Multivariate Classification of Mild and Moderate Hearing Loss Using a Speech-in-Noise Test for Hearing Screening at a Distance

Edoardo Maria Polo[1], Maximiliano Mollura[2], Riccardo Barbieri[2], and Alessia Paglialonga[3(✉)]

[1] DIAG, Università la Sapienza di Roma, 00185 Rome, Italy
[2] Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, 20133 Milan, Italy
[3] Cnr-Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni (CNR-IEIIT), 20133 Milan, Italy
alessia.paglialonga@ieiit.cnr.it

Abstract. In the area of smartphone-based hearing screening, the number of speech-in-noise tests available is growing rapidly. However, the available tests are typically based on a univariate classification approach, for example using the speech recognition threshold (SRT) or the number of correct responses. There is still lack of multivariate approaches to screen for hearing loss (HL). Moreover, all the screening methods developed so far do not assess the degree of HL, despite the potential importance of this information in terms of patient education and clinical follow-up. The aim of this study was to characterize multivariate approaches to identify mild and moderate HL using a recently developed, validated speech-in-noise test for hearing screening at a distance, namely the WHISPER (Widespread Hearing Impairment Screening and PrEvention of Risk) test. The WHISPER test is automated, minimally dependent on the listeners' native language, it is based on an optimized, efficient adaptive procedure, and it uses a multivariate approach. The results showed that age and SRT were the features with highest performance in identifying mild and moderate HL, respectively. Multivariate classifiers using all the WHISPER features achieved better performance than univariate classifiers, reaching an accuracy equal to 0.82 and 0.87 for mild and moderate HL, respectively. Overall, this study suggested that mild and moderate HL may be discriminated with high accuracy using a set of features extracted from the WHISPER test, laying the ground for the development of future self-administered speech-in-noise tests able to provide specific recommendations based on the degree of HL.

Keywords: Hearing loss · Hearing screening · Machine learning · Smartphone-based screening · Multivariate classifiers

S. Spinsante et al. (Eds.): HealthyIoT 2022, LNICST 456, pp. 81–92, 2023.
https://doi.org/10.1007/978-3-031-28663-6_7

1 Introduction

In the growing field of mobile health (mHealth), a number of smartphone hearing health apps have been developed for a variety of purposes, e.g., hearing screening, hearing aid management, patient education, and hearing rehabilitation [1–5]. Hearing screening is becoming increasingly popular as a means to increase awareness and identify the earlier signs of age-related hearing loss (HL), which would be typically left unnoticed otherwise [6, 7]. Among the validated apps introduced for adult hearing screening, some use pure-tone audiometry whereas others use speech-in-noise testing. The interest around speech-in-noise screening tests is growing as they can help detect real-life communication problems, for example difficulties in having conversations in noisy environments. Moreover, differently than pure tone audiometry, speech-in-noise tests are less sensitive to calibration procedures and can be performed in uncontrolled noise environments [8–11].

Recently, we have developed and validated a novel, automated speech-in-noise screening test viable for testing at a distance, e.g., through a web- or mobile-app, namely the WHISPER test (Widespread Hearing Impairment Screening and PrEvention of Risk) [12–16]. Differently than the majority of currently available speech-in-noise tests, the WHISPER test is minimally dependent on the listeners' native language, it is based on an optimized, efficient adaptive procedure, and it extracts a list of variables in addition to the speech recognition threshold (SRT), that is the most common variable used for speech-based screening [12–15, 17]. Multivariate approaches to HL identification such as the one used in the WHISPER test may help overcome the limitations of univariate approaches based on SRT only. In fact, individuals with normal hearing may have poor SRTs, whereas individuals with HL may be able to reach satisfactory speech recognition performance [18, 19]. Moreover, research has shown that features such as the subject's age or the average reaction time can help identify HL [13, 17, 18, 20, 21]. Nevertheless, multivariate approaches to HL identification and classification are not widely adopted yet.

In our previous studies, we have assessed the ability of multivariate approaches to identify HL of mild degree or higher, using both the former and the newer World Health Organization (WHO) definitions of HL (i.e., average value of pure-tone thresholds at 0.5, 1, 2, and 4 kHz (PTA) higher than 25 dB HL and higher than 20 dB HL, respectively [22, 23]). Specifically, in a preliminary sample of 148 participants (age = 52.1 ± 20.4 years; age range: 20–89 years; 46 males, 102 female), we showed that multivariate classifiers based on, for example, logistic regression (LR), support vector machines, k-nearest neighbors, or random forest were more accurate than univariate classifiers to identify HL of mild degree or higher, using the former WHO definition of HL [13, 15]. In the same sample of participants, we showed that LR was also able to accurately predict the self-perceived hearing handicap, as measured using the Hearing Handicap Inventory for the Elderly–Screening Version (HHIE-S) [17]. In a larger sample of 207 participants (age = 52 ± 20 years; age range: 20–89 years; 66 males, 141 female), we confirmed that multivariate classifiers could achieve high accuracy (up to 0.85 with RF) and we showed, using post-hoc explainability techniques, that he most important features for the identification of mild HL, using the newer WHO definition, were age, SRT, average reaction time, and percentage of correct responses [17].

In all the above studies, multivariate algorithms were characterized considering binary classification of two output classes, i.e., normal hearing vs HL (mild or higher). Whereas binary classification can be appropriate for general HL detection, nevertheless knowledge of the degree of HL (e.g., mild-to-moderate vs moderate) would be important, particularly for hearing screening delivered at a distance using unsupervised tests via web or smartphone. In fact, individuals with different degrees of HL should undergo different intervention strategies and should be provided with different follow-up information and educational content [24]. The aim of this study was to characterize, for the first time, multivariate approaches to identify mild and moderate HL (mild HL: 20 dB HL $<$ PTA $\leq$ 40 dB HL; moderate HL: PTA $>$ 40 dB HL) using the WHISPER test.

The article is organized as follows. Section 2 outlines the study participants and protocol and the data analysis approach used. Section 3 presents the results obtained in terms of univariate and multivariate feature characterization and classification performances for binary and multi-class classification (NH vs mild-to-moderate HL vs moderate HL). Section 4 discusses the obtained results in the context of the available literature. Finally, the conclusions of the study and the possible future developments are outlined in Sect. 5.

2 Methods

2.1 Participants and Procedure

The study sample included 350 participants (117 men, 223 women; age: mean 49 years, range: 18–89 years) tested during HL awareness events. The study dataset includes 442 records (92 participants tested in both ears, 258 in one ear).

Pure-tone audiometry was performed at 0.5, 1, 2, and 4 kHz (Amplaid 177+ by Amplifon, TDH49 headphones) and speech-in-noise testing using the WHISPER test. Testing was performed in a quiet room at hearing screening and awareness initiatives. The protocol was approved by the Politecnico di Milano Research Ethical Committee (Opinion No. 2/2019, Feb 19, 2019; renewed by Opinion No. 13/2022, Apr 13, 2022).

The WHISPER test is delivered on a touch-screen interface and is based on an adaptive procedure. Specifically, a sequence of meaningless vowel–consonant–vowel (VCV) syllables (e.g., *ata* and *asa*) are presented at varying signal-to-noise ratio (SNR) in a three-alternative multiple-choice paradigm. Further details on the WHISPER test are reported in [12, 15, 21]. The following features were extracted from the WHISPER test: SRT, number of correct responses (#correct), percentage of correct responses (%correct), average reaction time, and test duration.

2.2 Data Analysis

The ears tested were classified in three classes, following the WHO definitions of mild and moderate HL [22, 23]. Specifically, the following three classes were defined: (i) normal hearing (NH): PTA $\leq$ 20 dB HL; 299 ears (~68%); (ii) mild HL: 20 dB HL $<$ PTA $\leq$ 40 dB HL; 97 ears (~22%); and (iii) moderate HL: PTA $>$ 40 dB HL; 46 ears (~10%). Six input features were considered for classification, i.e., the five features extracted from the WHISPER test and the subject's age.

Univariate and Multivariate Characterization of Features. The Receiver Operating Characteristics (ROC) for binary classification (i.e., mild HL vs NH; and moderate HL vs NH) were computed for each of the six input features and for LR on two combinations of features, i.e.: (i) the full set of six features and (ii) a subset of features with AUC $\geq$ 0.80 for *both* mild HL vs NH and moderate HL vs NH classification. The LR algorithm was used following results from [15, 17].

The Shapiro-Wilk test was performed to check for normality of the distributions of the six input features in the three output classes. As the distributions were not normal, possible differences in median values between the three classes were assessed using the Kruskal-Wallis test with Bonferroni correction. A significance level $\alpha = 0.05$ was considered.

Binary and Multiclass Classification Performance. Classification performance was assessed by training a LR algorithm for binary classification (mild HL vs NL, moderate HL vs NH) and multi-class classification (NH vs mild HL vs moderate HL). The data set was randomly split into a training set including 80% of the sample (353 records) and a test set including the remaining 20% (89 records). Stratification was applied to maintain the same percentage of records in the two classes of the original data set in the training and test partitions. Class weights were applied to the data to compute LR coefficients to limit the effect of class imbalance, particularly for the moderate HL class. Data were standardized to zero mean and unit variance. Due to the relatively small size of the data set, 5-fold cross-validation was introduced on the training set to partially reduce the influence of the selected partition on the trained model.

3 Results

Figure 1 shows the distributions of the six input features in the three output classes (NH, mild HL, and moderate HL). Age, SRT, and average reaction time tended to increase with increasing degree of HL. All the observed differences in median values of age, SRT, and average reaction time were statistically significant, except for the age difference between mild and moderate HL. The features #correct and %correct tended to decrease with increasing degree of HL. All the observed differences in median values of #correct and %correct were statistically significant. The test duration tended to increase from NH to mild HL, but not from NH or mild HL to moderate HL.

Figure 2 shows the ROC estimated using, for each HL class, the feature with highest performance (age and SRT for mild and moderate HL, respectively) and using LR on (i) the full set of six features and (ii) a subset of features with AUC $\geq$ 0.80, i.e. age, SRT, and average reaction time. The univariate and multivariate performance of each feature and feature combinations for mild HL vs NH classification and for moderate HL vs NH classification is shown in Table 1 and Table 2, respectively.

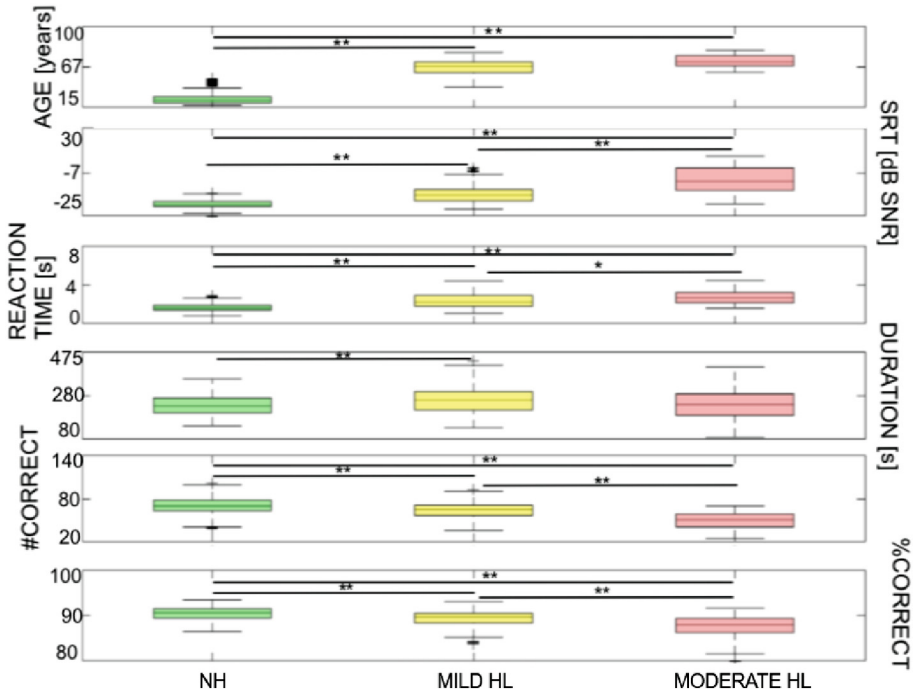

Fig. 1. Distribution of features in the three output classes: normal hearing, mild HL, and moderate HL. Statistically significant differences in median values between the classes are marked with * ($p < 0.05$) and ** ($p < 0.01$).

The feature with the highest performance for mild HL identification was age, whereas the one with highest performance for moderate HL identification was SRT (accuracy = 0.86 at the optimal cut-off value). For moderate HL identification, the performance of age was lower than that of SRT but still relatively high (accuracy = 0.82). The optimal cut-off values for age, SRT, average reaction time, and test duration increased from mild to moderate HL, whereas those for %correct and #correct decreased with increasing degree of HL, in line with the trends shown in Fig. 1. Using LR on combinations of three or six features did not lead to improved performance for mild HL identification, as shown in Table 1. For moderate HL identification, LR on three and on six features achieved improved performance (accuracy up to 0.90). In general, the highest performance for *both* mild HL and moderate HL identification was observed using LR on the full set of six features.

Table 3 shows the observed performance of LR for binary and multiclass classification performance, as measured in the training set (average ± s.d. from 5-fold cross validation) and in the test set. The observed accuracies were higher than 0.81 for binary classification and equal to 0.72 for multiclass classification, with no remarkable differences in performance between the average performance on the training set and the estimated performance on the test set, suggesting no overfitting effects. For binary classification, both sensitivity and specificity were high, indicating very good performance.

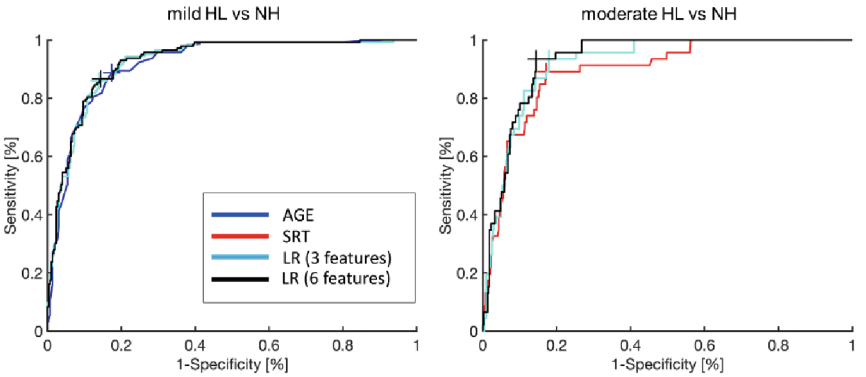

Fig. 2. ROC for binary classification (left-hand panel: mild HL vs NH; right-hand panel: moderate HL vs NH). The three ROC shown represent: (i) the feature with the highest classification performance, i.e. age for mild HL (dark blue) and SRT for moderate HL (red); (ii) LR of age, SRT, and average reaction time (light blue); and (iii) LR of all the input features (black).

Table 1. Univariate and multivariate performance for mild HL at the optimal cut-off value.

Feature	Sensitivity	Specificity	F1-score	Cut-off	AUC
Age	0.89	0.83	0.78	59 years	0.92
SRT	0.78	0.76	0.68	−12.14 dB SNR	0.85
Avg reaction time	0.74	0.75	0.64	1.9 s	0.80
Duration	0.57	0.61	0.48	252 s	0.58
%Correct	0.64	0.70	0.43	0.90	0.75
#Correct	0.67	0.66	0.44	66	0.74
LR (3 features)	0.86	0.86	0.79	–	0.92
LR (6 features)	0.87	0.86	0.80	–	0.93

The sensitivity in multiclass classification was lower compared to binary classification, in line with the higher number of classes. Nevertheless, multiclass classification performance was still good as sensitivity was around 0.70 and specificity was around 0.85. The lower values of sensitivity, specificity, and accuracy measured in the test set shown in Table 3 compared to those measured at the optimal cut-off value using the ROC (Table 1, Table 2) are related to the use of machine learning, as opposed to simple ROC analysis, and to the relatively small size of the dataset that leads to the observed variability in performance due to the underlying uncertainty in data. This variability is demonstrated by the observed standard deviation of the accuracy on the training set across 5-fold cross validation (s.d. up to $\pm$ 0.05). The higher values of F1-score observed for moderate HL vs NH classification compared to those shown in Table 2 may be related to the use of class weights that may have partially compensated the effect of class imbalance on F1-score estimates.

Table 2. Univariate and multivariate performance for moderate HL at the optimal cut-off value.

Feature	Sensitivity	Specificity	F1-score	Cut-off	AUC
Age	0.83	0.81	0.48	67 years	0.89
SRT	0.89	0.82	0.52	−7.48 dB SNR	0.89
Avg reaction time	0.83	0.72	0.38	2.20 s	0.83
Duration	0.41	0.70	0.19	274 s	0.49
%Correct	0.80	0.74	0.23	0.89	0.80
#Correct	0.82	0.78	0.24	60	0.88
LR (3 features)	0.93	0.82	0.52	–	0.92
LR (6 features)	0.93	0.86	0.57	–	0.93

Table 3. Binary and multiclass classification performance using LR.

	Accuracy (training)	Accuracy (test)	F1-score (training)	F1-score (test)	Sens (test)	Spec (test)
Mild HL/NH	0.84 ± 0.05	0.81	0.83 ± 0.06	0.79	0.83	0.80
Moderate HL/NH	0.85 ± 0.02	0.84	0.72 ± 0.03	0.72	0.89	0.84
Moderate HL/mild HL/NH	0.74 ± 0.04	0.72	0.63 ± 0.05	0.64	0.69	0.85

4 Discussion

The availability of methods for accurate identification of the degree of HL (i.e., mild vs moderate) following hearing screening via unsupervised tests delivered through web- or mobile- platforms would be important for tailoring clinical assessment and patient education. Nevertheless, current univariate approaches typically target mild HL. Also, there is still lack of multivariate approaches able to discriminate the degree of HL using a speech-in-noise screening test. In this study, we characterized the univariate and multivariate performance of a set of six features extracted from the WHISPER speech-in-noise screening test for the sake of identifying mild and moderate HL in unscreened adults.

Results in Fig. 1, Table 1, and Table 2 indicated that the univariate classification performance of the six features extracted from the WHISPER platform varied with varying degree of hearing loss. Specifically, the features with higher performance (i.e., AUC $\geq$ 0.80) for mild HL identification were age, SRT, and average reaction time. The features with higher performance for moderate HL identification were age, SRT, #Correct, average reaction time, and %correct. The highest accuracy at the optimal ROC point was observed using age and a cut-off value equal to 59 years for mild HL and

using SRT and a cut-off value equal to −7.48 dB SNR for moderate HL. Age and SRT were the features with higher performance for both mild and moderate HL (AUC ≥ 0.85), followed by average reaction time (AUC ≥ 0.80). The cut-off value for age, SRT, and average reaction time increased with increasing degree of HL. The feature with the lowest classification performance was test duration (AUC = 0.58 and 0.49 for mild and moderate HL, respectively).

The relationship between SRT, pure-tone thresholds, and age is well known. Age-related deficits in auditory and cognitive processing may play a role when speech is presented in background noise such as in the proposed screening test [18, 25, 26]. As shown in our earlier study, the interaction between age and PTA can accurately predict SRT [12], in line with the fact that the ability to properly recognize speech is the result of complex relationships between age, degree of HL, and cognitive abilities [27, 28]. The relevance of the average reaction time was also highlighted in previous studies in relation to mild HL detection [13, 17] and it is confirmed here for both mild and moderate HL classification. Regarding test duration, the univariate classification abilities were, in general, poor, with negligible differences in the distributions of test duration across the three classes. This may be interpreted in light of a compensation effect related to the adaptive nature of the WHISPER procedure. In fact, individuals with increasing degree of HL have in general poorer speech recognition abilities and worse cognitive abilities and, as such, they tend to exhibit longer reaction times when responding to a given stimulus in the trial. However, individuals with poorer speech recognition performance tend to go through a lower number of trials in the adaptive procedure as the staircase reaches convergence earlier if there is a high number of incorrect responses [12, 15, 17].

Multivariate characterization of features indicated that the classification performance obtained using LR on age, SRT, and average reaction time (i.e., the three features with AUC ≥ 0.80 for both mild and moderate HL) and the one obtained using LR on the full set of six features led to increased performance compared to the best univariate feature. LR on the six features yielded the highest performance for both mild and moderate HL classification. These results suggest that a multivariate approach may be more accurate than the best-performing univariate ones in discriminating different degrees of HL from the speech-in-noise test here used. The accuracy obtained by training a ML classifier using the six features was 0.82 and 0.87 for mild and moderate HL, respectively, suggesting high classification performance (Table 3).

The observed multivariate classification performance was equal to or higher than that observed in previous studies or with other speech-in-noise tests. For example, identification of mild HL using the SRT estimated from English digits-in-noise test yielded an accuracy equal to 0.82 [29]. In our previous study, using data from a smaller sample of 207 participants, we observed an accuracy equal to 0.86 for mild HL using the full set of six features [17]. The slightly lower accuracy observed in the current study may be related to differences in the underlying data and classification approach. Specifically, in [17] records with mild and moderate HL were aggregated in a single HL class, the output classes NH and HL were balanced (54% vs 46%), and age was more strongly correlated with HL. In a recent study, multiclass classification performance of the digits-in-noise test was assessed using a univariate approach based on the estimated SRT in a large sample of 3422 participants from the Rotterdam study. The observed accuracy at the

optimal ROC point was 0.72 for mild HL (42% of the sample) and 0.95 for moderate HL (12% of the sample) [30]. Another study assessed self-conducted SRT measured using the German matrix sentence test in home settings against two criteria for HL, i.e. (i) the earlier WHO criterion for mild HL (i.e., PTA > 25 dB HL), and (ii) the German criterion for hearing aid indication (i.e., pure-tone threshold > 30 dB in one or more frequencies between 500 Hz and 4 kHz), that is similar to a moderate HL criterion [31]. The study showed that the accuracy for criterion (i) was 0.74 whereas that of criterion (ii) was 0.76, i.e., lower than the performance here observed with our multivariate approach.

The study here shown has some limitations. First, the distribution of age and degree of HL in our sample may not reflect that of the general population. For example, in our sample we observed a prevalence of HL equal to 32% that is higher than the reported prevalence of hearing loss in adults, i.e., about 20% [32]. This sampling bias may be related to the experiment settings whereby data were collected primarily within the context of hearing screening and awareness initiatives for the general public. For similar reasons, the sample may have been biased towards higher age than that of the general population. It will be important in future studies to limit the sampling bias and assess the univariate and multivariate classification performance in a larger sample including a higher proportion of individuals with NH and a higher proportion of middle aged and young adults. In addition, our multivariate approach was based on a set of only six features extracted from the WHISPER test. It will be interesting to investigate further features, for example those related to psychometric functions estimated from the adaptive procedure, or individual performance in subsets of stimuli (e.g., high-frequency vs low-frequency stimuli), or more complex measures of reaction time. Inclusion of a cognitive testing module into the WHISPER platform could also help address in more detail the relationships between hearing sensitivity, speech recognition, reaction time, and aging. Last, but not least, in this study we focused on the WHISPER test only. It will be important to investigate univariate and multivariate classification performance towards mild and moderate HL using different automated speech-in-noise tests that may be delivered via web or smartphone.

5 Conclusions

In this study we assessed, for the first time, the ability of univariate and multivariate classifiers to identify mild and moderate HL in unscreened adults using a recently validated speech-in-noise test, the WHISPER test. The results showed that the features with highest performance in identifying HL were different between mild and moderate HL. Moreover, results showed that the performance of multivariate classifiers using the full set of available features was better than that of the best-performing univariate classifiers, reaching an accuracy equal to 0.82 and 0.87 for mild and moderate HL, respectively. The results of this study are encouraging and suggest that mild and moderate HL may be discriminated using a small set of features extracted from an automated speech-in-noise screening test, laying the ground for the development of future self-administered speech-in-noise tests viable for screening hearing and cognitive function at a distance and potentially able to provide specific recommendations based on the degree of HL. Access to a mobile application that in a few minutes can give a stratified indication on the

degree of HL, considering not only the SRT but a broader picture of the subject, could lead indeed to important benefits to individuals at risk of HL, who can quickly assess their hearing, with improved accuracy as multivariate approaches can help overcome limitations due to well-known mismatch between PTA and SRT in adults.

Acknowledgements. This research work was partially supported by Capita Foundation (project WHISPER, Widespread Hearing Impairment Screening and PrEvention of Risk, 2020 Auditory Research Grant). The authors would like to thank the Lions Clubs International and Associazione La Rotonda, Baranzate (MI) for their contribution in the organization and management of hearing screening and awareness initiatives. The authors are also grateful to all the students who contributed to data collection and to Marta Lenatti and Marco Zanet who contributed to the development of the WHISPER platform.

References

1. Swanepoel, D.W.: eHealth technologies enable more accessible hearing care. In: Seminars in Hearing, vol. 41, no. 02, pp. 133–140 (2020)
2. Paglialonga, A.: eHealth and mHealth for audiologic rehabilitation. In: Montano, J.J., Spitzer, J.B. (eds.) Adult Audiologic Rehabilitation, 3rd edn. Plural Publishing, San Diego (2020)
3. Paglialonga, A., Cleveland Nielsen, A., Ingo, E., Barr, C., Laplante-Lévesque, A.: eHealth and the hearing aid adult patient journey: a state-of-the-art review. BioMed. Eng. Online **17**, 101 (2018). https://doi.org/10.1186/s12938-018-0531-3
4. Paglialonga, A., Tognola, G., Pinciroli, F.: Apps for hearing science and care. Am. J. Audiol. **24**(3), 293–298 (2015). https://doi.org/10.1044/2015_AJA-14-0093
5. Bright, T., Pallawela, D.: Validated smartphone-based apps for ear and hearing assessments: a review. JMIR Rehabil. Assist. Tech. **3**(2), e13 (2016)
6. Davis, A., Smith, P., Ferguson, M., Stephens, D., Gianopoulos, I.: Acceptability, benefit and costs of early screening for hearing disability: a study of potential screening tests and models. Health Technol. Assess. **11**(42), 1–294 (2007)
7. Feltner, C., Wallace, I.F., Kistler, C.E., Coker-Schwimmer, M., Jonas, D.E.: Screening for hearing loss in older adults. JAMA **325**(12), 1202 (2021)
8. Leensen, M.C., de Laat, J.A., Dreschler, W.A.: Speech-in-noise screening tests by Internet, Part 1: test evaluation for noise-induced hearing loss identification. Int. J. Audiol. **50**(11), 823–834 (2011)
9. Smits, C., Kapteyn, T.S., Houtgast, T.: Development and validation of an automatic speech-in-noise screening test by telephone. Int. J. Audiol. **43**(1), 15–28 (2004)
10. Paglialonga, A., Grandori, F., Tognola, G.: Using the speech understanding in noise (SUN) test for adult hearing screening. Am. J. Audiol. **22**(1), 171–174 (2013). https://doi.org/10.1044/1059-0889(2012/12-0055)
11. Paglialonga, A., Tognola, G., Grandori, F.: A user operated test of suprathreshold acuity in noise for adult hearing screening: the SUN (speech understanding in noise) test. Comput. Biol. Med. **52**, 66–72 (2014). https://doi.org/10.1016/j.compbiomed.2014.06.012
12. Paglialonga, A., Polo, E.M., Zanet, M., Rocco, G., van Waterschoot, T., Barbieri, R.: An automated speech-in-noise test for remote testing: development and preliminary evaluation. Am. J. Audiol. **29**(3S), 564–576 (2020). https://doi.org/10.1044/2020_AJA-19-00071

13. Polo, E.M., Zanet, M., Lenatti, M., van Waterschoot, T., Barbieri, R., Paglialonga, A.: Development and evaluation of a novel method for adult hearing screening: towards a dedicated smartphone app. In: Goleva, R., Garcia, N.R.d.C., Pires, I.M. (eds) HealthyIoT 2020. LNICST, vol. 360, pp. 3–19. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-69963-5_1
14. Zanet, M., Polo, E.M., Rocco, G., Paglialonga, A., Barbieri, R.: Development and preliminary evaluation of a novel adaptive staircase procedure for automated speech-in-noise testing. In: Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2019). https://doi.org/10.1109/EMBC.2019.8857492
15. Zanet, M., et al.: Evaluation of a novel speech-in-noise test for hearing screening: classification performance and transducers characteristics. IEEE J. Biomed. Health Inf. **25**(12), 4300–4307 (2021). https://doi.org/10.1109/JBHI.2021.3100368
16. Paglialonga, A., et al.: WHISPER (widespread hearing impairment screening and prevention of risk): a new platform for early identification of hearing impairment and cognitive decline. In: Hearing Across the Lifespan Conference (HEAL), Cernobbio, Italy, 16–18 June 2022
17. Lenatti, M., Moreno-Sánchez, P.A., Polo, E.M., Mollura, M., Barbieri, R., Paglialonga, A.: Evaluation of machine learning algorithms and explainability techniques to detect hearing loss from a speech-in-noise screening test. Am. J. Audiol. **31**(3S), 961–979 (2022). https://doi.org/10.1044/2022_AJA-21-00194
18. Humes, L.E.: Understanding the speech-understanding problems of older adults. Am. J. Audiol. **22**(2), 303–305 (2013)
19. Killion, M.C., Niquette, P.A.: What can the pure-tone audiogram tell us about a patient's SNR loss? Hear. J. **53**(3), 46–48 (2000)
20. Nuesse, T., Steenken, R., Neher, T., Holube, I.: Exploring the link between cognitive abilities and speech recognition in the elderly under different listening conditions. Front. Psychol. **9**, 678 (2018)
21. Polo, E.M., Zanet, M., Paglialonga, A., Barbieri, R.: Preliminary evaluation of a novel language independent speech-in-noise test for adult hearing screening. In: Jarm, T., Cvetkoska, A., Mahnič-Kalamiza, S., Miklavcic, D. (eds.) EMBEC 2020. IFMBE Proceedings, vol. 80, pp. 976–983. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-64610-3_109
22. World Health Organization, Deafness and hearing loss. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. Accessed 24 Sept 2022
23. World Health Organization, World report on hearing. https://www.who.int/publications/i/item/world-report-on-hearing. Accessed 24 Sept 2022
24. US Preventive Services Task Force. Screening for Hearing Loss in Older Adults: US Preventive Services Task Force Recommendation Statement. JAMA **325**(12), 1196–1201 (2021)
25. Zurek, P., Delhorne, L.A.:Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment. J. Acoust. Soc. Am. **82**(5), 1548–1559 (1987)
26. Summers, V., Makashay, M.J., Theodoroff, S.M., Leek, M.R.: Suprathreshold auditory processing and speech perception in noise: hearing-impaired and normal-hearing listeners. J. Am. Acad. Audiol. **24**(4), 274–292 (2010)
27. Füllgrabe, C., Moore, B.C.J., Stone, M.A.: Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition. Front. Aging Neurosci. **6**, 347 (2015)
28. Smith, S.L., Pichora-Fuller, M.K.: Associations between speech understanding and auditory and visual tests of erbal working memory: effects of linguistic complexity, task, age, and hearing loss. Front. Psychol. **6**, 1394 (2015)
29. Watson, C.S., Kidd, G.R., Miller, J.D., Smits, C., Humes, L.E.: Telephone screening tests for functionally impaired hearing: current use in seven countries and development of a U.S. version. J. Am. Acad. Audiol. **23**(10), 757–767 (2012)

30. Armstrong, N.M., Oosterloo, B.C., Croll, P.H., Ikram, M.A., Goedegebure, A.: Discrimination of degrees of auditory performance from the digits-in-noise test based on hearing status, Int. J. Audiol. **59**(12), 897–904 (2020)
31. Ooster, J., Krueger, M., Bach, J.-H., Wagener, K.C., Kollmeier, B., Meyer, B.T.: Speech audiometry at home: automated listening tests via smart speakers with normal-hearing and hearing-impaired listeners. Trends Hear **24** (2020)
32. Stevens, G., Flaxman, S., Brunskill, E., Mascarenhas, M., Mathers, C.D., Finucane, M.: Global burden of disease hearing loss expert group: global and regional hearing impairment prevalence: an analysis of 42 studies in 29 countries. Eur. J. Pub. Health **23**(1), 146–152 (2013)