Evaluation of a Novel Speech-in-Noise Test for Hearing Screening: 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 viable for use in individuals of unknown language and in varying environments. We have developed a novel automated speech-in-noise test that uses stimuli viable for use in non-native listeners. The test reliability has been demonstrated in laboratory settings and in uncontrolled environmental noise settings in previous studies. The aim of this study was: (i) to evaluate the ability of the test to identify hearing loss using multivariate logistic regression classifiers in a population of 148 unscreened adults and (ii) to evaluate the ear-level sound pressure levels generated by different earphones and headphones as a function of the test volume. The multivariate classifiers had sensitivity equal to 0.79 and specificity equal to 0.79 using both the full set of features extracted from the test as well as a subset of three features (speech recognition threshold, age, and number of correct responses). The analysis of the ear-level sound pressure levels showed substantial variability across transducer types and models, with earphones levels being up to 22 dB lower than those of headphones. Overall, these results suggest that the proposed approach might be viable for hearing screening in varying environments if an option to self-adjust the test volume is included and if headphones are used. Future research is needed to assess the viability of the test for screening at a distance, for example by addressing the influence of user interface, device, and settings, on a large sample of subjects with varying 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 Capita Foundation through project WHISPER, Widespread Hearing Impairment Screening and PrEvention of Risk (2020 Auditory Research Grant) and 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. Lenatti, M. Mongelli, and A. Paglialonga are with the National Research Council (CNR), Institute of Electronics, Information Engineering and Telecommunications (IEIIT), Italy (e-mail: marco.zanet@ieiit.cnr.it; marta.lenatti@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).

R. Barbieri is with Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), 20133 Milan, Italy (e-mail: 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, including hearing screening, hearing aid fitting, and adult audiologic rehabilitation [1]-[3]. Since the early teleaudiology work by Cherry & Rubinstein [4], a variety of eHealth services have been created, including web- and mobile-based platforms [2], [5]. During the current COVID-19 pandemic, teleaudiology is even more necessary to ensure continuity of hearing care services under the existing social distancing measures [6]. In fact, key health authorities such as the US Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO) have advocated for ways to minimize physical contact between patients and healthcare providers [6]-[7]. 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 to increase access and cost-efficiency [2], [8].

Increasing access to hearing health care is a key challenge as there is substantial unmet need, particularly in 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 WHO as one of the top leading causes of number of years lived with disability globally [10]. Nevertheless, individuals with age-related hearing loss tend to seek help when it is too late or do not seek help at all [11]. This leads to decreased quality of life and increased healthcare costs [12] as untreated hearing loss may trigger a cascade of effects including isolation, depression, and cognitive decline [13]-[14].

Hearing screening can help increase awareness about hearing loss and its impact on communication, it can help identify individuals with hearing loss early, and enable timely intervention [11], [15]. Speech-in-noise tests are particularly appropriate in hearing screening 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), 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 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 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 wide-scale screening as these tests make use of speech material (e.g., words, digits) in specific languages and their performance in non-native listeners may be different than in native ones. For example, a study investigating the performance of native and non-native English speakers in the digits-in-noise test showed a significant effect of self-reported English competency on the speech recognition threshold (SRT) and suggested the use of dynamic cut-off values depending on self-reported language competency and age [23].

Recently, we have developed a novel speech-in-noise test for hearing screening that makes use of speech material viable for use in non-native listeners and is based on a novel, optimized staircase procedure [24]-[28]. Compared to conventional, fixed step size staircase procedures the new test is as accurate and repeatable and it is about two minutes shorter in individuals with normal hearing and with hearing loss [22], [29].

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 [24]. We used a univariate classifier based on the 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 (sensitivity = 0.7, specificity = 0.9), the area under the receiver operating characteristic (AUC) was equal to 0.84, and test reliability was high, with no observable perceptual learning effects [24], [29]. 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 [25], [30]. The first aim of this study was to evaluate the ability of the test to identify hearing loss in a population of unscreened adults using an original multivariate approach, based on a set of features in addition to the SRT (e.g., average reaction time, age, test duration, number and percentage of correct responses).

Regarding a potential application of the test for screening in various settings, in a preliminary study we showed that the test provided consistent results (repeatable estimates of SRT and similar test-retest repeatability) in controlled laboratory settings and in uncontrolled environmental noise settings [29]. However, a deeper analysis of the possible influence of the hardware and the environment is necessary. In fact, individuals performing the test on different devices either locally or 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 input-output 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). The second aim of this study was to characterize the ear-level actual sound pressure levels of the test across a range of consumer transducers, including commercially available headphones and earphones, to identify possible minimum requirements for transducers to be used in the test.

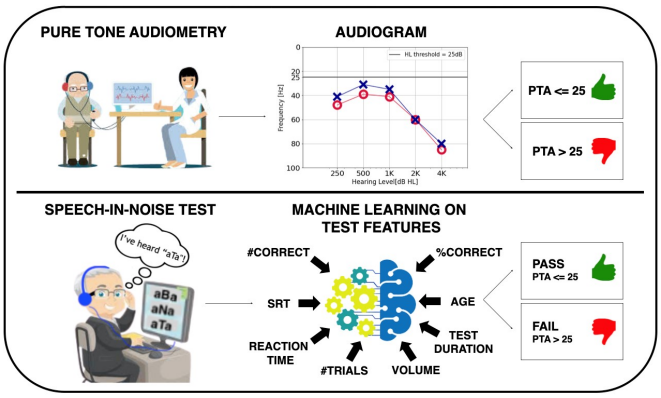

Fig. 1. (color online, b/w in print) Outline of the experiment to evaluate classification performance. Top panel: Pure-tone audiometry was performed at 0.5, 1, 2, and 4 kHz and the pure-tone average (PTA) was computed. Ears were classified in two classes: PTA ≤ 25 dB HL (no hearing loss) and PTA > 25 dB HL (hearing loss). Bottom panel: The speech-in-noise test was executed in a self-administered way and 8 features were extracted. Ears were classified by a machine learning approach into 'pass' and 'fail' using the PTA class as the target variable.

II. Materials and Methods

A. Experiment Setup and Speech-in-Noise Testing

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, at ear-level, the same cut-off value for PTA used by the WHO to define slight/mild hearing loss based on thresholds measured the better ear, as per the WHO definitions in force until Feb 28, 2021 [31], i.e.: PTA ≤ 25 dB HL (no hearing loss) and PTA > 25 dB HL (hearing loss).

Testing was performed in uncontrolled environmental noise settings in the context of opportunistic health screening initiatives (i.e. at community-based organizations for later life learning, health prevention and awareness events for the public) to reflect the potential target group of typical screening initiatives. The participant was seated comfortably in a dedicated room, with the experimenter supervising nearby.

The test is based on a user-operated speech-in-noise recognition task delivered in a three-alternatives multiple-choice format via a 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 used in some of the most commonly spoken languages worldwide (i.e., English, Spanish, French, Portuguese, German, and Italian) [26]. The test is based on a novel one-up/three-down staircase [27]-[28] that, in place of fixed, predetermined upward and downward step sizes (e.g., ±2 dB SNR) uses upward and downward steps that are adaptively determined at each trial based on the estimated psychometric curves of each stimulus in a way that the ratio between the downward and upward step approximates the optimal value of 0.7393 as suggested by [32].

Before taking the speech-in-noise test, participants were given the option to adjust the volume to a comfortable level using a graphical user interface. The default volume was set at 50% of the maximum allowed by the device and the chosen volume was saved as a percent value. To ensure appropriate implementation of the adaptive tracking algorithm, participants were not given the option to adapt the volume during 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.

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: 118 subjects; English: 10 subjects; Arabic: 6 subjects; Spanish: 4 subjects; French, Somali: 2 subjects; Albanian, Filipino, German, Moroccan Arabic, Igbo, and Efik: 1 subject). Of the 148 subjects, 98 (66%) had no hearing loss (average PTA = 11.69 dB HL) and 50 had hearing loss (average PTA = 38.25 dB HL), of which 34 (23%) mild-to-moderate and 16 (11%) moderate. 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.

B. *Evaluation of Classification Performance*

A set of eight features were extracted from the test software: SRT, i.e., the average of the SNR midpoints of the last 4 ascending runs; #trials, i.e. the total number of stimuli delivered; #correct, i.e. the number of stimuli correctly identified; %correct, i.e. the ratio #correct/#trials; average reaction time, i.e., the average of individual response time throughout the test; test duration, i.e. the time from the first stimulus presentation to the response to the last stimulus; self-adjusted volume, i.e. the percent value of the volume with reference to the maximum device volumes; and age (in years). A machine learning approach was then used to classify ears into 'pass' and 'fail' considering the PTA class (no hearing loss vs hearing loss) as the target variable.

To evaluate the classification performance of the test, a logistic regression algorithm was used in this study following a preliminary evaluation of different techniques. Specifically, we compared 7 supervised machine learning approaches (i.e., logistic regression, decision tree, support vector machines, k-nearest neighbors, ensemble logistic regression, random forests, and gradient boosting) using 5-fold cross-validation on the same dataset here used and we observed the following average AUC values on the test set using the full set of eight features: 0.89, 0.74, 0.88, 0.85, 0.88, 0.87, and 0.87, respectively [33].

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 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, a form of nested cross-validation [34]. 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. In addition, given the small sample size, we measured the variability of classification performance to address aleatoric uncertainty, i.e. the uncertainty of the model due to the inherent randomness of the data it was trained on [35]. 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.

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

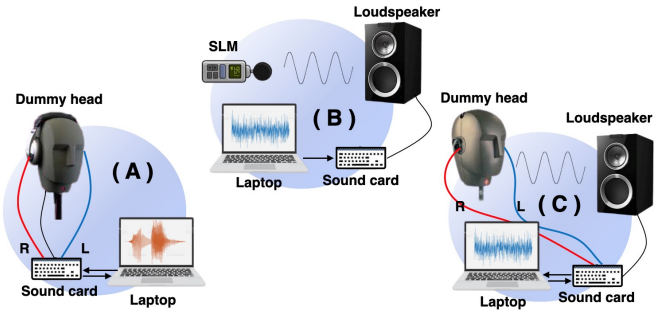

Fig. 2. (color online, b/w in print) Outline of the experiment to evaluate the sound levels of the test with different transducers. Panel (A): recording of the sequence of VCV stimuli via the dummy head using the different transducer models across the full range of volume output levels; Panel (B): calibration 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 white noise via the dummy head using the setup and output gain set in (B). R = right; L = left; SLM = sound level meter.

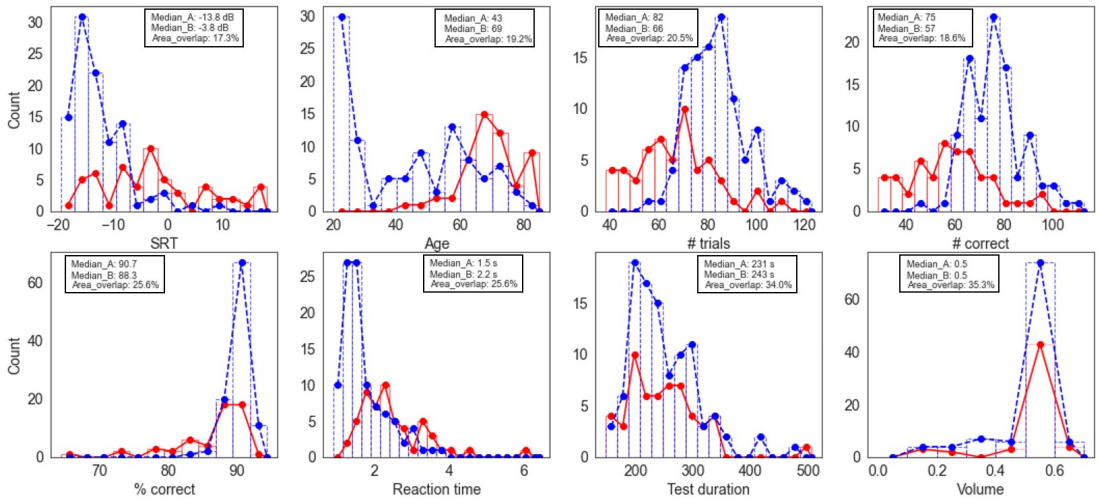

Fig. 3. (color online, b/w in print) Distributions (bar charts and frequency polygons) of the eight features in the two PTA classes: (A) PTA ≤ 25 dB HL (no hearing loss): dashed line; (B) PTA > 25 dB HL (hearing loss): continuous line. The text boxes show the median values in the two PTA classes and the percent overlap between the distributions.

The ear-level sound pressure levels of the test obtained with different consumer transducers across the full range of test volumes were measured in the lab using a dummy head. 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 with 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 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 SLM Software) at a distance of 1 meter from a loudspeaker. Finally, the SLM was removed and the dummy head was placed in the same position as the SLM and 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 hearing loss) are shown in Fig. 3. For each feature, the median values in the two classes as well as the percent overlap between the distributions of the two classes are also reported. The percent overlap was defined as the intersection area of the probability distributions estimated for the two classes. The distributions of SRT, age, #trials, #correct, %correct, and average reaction time were significantly different between the two classes (Wilcoxon rank sum test, $p << 0.001$) whereas the distributions of test duration and test volume were not significantly different between the two classes (Wilcoxon rank sum test: $p = 0.93$, $p = 0.06$, respectively).

The percent overlap ranged from about 17% (for SRT) to about 35% (for test volume) overall. The observed values of percent overlap confirmed that features such as test duration and volume had a similar distribution in the two classes (percent overlap equal to 34.0% and 35.3%, respectively). Features such as SRT, age, and #correct had more distinct distributions in the two classes, with percent overlap equal to 17.3%, 19.2%, and 18.6%, i.e., below 20%. Therefore, these features can be supposed to contribute more substantially to classification compared to features with higher percent overlap.

To assess the ability of features and features combinations to classify ears into the two PTA classes, the receiver operating characteristics were estimated using logistic regression on the whole dataset. The results are shown in Table I. The Table reports the values of AUC, sensitivity, and specificity obtained using each of the eight features separately, SRT and age (2 features), and SRT, age, and #correct (3 features), and all the 8 features. For the single features, the cut-off value for classification into the two classes is also shown. It is to note that the percent overlap observed with #trials was close to 20%, i.e.,

TABLE I
PERFORMANCE OF FEATURES AND FEATURES SETS
(AUC = AREA UNDER THE RECEIVER OPERATING CURVE; CUT-OFF = VALUE YIELDING THE REPORTED VALUES OF SENSITIVITY AND SPECIFICITY)

Input Feature(s)	AUC	Sensitivity	Specificity	cut-off
SRT	0.83	0.69	0.91	-7.48
Age	0.88	0.80	0.83	64
#trials	0.81	0.49	0.98	65
#correct	0.83	0.55	0.97	59
%correct	0.82	0.58	0.89	88.7
Reaction time	0.69	0.65	0.73	1.83
Test duration	0.50	0.31	0.83	200
Volume	0.58	0.40	0.65	0.51
2 features	0.90	0.67	0.89	-
3 features	0.91	0.82	0.84	-
8 features	0.93	0.80	0.90	-

TABLE II
CLASSIFICATION PERFORMANCE OVER 1000 ITERATIONS

Measure	8 features	3 features
Accuracy (training)	0.80 ± 0.02	0.81 ± 0.02
Accuracy	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

similar to the one observed with age and #correct. However, the Spearman's correlation coefficient between #trials and #correct was equal to 0.99 and therefore #trials was not included in the set of 3 features to limit multicollinearity. Vice versa, the correlation between the SRT, age, and #correct was lower (SRT-age: 0.69; SRT-#correct: -0.64; #correct-age: -0.53).

Table I shows that the classification models using 2, 3, and 8 features had higher AUC than each of the models using a single feature. Among the univariate models, the one with age had the highest performance, followed by SRT and #correct. Compared to the univariate classifiers, the multivariate models with 8 and 3 features reached higher AUC and, overall, both sensitivity and specificity were high (i.e., above 0.8).

B. Classification Performance

Table II shows the average and standard deviation of the observed performance measures (on training dataset: accuracy; on test dataset: accuracy, 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 i.e., the ones with lower percent overlap (Fig. 3) and higher classification performance (Table I): SRT, age, and #correct.

Overall, the average performance of the models with eight and three features was similar. Minor differences were observed, lower than 0.01, in terms of average AUC, sensitivity, specificity, and F1-score (Table II). The observed differences in performance measures between the two models were not statistically significant, except for the accuracy on the training test and AUC (t-test, $p < 0.001$ for both measures). 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: $p < 0.001$ in both models), 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 model with 8 and with 3 features, suggesting that the variability of the model performance 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 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'.

TABLE III
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

C. Sound Pressure Levels

Table III 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.

IV. DISCUSSION

The first aim of this study was to address the performance of a multivariate machine learning classifier to identify ears with hearing loss in a population of unscreened adults using a previously developed automated speech-in-noise test [24], [29]. Our dataset included 156 ears from 148 subjects. The distribution of ears in the two PTA classes was not balanced as 34% of ears had hearing loss. As the imbalance was relatively small, we did not apply dataset balancing techniques to support the future development of systems able to learn from real-world (i.e., imbalanced) input on a continuing basis. To limit the possible effect of class imbalance, in this study we assessed several performance measures in addition to accuracy as accuracy on imbalanced datasets may introduce some bias [36].

Using eight input features and a subset of three input features selected based on their ability to discriminate ears with and without hearing loss (Fig. 3, Table I), we observed similar average performance, as determined by running 1000 iterations of model optimization on different realizations of the training and test datasets (Table II). In addition, the variability of classification performance, as measured by the standard deviation of the performance measures across the 1000 iterations, was similar between the two models. Features such as SRT, age, and #correct had more distinct distributions in ears with and without hearing loss (Fig. 3) and better classification performance (Table I) 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 3 features model had the same accuracy (i.e., 0.79) and a slightly improved AUC compared to the 8 features model (0.90 vs 0.89).

Compared to our earlier investigations, where different classification algorithms were applied, the logistic regression model here used provided comparable or better classification performance. For example, when only the SRT was used to classify a subset of 106 ears from 98 subjects into pass and fail (cut-off SRT = -8 dB SNR, i.e., a cut-off value similar to the one here found), we observed an accuracy equal to 0.82, sensitivity equal to 0.70, specificity equal to 0.90, and AUC equal to 0.84 [24], in line with the results shown in Table I (first row). Similarly, when a decision tree algorithm with the full set of eight features was used on the same dataset here used, the average accuracy was 0.76, sensitivity 0.67, specificity 0.81, and the AUC was 0.74 [25]. The results of this study showed that logistic regression had better performance and that a subset of three features (SRT, age, and #correct) was appropriate for the sake of identifying ears with 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 subset of the data here used (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 [30]. 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 (Fig. 3, Tables I and II). It should be noted that in our sample age was the strongest predictor of hearing loss (Table I). In fact, participants with hearing loss were, on average, older than those with normal hearing (Fig. 3) as the experiment involved unscreened adults. Future studies, involving a higher number of individuals with hearing loss younger than 60 years old will help investigate further the role of each feature as a predictor of hearing loss.

Compared to the classification performance of other speech-in-noise tests based on multiple-choice recognition of short words, the logistic regression model using three features (SRT, age, and #correct) showed similar if not better performance in identifying ears with hearing loss. 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 the more distinctive features. However, further research is needed to demonstrate the viability of the proposed approach for adult hearing screening. Our sample was small and was not representative of the target population of our test. For example, our sample included a high proportion of female participants and was highly imbalanced with respect to spoken languages. It will be important to investigate test performance on a large, representative sample of participants, including subjects with varying degrees of hearing loss and across a larger set of native languages, also including tonal languages and languages that do not use the Roman alphabet, for example by using an online platform for wide-scale data collection. 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 mild-to-moderate hearing loss vs normal hearing, also addressing the revised cut-off used in the current WHO definition of normal hearing, introduced on March 2021 (i.e., PTA < 20 dB HL) [9], [37]. In the dataset used in this study, only 16 subjects 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. In addition, in view of a potentially broader application of the proposed approach as a tool for assessing benefit in individuals with hearing aids or cochlear implants, it would be interesting to investigate if the proposed system can be used as a measure of speech recognition performance in aided listening.

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 III). Considering the full range of

volumes shown in Table III, 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 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 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 [38]. 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 [39]-[40]. 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 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 better 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.

V. CONCLUSION

In summary, this study showed that the newly developed test combined with a multivariate classification algorithm may be a viable method for identifying hearing loss in varying settings and that the self-adjustment volume option may help compensate for the different transducers used. 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.

However, it is acknowledged that the results here reported do not fully demonstrate the viability of the test for screening at a distance as listening tests have been performed using only one headphone model. Real-world evaluation of the influence of transducer type and mode on the test results is necessary. More generally, it will be important to address test performance, including the possible influence of the user interface, device, transducers, 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, Geneva, Switzerland, vol. 295, p. 298, 2011. Available: http://whqlibdoc.who.int/publications/2011/9789240685215_eng.pdf
[11] 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.
[12] 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.
[13] 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.
[14] 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.
[15] A. Davis and P. Smith, "Adult hearing screening: health policy issues-what happens next?" *Am. J. Audiol.*, vol. 22, pp. 167–170, 2013.
[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] J.-M. Potgieter, D. W. Swanepoel, H. C. Myburgh, and C. Smits, "The South African English Smartphone Digits-in-Noise Hearing Test: Effect of Age, Hearing Loss, and Speaking Competence," Ear & Hearing, vol. 39, no. 4, pp. 656–663, Jul. 2018,
[24] 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.
[25] 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), Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 360, pp 3-19, 2021.
[26] 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.
[27] M. R. Leek, "Adaptive procedures in psychophysical research," *Percept Psychoph.*, vol. 63(8), pp. 1279–1292, 2001.
[28] 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.
[29] 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.
[30] 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, 2021.
[31] World Health Organization (WHO), Report of the informal working group on prevention of deafness and hearing impairment programme planning: Geneva, 18–21 June 1991. Available: https://apps.who.int/iris/handle/10665/58839
[32] M. A. García-Pérez, "Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties," Vision Res., vol. 38, no. 12, pp. 1861-81, Jun 1998.
[33] 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.
[34] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," PLoS ONE, vol. 14, no. 11, p. e0224365, Nov. 2019.
[35] R. Senge et al., "Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty," Information Sciences, vol. 255, pp. 16–29, Jan. 2014.
[36] H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009.
[37] L. E. Humes, "Examining the Validity of the World Health Organization's Long-Standing Hearing Impairment Grading System for Unaided Communication in Age-Related Hearing Loss," Am J Audiol, vol. 28, no. 3S, pp. 810–818, Oct. 2019.
[38] 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.
[39] Methods for Calculation of the Speech Intelligibility Index, ANSI Standard S3.5-1997.
[40] 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 as Graduate Research Fellow at the Italian National Research Council of Italy (CNR), Institute of Electronics, Information Engineering and Telecommunications (IEIIT), Milan, Italy. 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. She is currently working as Graduate Research Fellow at the Italian National Research Council of Italy (CNR), Institute of Electronics, Information Engineering and Telecommunications (IEIIT), Milan, Italy. Her research interests are related to machine learning methods and their application for the extraction of predictive and descriptive biomarkers in patients with chronic pathologies 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 in 2004. He worked for both Selex Communications and the Italian Telecommunications Consortium (CNIT) from 2001 to 2010. 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 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 interests 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.