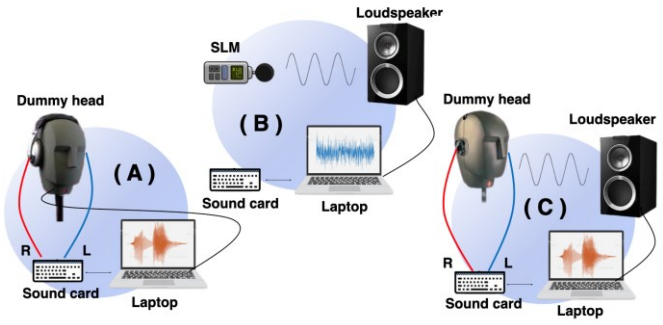

Fig. 2. Outline of the experiment to evaluate the sound levels of the test with different transducers. Panel (A): recording of the sequence of VCV stimuli via the dummy head using the different transducer models across the full range of volume output levels; Panel (B): calibration of the sound card by adjusting the output gain to reach a white noise level of 90 dB SPL at the SLM. Panel (C): recording of the sequence of VCV stimuli via the dummy head using the setup and output gain set in (B). R = right; L = left; SLM = sound level meter.

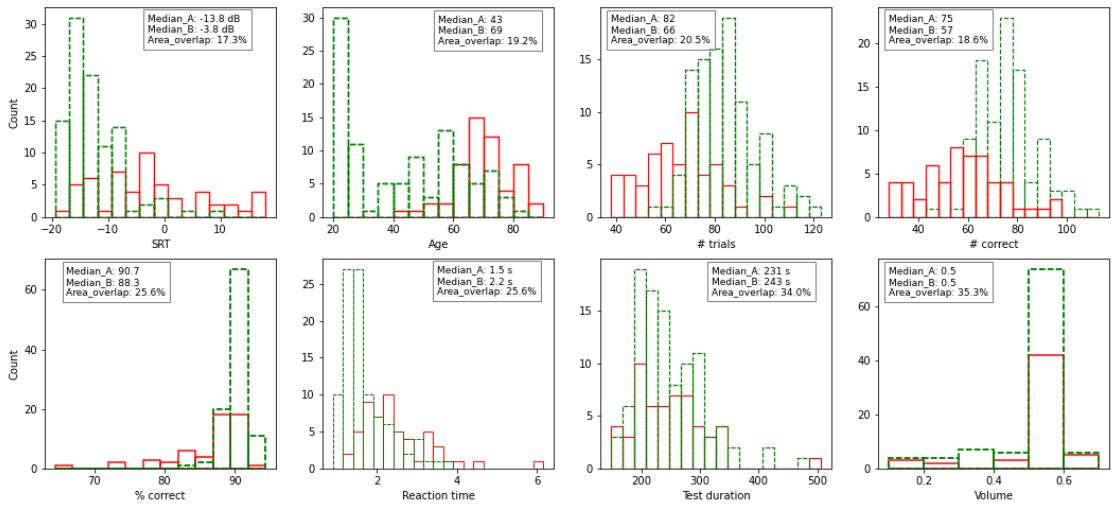

Fig. 3. Distributions of the eight features in the two PTA classes: (A) PTA ≤ 25 dB HL (no hearing loss): dashed line; (B) PTA > 25 dB HL (slight/mild hearing loss): continuous line. The text boxes show the median values in the two PTA classes and the percent overlap between the frequency distributions.

EarPods earphones, and Mpow BH319 wired In-ear earphones.

Then, the setup was calibrated to convert wave units into sound pressure levels (SPL), as shown in Fig. 2(B) and Fig. 2(C). Calibration was performed using white noise, as the recorded loudness of this wideband signal is minimally influenced by acoustical attenuation. The output gain of the sound card was adjusted to reach a white noise level of 90 dB SPL as measured by a Sound Level Meter (SLM; Brüel & Kjær Type 2250 Hand Held Analyzer with BZ-7222 Sound Level Meter Software) at a distance of 1 meter from a loudspeaker. Finally, the SLM was removed and the dummy head, placed in the same position as the SLM, was used to record the white noise at the adjusted sound card output gain. This calibration procedure was used to convert the level of the recorded sequences of VCV stimuli in dB SPL. A-weighted filtering was applied to the audio files to approximate the SPL perceived by the average human ear.

III. Results

A. Characterization of features

The distributions of the eight features in the two PTA classes (no hearing loss vs slight/mild hearing loss) are shown in Fig. 3. Fig. 3 also shows, for each feature, the median values in the two classes as well as the percent overlap between the distributions of the two classes, computed from the probability distributions. In general, the distributions of features such as SRT, age, #trials, #correct, %correct, and average reaction time were different between the two classes whereas features such as test duration and volume did not change substantially. Statistical analysis (Wilcoxon rank sum test for SRT, age, %correct, reaction time, test duration, volume; t-test for #trials and #correct) showed that the observed differences in SRT, age, #trials, #correct, %correct, and average reaction time between the two classes were statistically significant whereas the differences between test duration and test volume were not significant. The percent overlap ranged from about 17% to about 35% overall. The observed values of percent overlap suggest that features such as test duration and volume tended to have a similar distribution in the two classes, with percent overlap around 35%. Features such as SRT, age, and #correct had more distinct distributions in the two classes, with percent overlap below 0.2. Therefore, these features are likely to contribute more substantially to the classification algorithm compared to the features with higher percent overlap. Accordingly, the classification algorithm was built using both the full set of eight features and a reduced set of features, specifically SRT, age, and #correct.

B. Classification Performance

Table I shows the average and standard deviation of the observed performance measures (accuracy on training dataset, accuracy on test dataset, AUC, sensitivity, specificity, and F1-score) computed over 1000 iterations for the logistic regression classifier using both the full set of eight features as well as a reduced set of three features (SRT, age, and #correct).

Overall, the average performance of the models with eight and three features was strikingly similar. Minor differences were observed, lower than 0.01, in terms of average AUC, sensitivity, specificity, and F1-score. The observed differences were not statistically significant, except for the accuracy on the

TABLE I
CLASSIFICATION PERFORMANCE OVER 1000 ITERATIONS

Measure	8 features	3 features
Accuracy (training)	0.80 ± 0.02	0.81 ± 0.02
Accuracy (test)	0.79 ± 0.07	0.79 ± 0.07
AUC	0.89 ± 0.05	0.90 ± 0.05
Sensitivity	0.79 ± 0.13	0.79 ± 0.13
Specificity	0.79 ± 0.09	0.79 ± 0.10
F1-score	0.72 ± 0.09	0.72 ± 0.09

training test and AUC (t-test, $p < 0.05$). Slight, but statistically significant, differences in accuracy were observed between the training and test datasets (0.01 in the model with eight features and 0.02 in the model with three features; t-test, $p < 0.05$), suggesting a sufficiently stable performance and therefore limited overfitting effects. The AUC was about 0.9, with a slight increase in the model with three features, indicating very good classification performance. Sensitivity and specificity were, on average, around 0.79, which is a relatively high value considering the different nature of the new test (that measures the ability to recognize speech in background noise) and the target outcome, i.e. the degree of hearing impairment defined by the average pure-tone thresholds (that measure hearing sensitivity to detect simple frequency tone stimuli).

The observed values of standard deviation for all the performance measures were relatively low and very similar for the two models, suggesting that the variability of the model performances was inherently related to changes in the underlying datasets across the 1000 iterations rather than to the input features used by the algorithms. The observed standard deviations were, overall, smaller than 0.1 for all the measures, except for sensitivity for which it was about 0.13 for the two models. This is possibly due to the lower number of ears in the slight/mild hearing loss class compared to the no hearing loss class that may have led to a higher variability of the number of ears correctly classified as 'fail'.

C. Sound Pressure Levels

Table II shows the sound pressure levels measured with the different transducers here tested as a function of the test volume at the following percent volume levels: 25%, 50% (default settings), 75%, and 100%.

Overall, lower sound pressure levels were measured with the earphones compared to headphones, with maximum output levels of about 67 and 68 dB SPL with the Apple EarPods and Mpow in-ear, respectively. At the default volume level of 50%, earphones reached sound pressure levels lower than or equal to 50 dB SPL whereas headphones provided sound pressure levels in the range from 60 to 70 dB SPL. Among the headphones models here used, the highest output levels were observed with the Akg Y45 BT that provided an output equal to about 71 dB SPL at the default volume level of 50% and up to about 90 dB SPL when the maximum volume level was used. The Bose QuiteComfort II with noise canceling mode OFF and the Sony MDRZX110APW headphones used in the first part of this study showed similar characteristics, with differences below 1 dB across the volume range. Differences of about 20-22 dB were observed, at each of the tested volume levels, between the earphones and the Akg Y45 BT headphones.

TABLE II
TRANSDUCERS SOUND LEVELS (DB SPL) AS A FUNCTION OF THE PERCENT VOLUME LEVEL

Transducers models	25%	50%	75%	100%
Apple Earpods	35.30	48.06	58.09	66.58
Mpow BH319 wired In-ear	37.49	50.13	60.38	68.12
Bose QuiteComfort II (n.c. ON)	46.79	60.03	69.84	78.35
Sennheiser PC310	47.31	60.33	71.00	78.85
Sony MDR-7506	50.10	63.11	73.24	81.72
Bose QuiteComfort II (n.c. OFF)	52.23	65.26	76.00	84.01
Sony MDRZX110APW	52.97	66.03	76.01	85.20
Akg Y45 BT	57.26	70.93	80.92	89.44

IV. DISCUSSION

The first aim of this study was to address the performance of logistic regression for classifying ears with slight/mild hearing loss vs ears with no hearing loss in a population of unscreened adults using a previously developed automated speech-in-noise test [23], [28]. Using eight input features and a subset of three input features, we observed similar average performance, as determined by running 1000 iterations of model optimization on different realizations of the training and test datasets (Table I). In addition, the variability of classification performance, as measured by the standard deviation of the performance measures across the 1000 iterations, was strikingly similar between the two models. Features such as SRT, age, and number of correct responses in the test had more distinct distributions in ears with and without slight/mild hearing loss compared to features such as, e.g., average reaction time, test volume, and test duration. Accordingly, the simpler model using only SRT, age, and #correct as input features had a similar performance as the model using the full set of eight features. Specifically, the same accuracy was obtained (i.e., 0.79) and a slightly improved AUC for the model with three features compared to the one with eight features (0.90 vs 0.89).

Compared to our earlier investigations, where different classification algorithms were applied, the logistic regression model here used provided better classification performance. For example, when only the SRT was used to classify 106 ears from 98 subjects into pass and fail (cut-off SRT = -8 dB SNR), we observed an accuracy equal to 0.82, sensitivity equal to 0.70, specificity equal to 0.90, and AUC equal to 0.84 [23]. When a decision tree algorithm with full set of eight features was used on the same dataset here used (156 ears from 148 subjects), the average accuracy was 0.76, sensitivity was 0.67, specificity 0.81, and the AUC was equal to 0.74 [24]. The results of this study showed that the logistic regression algorithm outperformed the decision tree and that a subset of three features (SRT, age, and #correct) was appropriate for the sake of identifying ears with slight/mild hearing loss, yielding the same average performance and the same variability as the model with eight features. The importance of age, in addition to SRT, was suggested by a preliminary study where, using a generalized linear model with age and SRT as input variables on a dataset of 91 ears from 84 subjects we observed that the interaction between age and SRT (and not age as a single factor) was a significant predictor of hearing loss [29]. The results of the current study suggest that the number of correct responses obtained in the test, in addition to SRT and age, could be an important factor to determine the hearing loss class. Moreover, the degree of overlap between the distributions of #correct in the two classes was similar to the overlap observed for the distributions of SRT and age and was below 0.2.

Compared to the classification performance of other speech-in-noise tests based on multiple-choice recognition of short words, the logistic regression model using age in combination with two features extracted from the newly developed speech-in-noise test (i.e., SRT and #correct) showed similar if not better performance in identifying ears with hearing thresholds in the slight/moderate hearing loss range. For example, for the digits-in-noise test delivered by telephone a sensitivity equal to 0.75 and a specificity equal to 0.91 to identify ears with average hearing thresholds higher than 20.6 dB HL were observed [18]. Similarly, the sensitivity and specificity of the U.S. version of the digits-in-noise test were 0.8 and 0.83, respectively [19]. The Earcheck and the Occupational Earcheck online tests had a sensitivity of 0.51 and 0.92 and a specificity of 0.90 and 0.49, respectively, for the identification of ears with noise-induced hearing loss [20].

Taken as a whole, the results are encouraging as very good classification performance is obtained using logistic regression and a reduced set of features. However, further research is needed to demonstrate the viability of the proposed approach for adult hearing screening. It will be important to investigate test performance on a larger sample of participants, including subjects with varying degrees of hearing loss and across a larger set of native languages. Also, it would be interesting to investigate the ability of the proposed approach to identify the degree of hearing loss, e.g. moderate hearing loss (PTA > 40 dB HL) vs slight/mild hearing loss vs normal hearing. In the dataset used in this study, only 18 ears had moderate hearing loss therefore analyzing further data from a population of adults with hearing thresholds in the moderate hearing loss range would be crucial to investigate this aspect. It would also be important to address the performance of the test when delivered via interfaces that mimic those of web browsers or mobile devices and to compare it with other validated speech-in-noise screening tests, to fully understand the viability of the method in realistic settings in light of other currently available methods.

The second aim of this study was to address the characteristics of several consumer transducers to estimate the actual sound pressure levels of the test (i.e., the level at which the speech-in-noise stimuli are delivered to the users' ears) as a function of the test volume. In general, an increase in test volume from the default level of 50% to the maximum device level corresponded to an increase of about 18 dB in the actual sound pressure levels irrespectively of the transducer used. For example, the sound pressure levels increased from about 50 to about 68 dB SPL with the Mpow BH319 wired In-ear earphones and from about 71 to about 90 dB SPL with the Akg Y45 BT headphones (Table II). Considering the full range of volumes shown in Table II, i.e. from 25% to 100%, the resulting dynamic range at the level of the ears is about 32 dB. With the prospect of an application of the test for screening at a distance, a measured dynamic range of 32 dB suggests that the end users have relatively ample room to adjust the sound pressure levels of the test and reach a comfortable loudness level, which depends on the ear-level SPL all the other things being equal (hearing thresholds, environmental noise, and device characteristics). Individuals with slight/mild hearing loss, for example, would hear the test stimuli attenuated by at least 25 dB compared to individuals with hearing thresholds close to the ideal value of 0 dB HL and they may therefore feel the need to increase the test volume to reach a sufficiently audible level. It is worth noting that in speech-in-noise tests, such as the one here used, the test outcome (i.e., the SRT) is related mainly to the individual speech recognition performance and hearing thresholds rather than to the absolute sound pressure levels as long as these levels are set at a comfortable level [32]. Therefore, having a volume adjustment option incorporated in the test is important in view of future implementation of the test into a web or mobile app as users can, at least in part, compensate for possibly lower-than-ideal audibility of test stimuli due, for example, to environmental noise, higher individual hearing thresholds, or lower transducer gain.

However, it is important to notice that substantial variability of sound pressure levels was observed across the range of transducers here tested, with differences in SPL up to 22 dB. All the headphones here tested reached substantially higher sound pressure levels compared to earphones, with the Akg Yx45 BT headphones yielding the highest output levels across the range of test volumes. The two models of earphones had substantially similar characteristics, with differences in SPL of about 2 dB and maximum levels up to 68 dB SPL. These maximum levels are close to the average conversational speech levels. Conversational speech occurs at an average of 65 dB SPL and has a typical dynamic range of 30 dB, i.e. about 12 dB above and about 18 dB below the average [33]-[34]. Therefore, with commercially available earphones similar to the ones here tested and in the current settings, individuals with elevated hearing thresholds due to hearing loss and individuals performing the test in a noisy environment might not be able to reach sufficiently comfortable speech levels with the volume adjustment procedure even if the maximum test volume is set. In addition, it might happen that individuals undergoing the test in an unsupervised way on a web or mobile app may not use the volume adjustment option at all. For example, in this study we observed a median test volume of 0.5 both in the no hearing loss class (s.d. = 0.11) and in the slight/mild hearing loss class (s.d. = 0.12), with 83 out of 148 participants using the default volume level of 50%, corresponding to about 66 dB SPL with the Sony MDRZX110APW headphones. Thus, as a general criterion it would be safer to recommend use of headphones in place of earphones to enable higher sound pressure levels when the default volume settings are used. Specific benchmarks, in terms of minimum requirements for transducers characteristics to reach a desired ear-level SPL cannot be set as, in this study, the measured output levels were derived from a specific laptop model. In future studies, it will be important to assess the amount of change in the actual sound pressure levels when different devices are used, for example different computers, smartphones, and tablets. This will help identify minimum requirements for devices and transducers to deliver the test.

In summary, this study showed that the newly developed test combined with a multivariate classification algorithm may be a viable method for identifying slight/moderate hearing loss at a distance and that the self-adjustment volume option may help compensate for the different transducers used. However, results also indicate that it may be important to recommend use of headphones rather than earphones to ensure sufficiently high sound pressure levels, particularly in individuals with hearing loss. Future research is needed to investigate the viability of the test for screening at a distance. It will be important to address test performance, including the possible influence of the user interface, device, and settings, on a large sample of participants with varying degrees of hearing loss, also including subjects with moderate hearing loss.

ACKNOWLEDGMENT

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 unscreened adults. The Authors would also like to thank Anna Bersani, Carola Butera, and Antonio Carrella who contributed to the experiment on unscreened adults.

REFERENCES

[1] K. F. M. Tao, C. G. Brennan-Jones, D. M. Capobianco-Fava, D. M. P. Jayakody, P. L. Friedlandm, et al., "Teleaudiology services for rehabilitation with hearing aids in adults: A systematic review," *J. Speech Lang. Hear. Res.*, vol. 61, pp. 1831–1849, 2018.

[2] A. Paglialonga, A. Cleveland Nielsen, E. Ingo, C. Barr, and A. Laplante-Lévesque, "eHealth and the hearing aid adult patient journey: A state-of-the-art review," *BioMed. Engin. Online*, vol. 17, no. 101, 2018.

[3] A. Paglialonga., "eHealth and mHealth for Audiologic Rehabilitation," in *Adult Audiologic Rehabilitation,* 3rd ed. J. J. Montano, J. B. Spitzer Eds. San Diego: Plural Publishing, 2020, pp. 491-512.

[4] R. Cherry and A. Rubinstein, "The effects of telephone intervention on success with amplification," *Ear Hear.*, vol. 15, pp. 256–261, 1994.

[5] A. Paglialonga, G. Tognola, and F. Pinciroli, "Apps for hearing science and care," *Am. J. Audiol.*, vol. 24, pp. 293–298, 2015.

[6] D. W. Swanepoel and J. W. Hall, "Making audiology work during COVID-19 and beyond," *The Hearing Journal Online Only Blog*. April 21, 2020 Available: https://journals.lww.com/thehearingjournal/blog/OnlineFirst/pages/post.aspx?PostID=59

[7] B. Ballachanda, H. Abrams, J. W. Hall, V. Manchaiah, D. Minihane, S. et al., "Tele-audiology in a pandemic and beyond: Flexibility and suitability in audiology practice," *Audiology Today*, July/August 2020. Available: https://www.audiology.org/audiology-today-julyaugust-2020/tele-audiology-pandemic-and-beyond-flexibility-and-suitability

[8] G. Tognola, A. Paglialonga, E. Chiaramello, and F. Pinciroli, "eHealth for hearing – New views and apps practicalities," *Eur. J. Biomed. Inform.*, vol. 11, en43–en49, 2015.

[9] World Health Organization (WHO), *Deafness and hearing loss*, Fact Sheet n. 300. Available: http://www.who.int/mediacentre/factsheets/fs300/en/

[10] World Health Organization (WHO), World report on disability. WHO press, Geneve, Switzerland, vol. 295, p. 298, 2011. Available: http://whqlibdoc.who.int/publications/2011/9789240685215_eng.pdf

[11] N. S. Reed, A. Altan, J. A. Deal, C. Yeh, A. D. Kravetz, et al., "Trends in health care costs and utilization associated with untreated hearing loss over 10 years,". *JAMA Otolaryngol. Head and Neck Surg.*, vol. 145, pp. 27–34, 2019.

[12] D. S. Dalton, K. J. Cruickshanks, B. E. Klein, R. Klein., T. L. Wiley, et al., "The impact of hearing loss on quality of life in older adults," Gerontologist, vol. 43(5), pp. 661–668, 2003.

[13] H. R. Davies, D. Cadar, A. Herbert, M. Orrell, A. and Steptoe, "Hearing impairment and incident dementia: findings from the english longitudinal study of ageing," *J. Am. Ger. Soc.*, vol. 65(9), pp. 2074–2081, 2017.

[14] A. Davis and P. Smith, "Adult hearing screening: health policy issues-what happens next?" *Am. J. Audiol.*, vol. 22, pp. 167–170, 2013.

[15] A. Davis., P. Smith, M. Ferguson, D. Stephens, and I. Gianopoulos, "Acceptability, benefit and costs of early screening for hearing disability: a study of potential screening tests and models," *Health Technology Assessment*, vol. 11, no. 42, 2007.

[16] L. E. Humes, "Understanding the speech-understanding problems of older adults," *Am J Audiol*, vol. 22, pp. 303–305, Dec 2013.

[17] M. C. Killion and P. A. Niquette, "What can the pure-tone audiogram tell us about a patient's SNR loss?" *Hear J* , vol. 53, pp. 46–53, 2000.

[18] C. Smits, T. Kapteyn, and T. Houtgast, "Development and validation of an automatic speech-in-noise screening test by telephone," *Int. J. Audiol.*, vol. 43, pp. 1–28, 2004.

[19] C. Watson, G. Kidd, J. Miller, C. Smits, and L. E. Humes, "Telephone screening tests for functionally impaired hearing: Current use in seven countries and development of a US version" *J. Am. Acad. Audiol.*, vol. 23, pp. 757–767, 2012.

[20] M. C. Leensen, J. A de Laat, A. F Snik, and W. A. Dreschler, "Speech-in-noise screening tests by internet, part 2: improving test sensitivity for noise-induced hearing loss," *Int. J. Audiol.*, vol. 50, pp. 835–848, 2011.

[21] P. Blamey, J. Blamey, and E. Saunders, "Effectiveness of a Teleaudiology approach to hearing aid fitting," *J Telemed. Telecare*, vol. 2, pp. 474–478, 2015.

[22] A. Paglialonga, G. Tognola, and F. Grandori, "A user-operated test of suprathreshold acuity in noise for adult hearing screening: The SUN (Speech Understanding in Noise) test," *Comput Biol Med*., vol. 52, pp. 66–72, 2014.

[23] A. Paglialonga, E. M. Polo, M. Zanet, G. Rocco, T. van Waterschoot, et al., "An automated speech-in-noise test for remote testing: development and preliminary evaluation," *Am J Audiol.*, vol. 29, pp. 564–576, 2020.

[24] E. M. Polo, M. Zanet, M. Lenatti, T. van Waterschoot, R. Barbieri, A. and Paglialonga, "Development and evaluation of a novel method for adult hearing screening: Towards a dedicated smartphone app," Proc. 7th EAI Intern. Conf. IoT Technologies for HealthCare (EAI HealthyIoT 2020), Dec 2-4, 2020, virtual conference.

[25] G. Rocco, "Design, implementation, and pilot testing of a language-independent speech intelligibility test," M.Sc. dissertation, Dept. Electronics. Information and Bioengineering, Politecnico di Milano, Milan, Italy, Apr. 2018.

[26] M. R. Leek, "Adaptive procedures in psychophysical research," *Percept Psychoph.*, vol. 63(8), pp. 1279–1292, 2001.

[27] E. M. Polo, M. Zanet, R. Barbieri, and A. Paglialonga, "Development and Characterization of a Novel Adaptive Staircase for Speech Recognition Testing," *BioMed Engin OnLine*, submitted for publication.

[28] M. Zanet, E. M. Polo, G. Rocco, A. Paglialonga, and R. Barbieri, "Development and preliminary evaluation of a novel adaptive staircase procedure for automated speech-in-noise testing," *Proc. 41st Annual Intern. Conf. IEEE Eng. Med. Biol. Soc.*, Berlin, Germany, July 23-27 2019, pp. 6991-6994.

[29] E. M. Polo, M. Zanet, A. Paglialonga, and R. Barbieri, "Preliminary evaluation of a novel language independent speech-in-noise test for adult hearing screening," *Proc. 8th Eur. Med. Biol. Eng. Conf. (EMBEC), IFMBE Proceedings*, vol. 80, pp. 976-983.

[30] World Health Organization (WHO), *Grades of hearing impairment*, Available: https://www.who.int/pbd/deafness/hearing_impairment_grades/en/#

[31] M. Lenatti, "Automated detection of hearing loss by machine learning approaches applied to speech-in-noise testing for adult hearing screening," M.Sc. dissertation, Dept. Electronics. Information and Bioengineering, Politecnico di Milano, Milan, Italy, Dec. 2020.

[32] D. R. Moore, M. Edmondson-Jones, P. Dawes, H. Fortnum, A. McCormack, et al., "Relation between Speech-in-Noise Threshold, Hearing Loss and Cognition from 40–69 Years of Age,". PLoS ONE 9, e107720, 2014.

[33] *Methods for Calculation of the Speech Intelligibility Index*, ANSI Standard S3.5-1997.

[34] K. S. Pearsons, R.L. Bennett, and S. Fidell, "Speech Levels in Various Noise Environments (US Environmental Agency, Office of Health and

Ecological Effects, Office of Research and Development, Washington, DC)," 1977.

Marco Zanet was born in Rho, Italy, in 1994. He holds a M. Sc. in biomedical engineering from Politecnico di Milano, Italy (2019).

He worked as an Intern in the Product Development Group at Amplifon S.p.A., Milan, Italy and is currently working as Graduate Research Fellow at the talian National Research Council of Italy (CNR), Institute of Electronics, Information Engineering and Telecommunications (IEIIT), Milan, Italy. He is working on the "Segment-based Collective Intelligence for Population Health Improvement (SCI-PHI)" project and he is interested in the application of Machine Learning for Healthcare, eHealth, biosignal processing and audiology.

Edoardo Maria Polo was born in Milan, Italy, in 1994. He earned the M.Sc in biomedical engineering from Politecnico di Milano, Italy, in April 2019.

In November 2019, he was enrolled in the ABRO PhD in Bioengineering at University of Rome "La Sapienza", Rome, Italy. During 2020, he also worked as assistant professor for two courses of the Biomedical Engineering program at Politecnico di Milano, regarding medical informatics and bioengineering of neurosensory systems. His research activity deals with biomedical signal processing as a tool to unravel the impact of hearing problems on listening effort and our perception of emotions.

Marta Lenatti was born in Sondrio, Italy in 1996. She received the M.Sc. in biomedical engineering from Politecnico di Milano, Italy, in December 2020. Her research interests are related to machine learning methods and application and eHealth in audiology.

Toon van Waterschoot (S'04, M'12) received MSc (2001) and PhD (2009) degrees in Electrical Engineering, both from KU Leuven, Belgium, where he is currently an Associate Professor and Consolidator Grantee of the European Research Council (ERC). He has previously also held teaching and research positions at Delft University of Technology in The Netherlands and the University of Lugano in Switzerland. His research interests are in signal processing, machine learning, and numerical optimization, applied to acoustic signal enhancement, acoustic modeling, audio analysis, and audio reproduction.

He has been serving as an Associate Editor for the Journal of the Audio Engineering Society and for the EURASIP Journal on Audio, Music, and Speech Processing. He is a Director of the European Association for Signal Processing (EURASIP), a Member of the IEEE Audio and Acoustic Signal Processing Technical Committee, a Member of the EURASIP Special Area Team on Acoustic, Speech and Music Signal Processing, and a Founding Member of the EAA Technical Committee in Audio Signal Processing. He is a member of EURASIP, IEEE, ASA, and AES.

Maurizio Mongelli obtained his Ph.D. Degree in Electronics and Computer Engineering from the University of Genoa (UniGe) in 2004. The doctorate was funded by Selex Communications S.p.A. (Selex). He worked for both Selex and the Italian Telecommunications Consortium (CNIT) from 2001 to 2010. During his doctorate and in the following years, he worked on the quality of service for military networks with Selex. From 2007 to 2008, he coordinated a joint laboratory between UniGe and Selex, dedicated to the study and prototype implementation of Ethernet resilience mechanisms. He was the CNIT technical coordinator of a research project concerning satellite emulation systems, funded by the European Space Agency; spent three months working on the project at the German Aerospace Center in Munich. Since 2012 he is a researcher at the Institute of Electronics, Information Engineering and Telecommunications (IEIIT) of the National Research Council of Italy (CNR) in Genua, Italy, where he deals with machine learning applied to bioinformatics and cyber-physical systems, having the responsibility and coordination, for the CNR part, of funded projects (5, of which 1 at European level) in these sectors. He is co-author of over 100 international scientific papers and 2 patents.

Riccardo Barbieri (M'00, SM'08) received the M.S. degree in electrical engineering from the University of Rome "La Sapienza", Rome, Italy, in 1992, and the Ph.D. in biomedical engineering from Boston University, Boston, MA, USA, in 1998.

He is currently an Associate Professor in the Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy. His research interests include the development of signal processing algorithms for the analysis of biological systems, with focuses on computational modeling of neural information encoding, and on the application of nonlinear and multivariate statistical models to characterize heart rate variability and cardiovascular control dynamics.

Dr. Barbieri is a member of the American Association for the Advancement of Science, the European Society of Hypertension, the Society for Neuroscience, and the Engineering in Medicine and Biology Society.

Alessia Paglialonga received the M.Sc. (2005) and PhD (2009) degrees in biomedical engineering, both from Politecnico di Milano, Italy.

She is currently a researcher at the National Research Council of Italy (CNR), Institute of Electronics, Information Engineering and Telecommunications (IEIIT), Milan, Italy, Adjunct Professor at Politecnico di Milano, Italy, and Visiting Scientist at Ryerson University, Toronto, Canada. She has previously also held teaching and research positions at Politecnico di Milano and CNR. Her research interest include eHealth, audiological technology, predictive health modeling, biosignal processing.

She is serving as Associate Editor for BioMedical Engineering Online (BMC, Springer Nature) and the International Journal of Audiology (Informa Healthcare). She is member of the European Society of Cardiology and the European Alliance for Innovation.