




An adaptive protocol to assess physiological responses as a function of task demand in speech-in-noise testing

Edoardo Maria Polo ^a, Davide Simeone ^{a,b} , Maximiliano Mollura ^a , Alessia Paglialonga ^b ,^{1,*},
Riccardo Barbieri ^a ¹

^a Politecnico di Milano, Piazza Leonardo da Vinci, 32, Milan, 20133, Italy

^b Cnr-Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni (CNR-IEIIT), Piazza Leonardo da Vinci, 32, Milan, 20133, Italy

ARTICLE INFO

Keywords:

Listening effort
Physiological signals
Speech communication
Speech-in-noise testing

ABSTRACT

Background: Acoustic challenges impose demands on cognitive resources, known as listening effort (LE), which can substantially influence speech perception and communication. Standardized assessment protocols for monitoring LE are lacking, hindering the development of adaptive hearing assistive technology.

New Method: We employed an adaptive protocol, including a speech-in-noise test and personalized definition of task demand, to assess LE and its physiological correlates. Features extracted from electroencephalogram, galvanic skin response, electrocardiogram, respiration, pupil dilation, and blood volume pulse responses were analyzed as a function of task demand in 21 healthy participants with normal hearing.

Results: Heightened sympathetic response was observed with higher task demand, evidenced by increased heart rate, blood pressure, and breath amplitude. Blood volume amplitude and breath amplitude exhibited higher sensitivity to changes in task demand.

Comparison with Existing Methods: Notably, galvanic skin response showed higher amplitude during low task demand phases, indicating increased attention and engagement, aligning with findings from electroencephalogram signals and Lacey's attention theory.

Conclusions: The analysis of a range of physiological signals, spanning cardiovascular, central, and autonomic domains, demonstrated effectiveness in comprehensively examining LE. Future research should explore additional levels and manipulations of task demand, as well as the influence of individual motivation and hearing sensitivity, to further validate these outcomes and enhance the development of adaptive hearing assistive technology.

1. Introduction

Everyday's life activities may pose listening challenges when the conditions are acoustically adverse (e.g., noisy environments, multiple talkers) or when the hearing ability of the listener is decreased due to elevated pure-tone thresholds or decline in suprathreshold auditory processing abilities (Pichora-Fuller et al., 2016). Listening in challenging conditions requires the allocation of executive cognitive resources that may vary as a function of the acoustic, linguistic, and cognitive demand of the task, as well as with varying listeners' abilities (Peelle, 2018). The recruitment of cognitive resources required in challenging listening tasks is called listening effort (LE) (Pichora-Fuller et al., 2016). To date, there are no standardized protocols to assess the individual LE

required during real-world listening tasks. Since a wide range of factors may affect LE measured in experimental settings, especially in terms of speech stimuli (e.g., vowel consonant vowel (VCV), disyllabic words, digits, sentences) and experimental protocols (e.g., noise levels, type of signals recorded, type of task), the investigation of LE requires the observation of multiple and different domains. Moreover, the relationship between LE and listener's performance is complex. In LE conditions, a semantic message can be understood with appropriate recruitment of cognitive resources, but a constant increased cognitive load can lead to stress and fatigue, influencing individual performance (Hétu et al., 1988; Pichora-Fuller et al., 2016)

* Correspondence to: Alessia Paglialonga, Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, Consiglio Nazionale delle Ricerche (CNR-IEIIT), c/o DEIB (Ed. 21), Politecnico di Milano, Piazza Leonardo da Vinci, 32, I-20133 Milano, Italy.

E-mail addresses: edoardomaria.polo@polimi.it (E.M. Polo), davidesimeone@cnr.it (D. Simeone), maximiliano.mollura@polimi.it (M. Mollura), alessia.paglialonga@cnr.it (A. Paglialonga), riccardo.barbieri@polimi.it (R. Barbieri).

¹ Alessia Paglialonga and Riccardo Barbieri contributed equally to this work.

LE can be quantified using self-report, behavioral and physiological measures. The latter are particularly promising to monitor LE in real-world settings and are investigated in this paper. Since the mental effort is usually expressed as autonomic nervous system (ANS) activation, some physiological signals such as skin conductance, pupil diameter and heart-related features are commonly used to measure effort investment during listening. Also, some neuroanatomical markers associated to LE have been addressed in literature. Higher activation of the prefrontal cortex, premotor cortex and cingulo opercular network was observed with degraded speech (Peelle, 2018). For example, in the study reported in Seeman and Sims (2015) sentences were presented at four different fixed signal-to-noise ratio (SNR) levels, resulting in smaller heart-rate variability for smaller SNRs and for increased task complexity (diotic vs dichotic listening) during a key word recognition speech-in-noise test (SNT), while galvanic skin response (GSR) was elevated for greater task complexity only. In Plain et al. (2020), it was found that cardiac pre-ejection period (PEP) reactivity varied linearly with poorer SNRs during a speech-in-noise task including short Dutch sentences, which was completed at six fixed SNRs between -1 and -21 dB SNR, distributed in 4 dB steps. In McMahon et al. (2016), sentences spoken by a native Australian-English female were presented in the presence of four-talker babble noise at 15 different levels between -7 and $+7$ dB SNR, during two different sessions that used 16-channel (highly intelligible) an 6-channel (moderately intelligible) vocoded material, respectively. Pupil size was measured while alpha band activity from three parietal electrodes was extracted from electroencephalographic (EEG) data, and both measures significantly decreased with higher SNRs for the 16-channel vocoding, while this relationship was not observed for the moderately intelligible condition.

While most studies analyzed the response of physiological signals at pre-determined SNR levels, a tailored assessment of physiological responses in LE conditions would require the adoption of an adaptive approach to assess how LE changes depending on the individual auditory performance. For example, in Petersen et al. (2015) two, four or six monosyllabic Swedish digits spoken by a female talker were presented at -4 , 0 and $+4$ dB with respect to the speech reception threshold of each subject, previously found through an adaptive tracking procedure targeting 80% intelligibility. In the least intelligible condition, average alpha power across 31 centro-parietal electrodes resulted to increase with hearing loss in the low and intermediate memory load conditions, while for the high memory load it dropped for the mild and moderate hearing loss subjects. In Zekveld et al. (2013), everyday Dutch sentences were presented in background-interfering speech at 29% and 71% speech recognition thresholds (SRTs), after two adaptive procedures were applied to find the SRTs. Peak and mean pupil dilation compared to baseline were higher in the 29% condition than in the 71% condition, indicating higher processing load. While the overall evidence summarized above suggests that physiological signals may help assess LE associated with speech recognition tasks, inconsistent relationships between and within subjective, behavioral and physiological measures of LE during auditory tasks at different demands were reported in literature (e.g., Strand et al., 2018; Alhanbali et al., 2019). This phenomenon was explained in previous studies with the idea of a “multi-dimensional” model of LE (Shields et al., 2023), which indicates that different measures (both between and within groups) may capture different aspects of LE, spatially and temporally. First, the degree of agreement between measures and their sensitivity to detect LE are dependent from the experimental conditions (Shields et al., 2023). Second, different physiological measures can capture different aspects of LE and different processing stages, such as attending to, processing, or adapting to auditory stimuli. The abovementioned temporal and spatial stratification of LE measures would therefore explain the absence of strong and consistent correlations between the examined measures and the lack of an optimal measure of LE. These assumptions would require to move towards more ecological studies outside the laboratory and try different approaches from the ones commonly adopted.

The aim of this study is to investigate multi-domain physiological indicators linked to task demand during an adaptive SNT, carefully designed to reduce task complexity, cognitive load, and testing time (Zanet et al., 2021; Paglialonga et al., 2020). The rationale of the adaptive approach here used is to personalize the SNR of stimuli around the individual SRT and reduce the likelihood of trials becoming overly challenging over time, thus minimizing the risk of disengagement that may alter the physiological responses (Pichora-Fuller et al., 2016), and at the same time enabling the assessment of task demand as a function of the individual SRT. The preliminary findings of this research, as outlined in Polo et al. (2022), pertain solely to the cardiovascular aspect. In this study, to comprehensively assess physiological indicators of LE, six physiological signals are recorded, overall covering the cardiovascular, central, and autonomic domains. The primary research questions are: (1) Is it possible to observe changes in physiological responses as a function of task demand during an adaptive speech-in-noise test? and (2) How do different physiological indicators—across cardiovascular, central, and autonomic domains—correlate with task demand during the adaptive SNT?

2. Materials and methods

2.1. Speech-in-noise test

The adaptive SNT used in this study was recently developed and validated on a population of 417 individuals with varying degrees of hearing sensitivity (Paglialonga et al., 2020, 2023; Polo et al., 2023). The test employs a corpus of 12 nonsense VCV stimuli containing spoken consonants (/b, d, f, g, k, l, m, n, p, r, s, t/) in the context of the vowel /a/ (e.g., aba, ada) recorded from a male professional native English speaker. A three-alternative forced-choice (3AFC) task is employed, with response options defined using a maximal opposition criterion to maximize perceptual distinctions in manner, voicing, and place of articulation (Paglialonga et al., 2013, 2014; Vaez et al., 2014). The rationale behind this specific test design was to reduce reliance on higher-level cognitive processing, mitigate the impact of factors like subjects' educational background, literacy, or native language on test results, and limit possible state anxiety related to task execution (Mattys et al., 2009; Cooke et al., 2010; Roup et al., 2020). VCVs were presented in filtered speech-shaped noise (onset: 500 ms; offset: 100 ms). The noise was computed by filtering a Gaussian white noise by using the international long-term average speech spectrum (Byrne et al., 1994) and a low-pass filter with a cut-off frequency of 1.4 kHz and a roll-off slope of 100 dB/octave. This noise was then attenuated by 15 dB to create a noise floor as suggested in Leensen et al. (2011).

The SNT uses a one-up/three-down (1U3D) adaptive procedure to maximize effectiveness, precision, and convergence to the SRT target at 79.4% intelligibility (Schlauch and Rose, 1990; Shelton and Scarrow, 1984; Leek, 2001). The test consists of a single block, and at each trial one VCV from the whole corpus is randomly picked and presented, while the order of the three alternatives displayed on the screen is randomly determined. The test employs a recently validated, optimized staircase procedure with varying upward and downward steps in SNR computed from the predicted intelligibility of each VCV in the set (Zanet et al., 2019; Paglialonga et al., 2020; Zanet et al., 2021). This approach uses the Short-Time Objective Intelligibility (STOI) values to predict VCV intelligibility (Taal et al., 2010) - a computational metric highly correlated with, but distinct from, actual speech intelligibility (Rocco et al., 2023). During each trial:

1. The adaptive algorithm selects a target STOI value (for the first trial, a STOI close to 100% is used)
2. A VCV stimulus is randomly chosen and presented at the SNR level corresponding to the current STOI value on its STOI-derived psychometric function

3. Following the 1U3D rule:

- After each incorrect response, the STOI value increases
- After three consecutive correct responses, the STOI value decreases

The obtained STOI value is used as the target STOI value defined in (1). The ratio between downward and upward steps in STOI values is set in a way that the ratio between downward and upward steps in SNR is close to 0.7393 as recommended by [García-Pérez \(1998\)](#). This design is particularly efficient because, while the STOI value governs the adaptive procedure, the actual SNR at which each stimulus is presented may differ due to their individual psychometric functions. The test continues until 12 reversals are achieved, with SRT calculated as the mean SNR of the stimuli presented during the last four ascending runs of STOI values, as recommended by [García-Pérez \(1998\)](#).

Participants had the option to fine-tune the stimulus volume at a comfortable level during an initial training phase before starting the test ([Zanet et al., 2021](#); [Paglialonga et al., 2020](#); [Zanet et al., 2019](#)). The features extracted from the SNT include the SRT, the correctness of the responses given at each trial and the reaction time for each trial, measured as the time elapsed between the onset of the stimulus and the subject's click on one of the three alternative buttons displayed on the screen.

2.2. Experimental protocol

Participants were 21 healthy young adults (13 female, 8 male; mean age = 26.2 ± 1.47 years) with normal hearing (pure-tone average thresholds across 500, 1000, 2000, and 4000 Hz < 20 dB HL). To mitigate the potential influence of caffeine and nicotine on physiological signals, participants were instructed to abstain from drinking coffee and from smoking for a minimum of two hours before the experiment. Sensors for acquiring electrocardiographic (ECG), blood volume pulse (BVP), GSR, pupil dilation (PUPIL), EEG, and respiration (RESP) signals were affixed to the subjects. Initially, a two-minute monitoring period was conducted while the subjects were instructed to fix their gaze on a gray screen to establish baseline values for the physiological signals. Following the baseline measurement, the SNT described in Section 2.1 was administered and signals were monitored throughout the test execution.

In the adaptive staircase procedure, trials were categorized as either low or high demand based on their SNR relative to the subject's SRT. Trials with an SNR higher than the SRT + 2 dB were labeled as low demand, while trials with an SNR equal to or lower than SRT + 2 dB were labeled as high demand. The cut-off SNR of SRT + 2 dB was chosen to balance two objectives: (i) Ensuring a sufficient number of high demand trials for the analysis, and (ii) Keeping the difference between the cut-off SNR and the SRT below 3 dB SNR, which is the average difference in SNR that leads to measurable changes in speech intelligibility ([McShefferty et al., 2015](#)). To ensure the stability and comparability of physiological signal analysis between the two task demand levels, we focused on time segments that contained consecutive trials of the same difficulty level (i.e., low (L) or high (H) demand). We considered all segments with consecutive trials at the same difficulty level and identified the longest segment for each level of demand. Then, we extracted the longest available segment at each of the two difficulty levels and, by considering the lower duration between the two, we truncated the longer segment to match the duration of the shorter one, ensuring equal duration for the selected time segments at the two task demand levels. In both demand phases, we prioritized time segments from the later part of the phase. This choice was made to minimize the influence of the initial familiarization period, as the staircase procedure begins at an easy level for the subject. Most consecutive low demand trials occur at the beginning of the test, during the initial descent before the SNR settles near the SRT. By selecting the last consecutive low demand trials before the subject's performance reached the cut-off SNR of SRT + 2 dB, we aimed to capture the most representative and stable

low demand time segments for analysis. Moreover, we have likewise chosen a duration equivalent to that of the test phases (i.e., L and H task demand) for the baseline, enabling a comparison across the three components. During the experiment, two minutes of baseline data were initially acquired before the SNT. The baseline window was defined as the segment of time immediately preceding the appearance of the test screen and the duration of the baseline segment was chosen to match the duration of the L and H phases in each subject. [Fig. 1](#) illustrates an example of an adaptive procedure from a single participant and highlights the identified low and high task demand windows. The solid gray line depicts the target STOI value, whereas the red line represents the SNR in dB used at each trial, as reported in [Rocco et al. \(2023\)](#). Regarding the task demand conditions, in the example four consecutive windows for low (i.e., 1st L and 2nd L) and high task demand (i.e., 1st H and 2nd H) are identified, and two windows of equal duration are selected for the analysis (i.e., the final part of the 1st L window and the 2nd H window).

2.3. Apparatus

Pure tone audiometry was measured using a clinical audiometer (Amplaid 177+, Amplifon with TDH49 headphones). The SNT was administered through a desktop computer using UXD CT887 headphones and participants responded using a mouse. The ECG, BVP, GSR, and RESP signals were acquired using the Procomp Infinity device. EEG data were collected using a DSI 24 headset equipped with 19 dry electrodes positioned according to the international 10–20 system. The headset featured a 300 Hz sampling rate and integrated a 16-bit analog-to-digital (A/D) converter for precise signal conversion. For PUPIL data acquisition, a Tobii Pro X2 Compact eye-tracker, operating at a sampling frequency of 60 Hz, was employed. In relation to visual aspects, particularly the pupillary signal, we followed established guidelines ([Laeng and Endestad, 2012](#)). To minimize external light interference, laboratory windows were darkened, and uniform artificial lighting was maintained throughout the recording sessions. The screen brightness was consistently maintained at 3/4 of the maximum brightness for all subjects. The experiment was conducted at the SpinLab of Politecnico di Milano, with all participants providing informed consent prior to participation. The protocol was approved by the Politecnico di Milano Research Ethics Committee (Opinion No. 29/2021).

2.4. Signal processing and feature extraction

The signal processing and feature extraction for each physiological signal are reported here below.

2.4.1. ECG

The ECG signal, sampled at 2048 Hz, was pre-processed using a fourth-order zero-phase low-pass Butterworth filter and down-sampled to 250 Hz. The ECG signal was analyzed to extract HRV features, which provide information about the ANS activity. The R peaks, representing the depolarization of the ventricles and the most prominent feature of the ECG signal, were detected using the Pan-Tompkins algorithm ([Sedghamiz, 2014](#)). The RR interval, defined as the time between two consecutive R peaks, was then calculated. Any errors in R peak detection were manually corrected using an in-house software. To ensure precise HRV measurements, we adopted the Point process framework, which is especially suited for short time intervals such as the task demand phases here defined, that are substantially shorter than the conventional 5-minute windows typically used for HRV analysis, as suggested by [Sassi et al. \(2015\)](#). The RR series served as input for the Point process framework, which models heartbeats as a stochastic point process. This approach allows continuous estimation of the average inter-beat interval and the associated spectral indices, providing real-time assessments of various HRV parameters ([Chen et al., 2009, 2010](#); [Barbieri et al., 2005](#)).

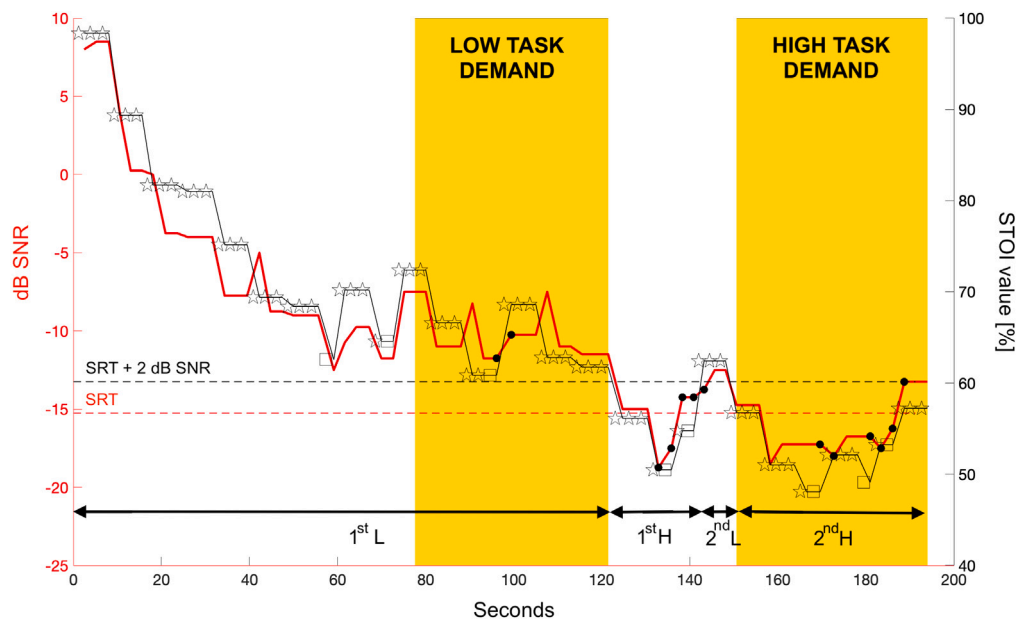


Fig. 1. Example of adaptive trial results and corresponding low and high task demand windows (in yellow) from one participant. The gray continuous line represents STOI (Short-Time Objective Intelligibility) values. The continuous red line tracks the SNR values, with stars indicating correct responses and squares indicating errors. The solid markers denote the values of SNRs in the last 4 ascending runs of the black continuous line. The SRT (horizontal dashed red line) is computed as the average of the SNRs of the solid markers. The figure also highlights four consecutive trial windows for low (1st L and 2nd L) and high task demand (1st H and 2nd H). (color online; b/w in print).

From the modeled RR signal, the following features were extracted, sampled at 10 Hz, and subsequently averaged within the temporal windows of interest:

- μ_{RR} : Average of RR interval.
- σ_{RR}^2 : Variance of RR interval.
- RRTOT: Total power of the continuous RR interval spectrum.
- Very Low (<0.04 Hz), Low (0.04–0.15 Hz), and High (0.15–0.4 Hz) frequency power of the continuous RR interval spectrum (RRVLF, RRLF, RRHF), the normalized values RRLF_n and RRHF_n, computed relative to total power, excluding very low-frequency power, and the ratio RRLF/RRHF (RRLF_{toRRHF}).

These metrics were selected due to their capacity to provide comprehensive insights into ANS activity. Mean RR (μ_{RR}) and variance σ_{RR}^2 reflect overall HRV, while LF and HF powers are indicative of sympathetic and parasympathetic influences, respectively. The LF/HF ratio serves as a measure of sympathovagal balance (Sassi et al., 2015).

2.4.2. EEG

EEG signals were preprocessed using the EEGLAB toolbox in MATLAB (Delorme and Makeig, 2004). This involved filtering the data within the 1 Hz to 45 Hz range using a finite impulse response, zero-phase filter, and applying a notch filter at 50 Hz to eliminate power line interference. Data from the problematic Pz electrode were temporarily excluded and later interpolated. Subsequently, Independent Component Analysis (ICA) with the Extended Infomax algorithm was employed (Delorme et al., 2012).

The quality of the extracted components was thoroughly evaluated using the ICLABEL plugin (Pion-Tonachini et al., 2019), and any components identified as artifacts were removed based on default threshold values. The Common-Average Referencing (CAR) method was employed to reduce common noise in the recorded signals. The selection of this EEG signal processing pipeline was based on comparisons detailed in Cassani et al. (2022).

In this study, straightforward and easily interpretable features were extracted:

- The Power Spectral Density (PSD) in various frequency bands (i.e., α , β and θ) for the frontal (F) and parietal (P) regions were computed and normalized by the total power spectral density (1–45 Hz).
- The attention index, computed as the ratio between the PSD in the β frequency band and the PSD in the θ frequency band for the frontal and parietal regions. This ratio tends to increase during attentive states, providing a measure of attention throughout the trials (Cómez et al., 1998; Farabbi and Mainardi, 2022).
- The engagement index, the ratio between the PSD in the β frequency band and the PSD in the α frequency band for the frontal and parietal regions. This ratio is useful to investigate whether the change in task demand impacted not only attention but also the degree of involvement (Coelli et al., 2017).

The EEG features selected, such as PSD in alpha, beta, and theta bands, as well as attention and engagement indices, are widely used to assess cognitive states during various tasks (Klimesch, 1999; Berka et al., 2007). By focusing on frontal and parietal regions, we aimed to capture the most relevant EEG correlates of LE (Clayton et al., 2015).

2.4.3. GSR

To isolate the Phasic Component of the GSR signal (sampled at 256 Hz), we applied a fourth order low-pass Butterworth filter with a cutoff frequency of 2 Hz. The signal was then downsampled to 5 Hz. We employed a median filter within a 4-second window around each sample, following the methods outlined in Bakker et al. (2011) and Greco et al. (2016). This process yielded a median signal, which we subtracted from the filtered signal to obtain the Phasic Component. GSR peaks, indicating spikes in eccrine gland activity, were identified by locating local maxima in the filtered signal between the onset (amplitude > 0.01 μ S) and offset (amplitude < 0.01 μ S) of the Phasic Component, as reported in Braithwaite et al. (2013). The following features were extracted from the GSR signal as suggested by Picard et al. (2001), Kim and Andre (2008), Fleureau et al. (2012), Lisetti and Nasoz (2004) and Frantzidis et al. (2010):

- Avg amplitude peaks: Average amplitude over identified peaks and their standard deviation (Sd amplitude peaks).

- Avg rise time: Average time between onsets and peaks.
- Avg recovery time: Average time between peaks and offsets.
- N peaks: Number of identified peaks in the time window.
- Mean of the low-pass filtered GSR, considering both tonic and phasic components (Avg GSR) and its standard deviation (Sd GSR).
- Env: Mean of the envelope of the phasic component.

The GSR features, including amplitude, rise time, recovery time, and number of peaks, as well as the mean and standard deviation of the filtered GSR signal, are commonly used to assess sympathetic nervous system activity and emotional arousal (Boucsein, 2012; Critchley, 2002). They were selected to investigate how sympathetic arousal varies with task demand during the SNT. An increase in the amplitude and number of GSR peaks suggests a higher level of sympathetic activation and emotional arousal. Shorter rise times and recovery times indicate a more rapid and intense physiological response to stimuli, which is often associated with increased stress and cognitive workload. Additionally, an increase in the mean and standard deviation of the filtered GSR signal reflects an overall increase in skin conductance level and its variability, which can be indicative of heightened sympathetic arousal and emotional reactivity.

2.4.4. BVP

The BVP signal (sampled at 2048 Hz) was filtered using a fourth-order low-pass Butterworth filter with a 25 Hz cutoff frequency and downsampled to 250 Hz. By synchronizing the BVP signal with the ECG signal, we successfully extracted systolic, diastolic, and onset amplitudes. The systolic and diastolic values were discerned from the peaks and troughs between R-peaks, while the onset values were identified at the inflection points. This analysis yielded two key features:

- Mean Volume Amplitude Index (VA)
- Mean Pulse Arrival Time (PAT), computed with respect to onsets relative to diastoles or systoles for enhanced reliability in turbulent signal segments.

VA and PAT derived from the BVP signal are indicators of peripheral blood flow and vascular tone, respectively. These parameters have been correlated with various psychophysiological states, including emotional arousal and mental stress (Parreira et al., 2023; Chakraborty et al., 2024). An increase in VA suggests enhanced peripheral blood volume, implying reduced peripheral blood pressure, which may indicate elevated parasympathetic activity and low arousal. Conversely, a decrease in PAT signifies increased pulse wave velocity, often associated with heightened vascular tone and sympathetic activation.

2.4.5. RESP

The respiration signal, sampled at 256 Hz, was processed using the Parks–McClellan algorithm (Rabiner et al., 1978) by applying a zero-phase digital low-pass filter with a 1 Hz cutoff frequency to isolate the desired frequency components in the signal. After filtering, similar to the univariate approach for the RR series of the ECG, we employed a bivariate autoregressive point process model to estimate the autonomic regulation of heartbeat influenced by respiratory changes (Chen et al., 2009). This model separated the self-regulatory process from the effects of Respiratory Sinus Arrhythmia (RSA) on the Autonomic Nervous System's (ANS) feedback branch. The modeling enabled us to create time and frequency representations of the RR and RESP series, along with their corresponding cross-spectrum. Additionally, we calculated the Coherence in Time and Frequency (COH(t,f)) between the RR and RESP series, serving as an indicator of the robustness of the coupling between these two time series.

The following features were computed from RESP signals through the bivariate point process:

- μ_{RESP} : Average amplitude of the modeled resp series.

- σ^2_{RESP} : Variance amplitude of the resp series.
- $RESP_{HF}$: High-frequency (0.15–0.4 Hz) power in the continuous resp spectrum.
- RSA_{HF} : RSA gain in the high-frequency range.
- COH_{HF} : Coherence of the two time series calculated in the high-frequency range.
- $f_{max_{HF}}$: RESP frequency computed at the point of maximum coherence in the high-frequency band.

These features reflect the depth and variability of breathing, the strength of RSA, and the degree of cardiorespiratory coupling. RSA, in particular, has been extensively studied as an index of vagal tone and has been shown to be sensitive to various types of stressors and mental effort (Houtveen et al., 2002).

2.4.6. PUPIL

In the initial data analysis, we identified blinks based on sample diameters (<2 mm or >8 mm) and artifacts (abrupt changes > 0.375 mm within 20 ms intervals) in accordance with Partala and Surakka (2003) and Pong and Fuchs (2000). We estimated missing data and/or blinks using cubic spline interpolation. To eliminate high-frequency noise, a fourth-order zero-phase low-pass Butterworth anti-aliasing filter with a 5 Hz cutoff frequency was applied, preserving relevant signal information. Subsequently, we downsampled the signal to 10 Hz, an appropriate rate for pupillometric data analysis. Spectral analysis was conducted by computing Welch's periodogram on the detrended signal using a 1.875-second Hamming window with a 50% overlap. After processing the signals from each eye, we averaged the samples to obtain a signal representing the mean diameter of both eyes, which was then used to compute the following features:

- Mean diameter (AVD) and its standard deviation (SDD).
- PSD of diameter in low (0.05–0.15 Hz) (DLF), high (0.15–0.45 Hz) (DHF), and very high (0.45–1.5 Hz) (DVHF) frequency ranges, along with the balance index (DLF/DHF).

Pupillometric features, including mean diameter (AVD), its standard deviation (SDD), and PSD in different frequency ranges, are widely used as indicators of cognitive workload, attention, and arousal (Beatty, 1982; Zekveld et al., 2018; Duchowski et al., 2018). These features were selected to characterize the pupillary response to changes in task demand and gain insights into cognitive workload and arousal associated with LE. An increase in pupil diameter is generally associated with increased cognitive workload, attentional allocation, and arousal. The standard deviation of the pupil diameter reflects the variability in pupil size, which can be indicative of changes in cognitive processing and fluctuations in arousal. The PSD of the pupil diameter signal in different frequency ranges provides information about the temporal dynamics of the pupillary response.

2.5. Statistical analysis

The SNT performance variables, i.e., average reaction time and percentage of correct responses, were analyzed within each task demand window. The average reaction time was computed as the mean value of the reaction time of all trials in each task demand window. The Shapiro–Wilk test was employed to assess the normality of data distributions for the low and high task demand phases. The Wilcoxon signed-rank test was then applied since the distributions were not normal.

Concerning the physiological features, as they needed to accommodate not only the two high and low task demand windows but also the baseline, the Shapiro–Wilk test was used to assess normality of data. For normal distributions, an ANOVA test was employed, while for not normal distributions, the Friedman's test was performed. Specifically, if a comparison yielded statistical significance, multiple comparisons were conducted to evaluate possible pairwise differences. Tukey's correction

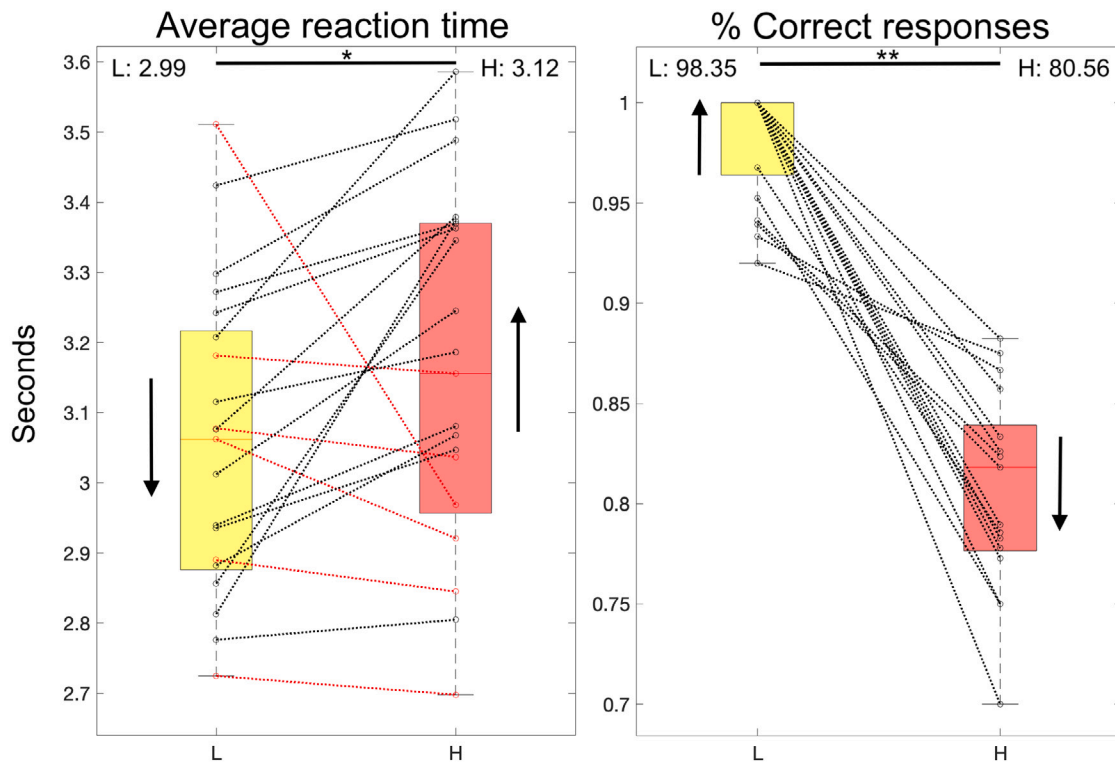


Fig. 2. Boxplots of the average reaction time and percentage of correct responses in the low (L) and high (H) task demand phases. In each plot, the average values for the L and H windows are also reported. Dotted lines connect data points from individual subjects. The up and down arrows next to the boxplots indicate whether the trend for each feature is increasing or decreasing in either of the task demands. Expected trends highlighted in black and opposite trends highlighted in red for the minority of subjects. Statistically significant differences are marked with * ($p < 0.05$) and ** ($p < 0.01$) (color online; b/w in print).

was applied to all comparisons.

Furthermore, to provide an intuitive visualization of how physiological variables change across different task demand levels, 3D boxes were generated using the most informative features in terms of physiological patterns, by considering both the features showing statistically significant differences and features extracted from the three domains here considered, cardiovascular, central, and autonomic. 3D boxes are constructed with the feature median as the center of each box, where the box's sides represent the median absolute deviation.

3. Results

Fig. 2 displays the distributions of the SNT features (i.e., average reaction time and percentage of correct responses) observed in the low and high task demand (mean window duration: 64.41 ± 17.49 s) windows. In the low task demand phase, a lower average reaction time and a higher percentage of correct responses are observed compared to the high task demand window. All the observed differences were statistically significant (average reaction time: $p = 0.011$; percentage of correct responses: $p = 5.88e-5$). These trends are consistent with an increased task difficulty in the high task demand window (lower SNR) that is associated with longer reaction times (in 15/21 subjects) and lower speech recognition performance, as reflected by a lower percentage of correct responses (in 21/21 subjects). **Table A.1** in the **Appendix** provides detailed information for each subject, including the number of trials, mean SNR, mean STOI values for the low and high task demand windows as well as individual SRT.

Table 1 shows the distributions (median values and median absolute deviations) of the features extracted from all the physiological signals at the baseline (B) and in the low (L) and high (H) task demand phases.

Fig. 3 shows the boxplots of the six features that exhibit statistically significant differences between the low and high task demand phases, as reported in the right-most column in **Table 1** (i.e., μ_{RR} , Env, PAT,

Avg GSR, VA and μ_{RESP}), along with the trends observed for each subject. The results observed for each physiological signal are summarized here below:

- **ECG:** By assessing the cardiac characteristics obtained through point process, we observed that the mean RR interval (μ_{RR}) emerges as a significant metric as it effectively distinguishes between the low and high task demand phases (**Fig. 3, Table 1**). Notably, in the high task demand phase an increase in heart rate (lower μ_{RR}) is observed, even when compared with the initial baseline phase, although statistical significance is not reached. Similar trends are observed for σ_{RR}^2 , although the observed differences are not statistically significant. Nonetheless, this describes a more consistent and stable pattern of RR interval distribution during phases of increased task demand, as indicated by the narrower range depicted in the boxplots in **Fig. 3**. Furthermore, the RR spectral power in the whole band (RRTOT) and in the very low frequency range (RRVLF) demonstrate a statistically significant decrease in the high demand phase compared to the baseline (**Table 1**). No statistically significant differences were observed in the other ECG spectral features.
- **EEG:** As shown in **Table 1**, the spectral power of the more relevant frequency bands (i.e., α , β , θ) demonstrates a pronounced increase during the two task demand phases compared to the baseline, within both parietal (P) and frontal (F) regions. The attention index β/θ displays a significant increase during the low task demand phase compared to the baseline in both regions, whereas it significantly increases in the high task demand phase compared to the baseline only in the parietal region. Furthermore, the measure of cognitive engagement β/α experiences a substantial increase in both low and high task demand phases compared to the baseline (**Table 1**). However, no statistically significant differences in EEG features between the two task demand phases are observed.

Table 1

Median and median absolute deviation of the features extracted from physiological signals in the Baseline (B), low (L), and high (H) task demand windows. Statistically significant differences are reported in the last column. Further details on the statistical analysis are reported in [Table A.2 in the Appendix](#).

	B	L	H	*
ECG				
μ_{RR} [s] 10^{-2}	84.60 (9.44)	83.60 (8.81)	81.80 (8.53)	L-H
σ_{RR}^2 [s ²] $\times 10^{-3}$	0.88 (0.79)	0.80 (0.63)	0.62 (0.61)	-
RRTOT [s ²] $\times 10^{-2}$	0.46 (0.63)	0.36 (0.19)	0.21 (0.17)	B-H
RRVLF [s ²] $\times 10^{-2}$	0.13 (0.56)	0.15 (0.14)	0.07 (0.09)	B-H
RRLF [s ²] $\times 10^{-3}$	0.90 (1.40)	0.62 (0.97)	0.79 (0.78)	-
RRHF [s ²] $\times 10^{-3}$	0.41 (0.60)	0.42 (0.32)	0.31 (0.37)	-
RRLF _n	0.63 (0.15)	0.57 (0.12)	0.54 (0.17)	-
RRHF _n	0.37 (0.15)	0.43 (0.12)	0.34 (0.13)	-
RRLFtoHF	2.45 (3.31)	1.71 (1.62)	2.67 (2.06)	-
EEG				
PSD α F $\times 10^{-2}$	2.07 (2.35)	4.97 (3.21)	4.25 (2.81)	B-L,B-H
PSD β F $\times 10^{-2}$	2.03 (3.45)	14.07 (7.92)	7.55 (7.84)	B-L,B-H
PSD θ F $\times 10^{-2}$	7.80 (3.32)	11.40 (2.84)	11.04 (3.57)	B-L
PSD α P $\times 10^{-2}$	2.48 (3.27)	6.41 (4.28)	4.84 (4.09)	B-L,B-H
PSD β P $\times 10^{-2}$	2.48 (4.39)	11.98 (7.54)	10.99 (7.53)	B-L,B-H
PSD θ P $\times 10^{-2}$	8.89 (3.55)	11.97 (3.21)	11.18 (3.72)	-
β/θ F	0.11 (0.74)	1.01 (0.81)	0.97 (1.51)	B-L
β/θ P	0.20 (0.48)	0.94 (0.60)	0.50 (0.69)	B-L,B-H
β/α F	0.59 (1.79)	1.75 (1.24)	1.56 (2.70)	B-L,B-H
β/α P	0.56 (0.88)	1.56 (0.56)	1.41 (0.86)	B-L,B-H
GSR				
Avg amplitude peaks [μ S] $\times 10^{-2}$	1.73 (4.85)	2.51 (5.41)	0.36 (5.69)	-
Sd Amplitude peaks [μ S] $\times 10^{-2}$	1.50 (5.83)	2.21 (6.79)	0 (8.61)	-
Avg rise time [s]	0.82 (0.57)	0.82 (0.44)	0.63 (0.59)	-
Avg recovery time [s]	2.38 (3.13)	3.17 (5.27)	1.95 (2.43)	-
N peaks	2 (2.60)	2 (3.30)	1 (2.77)	-
Avg GSR [μ S]	1.81 (2.13)	2.12 (2.22)	1.40 (2.32)	B-L,L-H
Env [μ S] $\times 10^{-2}$	1.88 (3.78)	2.27 (4.25)	0.67 (4.97)	L-H
BVP				
VA [a.u.]	4.88 (1.89)	5.82 (2.11)	3.99 (1.73)	B-H,L-H
PAT [s] $\times 10^{-2}$	29.96 (1.93)	30.21 (1.95)	29.13 (1.93)	L-H
RESP				
μ_{RESP} [a.u.]	31.07 (3.25)	31.30 (3.39)	31.39 (3.54)	B-H,L-H
RSA_{HF} [a.u./ms]	1.85 (2.60)	4.25 (3.03)	2.78 (2.03)	-
$RESP_{HF}$ [a.u.] $\times 10^{-3}$	0.35 (1.65)	0.48 (12.20)	0.45 (0.93)	-
COH_{HF}	0.55 (0.11)	0.60 (0.11)	0.68 (0.10)	-
$fmax_{HF}$ [Hz]	0.31 (0.05)	0.30 (0.04)	0.32 (0.04)	-
PUPIL				
AVD [mm]	2.94 (0.36)	2.99 (0.30)	3.04 (0.26)	-
SDD [mm]	0.22 (0.10)	0.17 (0.03)	0.17 (0.07)	B-L,B-H
DLF [mm ²]	0.92 (3.33)	0.41 (0.23)	0.28 (0.10)	B-L,B-H
DHF [mm ²]	0.78 (1.25)	0.55 (0.37)	0.50 (0.38)	-
DVHF [mm ²]	0.49 (0.50)	0.28 (0.18)	0.29 (0.34)	B-L
DLFtoHF	0.98 (1.20)	0.63 (0.48)	0.58 (0.29)	B-L,B-H

- GSR: For most of the features extracted from the GSR signal, no statistically significant differences are observed across the three phases. A noteworthy exception lies in the mean value of the signal (Avg GSR) within the three time windows. Specifically, this value is markedly higher during the low task demand phase compared to the other two phases, and lower in the high task demand phase compared to the baseline. A similar pattern is observed in the Env feature, portraying the envelope of the phasic component of the signal. In general, more conventional GSR features, such as the number and amplitude of peaks (e.g., N peaks and Avg amplitude peaks) show trends that are aligned with those of Avg GSR and Env, collectively indicating an increased GSR during the low task demand phase compared to the high task demand phase.
- BVP: Notably, the two features (i.e., VA and PAT) extracted from the BVP signal show statistically significant differences between the low and high task demand phases. Specifically, lower values of VA and PAT during the high task demand phase are observed compared to the baseline and to the low task demand phase. VA

is especially noteworthy, as the majority of subjects (18 out of 21) exhibit a consistent trend between the low and high task demand phases, as illustrated in [Fig. 3](#).

- RESP: The average breath amplitude (μ RESP) exhibits a significant increase during the high task demand phase. Specifically, this feature, following VA and Avg GSR (3 out of 21 subjects each), stands out as having one of the smallest numbers of subjects (4 out of 21) displaying a trend opposite to the average pattern. This emphasizes its ability to discriminate physiological responses across different task demand phases. Specifically, at high task demand, subjects tend to engage in more intense and more closely spaced breaths. However, the observed inter-individual changes are very small ([Fig. 3](#)). The remaining RESP features (i.e., $RESP_{HF}$, COH_{HF} and $fmax_{HF}$) do not demonstrate remarkable differences across all three phases, except for RSA_{HF} . This parameter, representing an estimate of the vagal tone (i.e., respiratory sinus arrhythmia), displays a higher median value during the low and high task demand phases, although the observed differences are not statistically significant.

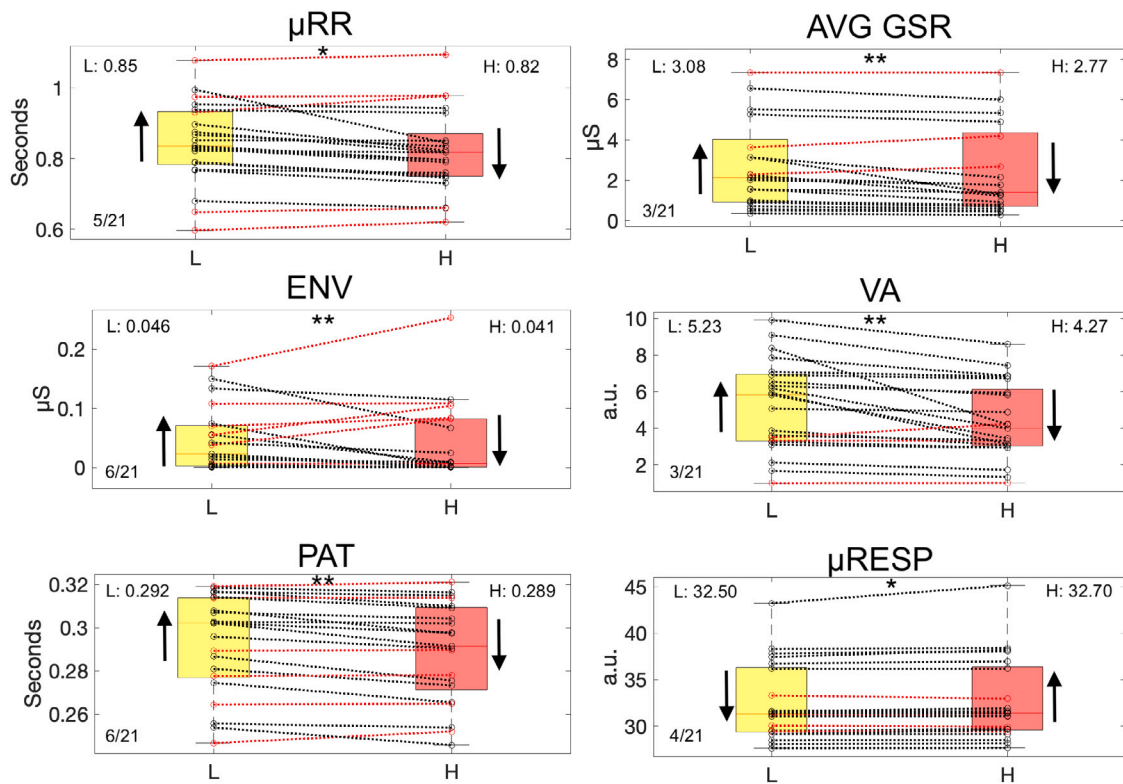


Fig. 3. Boxplots of the features that showed statistically significant differences between low task demand (L) and high task demand (H) phases (μ RR, Env, PAT, Avg GSR, VA and μ RESP). In each plot above, the means for the features divided by L and H are reported. Dotted lines connect data points from the same subjects. The up and down arrows next to the boxplots indicate whether the trend for each feature is increasing or decreasing in either of the task demands. The expected trends are highlighted in black, while opposite trends are shown in red. In the bottom right, the number of subjects out of a total of 21 who follow the trend is indicated. Statistically significant differences are marked with * ($p < 0.05$) and ** ($p < 0.01$) (color online; b/w in print).

- PUPIL: The PUPIL-amplitude related metrics (i.e., AVD and SDD) seem to be related to no significant distinction between the two task demand phases (Table 1), with SDD showing a trend towards lower values during SNT execution compared to the baseline. Regarding frequency-related metrics (i.e., DLF, DHF, DVHF and DLFtoHF), spectral power consistently shows lower levels during the two test phases compared to the baseline, with no significant differences between the two task demand phases.

Fig. 4 panel (a) depicts the features that demonstrate statistically significant differences between high and low task demand phases, with the fewest number of subjects exhibiting opposite trends, using 3D boxes. These features include one cardiovascular measure (VA) and two autonomic measures (μ RESP and AVG GSR). In panel (b), 3D boxes are presented to illustrate one representative feature from each domain: cardiovascular (VA) and autonomic (AVG GSR), and central (β/θ P). VA and AVG GSR showed statistically significant differences between low and high task demand. Regarding the central domain, β/θ P was selected because it showed significant differences between baseline and each of the two task demand conditions and because it is an interpretable measure of attention. Fig. 4 clearly shows that the two examined test phases exhibit distinct physiological patterns in all the three domains. Specifically, during the low task demand phase, several distinct physiological changes can be observed compared to the high task demand phase. Firstly, heightened attention in the parietal region is evident, as indicated by the increased β/θ P ratio in the parietal EEG signal (panel (b)). Secondly, elevated autonomic arousal is observed, as measured by the increased average galvanic skin response (AVG GSR) feature (panels (a) and (b)). Finally, an increase in peripheral blood volume amplitude is observed, as evidenced by the higher VA value (panels (a) and (b)).

4. Discussion

This study aimed at developing a novel adaptive, multi-domain protocol to assess physiological variables associated with LE as a function of task demand in the context of a speech-in-noise recognition task. Specifically, a validated adaptive SNT was employed and two levels of task demand (i.e., low and high) were defined based on the individual SRT. A large set of physiological variables reflecting the cardiovascular, autonomic, and central domains were extracted from the ECG, EEG, GSR, BVP, RESP, and PUPIL signals and used to address differences in physiological responses during high and low task demand phases in a sample of healthy normal-hearing participants.

4.1. Protocol design

The literature extensively discusses LE, defined as the conscious allocation of mental resources to overcome challenges and achieve goals during auditory tasks (Pichora-Fuller et al., 2016). As task demand increases, the corresponding LE is expected to increase, reaching a certain upper limit that is determined by different factors, including the individual's capability and the perceived importance of the task (Brehm and Self, 1989). Moreover, LE is influenced by various aspects including, but not limited to psychosocial considerations (Pichora-Fuller et al., 2016) and fatigue (Hornsby et al., 2016), that is a result of sustained effort during challenging listening tasks. Our protocol design is focused on task demand and is inherently kept short to limit the effects of possible fatigue and task complexity. The stimuli used in the SNT (i.e., VCVs) consist of nonsense words and are presented in alternative-choice task, minimizing the involvement of cognitive processes, as opposed to more complex stimuli and tasks, e.g. open set sentence recognition (Zekveld et al., 2013). Furthermore, the adaptive nature

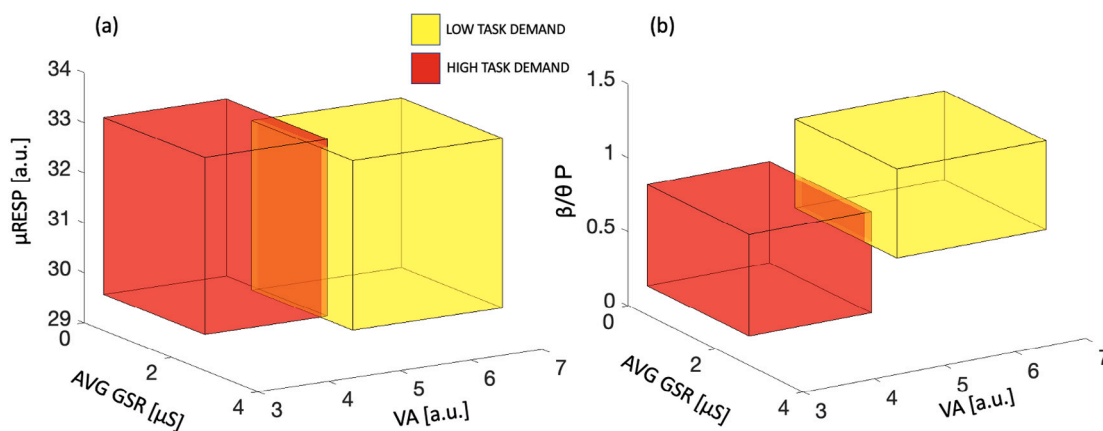


Fig. 4. 3D boxes display features associated with low task demand (L, in yellow) and high task demand (H, in red) phases. Panel (a) presents the features showing the most consistent trends across subjects: cardiovascular (VA) and autonomic (μRESP and AVG GSR), which demonstrated significant differences between demand levels with minimal individual variability. Panel (b) illustrates one representative feature from each physiological domain: cardiovascular (VA), central ($\beta/\theta P$), and autonomic (AVG GSR). The boxes are constructed by centering the rectangles around the coordinates of the corresponding median values, with the length of the sides of the rectangles set equal to the median absolute deviation. (color online; b/w in print).

of the protocol, including both SNT execution based on individual responses and task demand definition based on the individual SRT, ensures that it aligns with the subject's speech recognition performance, preventing it from becoming overly challenging and prone to errors, which may be observed when the task becomes too demanding in relation to the individual's motivation to engage in the test. The distributions in Fig. 1 are consistent with the increase in task difficulty in the high task demand window compared to the low task demand window as the individual performance is substantially worse (lower percentage of correct responses observed in all participants) and that participants need more time to execute the task (higher reaction times in most of the participants, which may be considered an indirect indicator of possibly increased LE).

4.2. Physiological responses

The outcomes presented in Section 3 reveal a consistent pattern across signals and features indicating a substantial rise in sympathetic activation (heightened physiological arousal) during the high task demand phase as shown, e.g., by an increase in GSR amplitude-related features and heart acceleration and demonstrate distinct physiological responses that can be differentiated based on each subject's SRT, even across a range of SNRs within both low and high demand time windows. While the previous observation answers to the first research question of this study, the analysis of the indicators that can help better discriminate LE in different task demand windows, as measured through the proposed protocol, are discussed in detail in the following sections.

4.2.1. ECG and RESP

Looking at the cardiovascular assessment, time-domain features such as μRR , VA, and PAT were effective in distinguishing high from low task demand. Specifically, μRR , representing the average interval of the modeled RR series, significantly decreased during the high task demand phase, indicating an acceleration in heartbeat during the most challenging phase of the test. A similar trend is evident in the PAT feature, representing the time between an R-peak and the onset of the corresponding BVP signal, signifying an acceleration of the pressure wave from the heart to the periphery—a reflection of sympathetic activation related to high task demand. The VA feature showed a similar pattern and decreased in 18 out of 21 participants in the high task demand phase, highlighting its role as a robust physiological variable describing task demand in auditory tasks. VA denotes the amplitude modulation of the BVP signal, representing the volume

of blood on the periphery. A lower BVP amplitude value is linked to greater arterial blood pressure (Tusman et al., 2018), associated with vasoconstriction. The frequency-domain features extracted from the ECG signals show trends that are overall aligned with the increased sympathetic activation suggested by the time-domain features (e.g., lower RR VLF, lower RRTOT, and higher RRLf to HF), although, no statistically significant differences were observed. Specifically, total spectral power (RRTOT) can be considered indicative of heart rate variability, generally associated with healthy parasympathetic driven activation, and RRLf to HF is related to the sympathovagal cardiac balance whereas the interpretation of RR VLF remains a subject of debate, requiring further clarification. As seen in our study, certain research connects RR VLF power with parasympathetic activity (Taylor et al., 1998). Interestingly, here we observe a decrease in RR VLF power during the high task demand phase, akin to RRTOT. Overall, from a cardiovascular standpoint, these findings consistently suggest that the high task demand phase is characterized by heightened sympathetic activation compared to the low task demand phase.

A peculiar aspect of this study is the introduction of the point process paradigm, a novel approach in the field of LE research. The point process enables real-time tracking of HRV indices, overcoming the usual limitations associated with their calculation using conventional algorithms in windows shorter than 5 min (Sassi et al., 2015). Prior studies have demonstrated that time- and frequency-domain characteristics derived from ECG and BVP signals through point process analysis can proficiently differentiate physiological reactions during brief auditory and auditory/visual tasks based on arousal and valence (Polo et al., 2024b), within time frames shorter than those investigated in this study (e.g., 45 s). Granting the opportunity to devise shorter protocols is essential to avoid conflating effort with fatigue and to ensure real-time assessment of physiological responses, for example using the adaptive SNT here used. Citing existing literature, the use of HRV measures in investigating LE is commonly limited and frequently falls short in distinguishing between HRV variables and levels of task demand (Mackersie and Cones, 2011; Francis et al., 2021; Cvijanović et al., 2017). In literature, the predominant feature often employed is the average heart rate of a small number of beats. The cardiovascular analysis conducted in our study emphasizes the significance of addressing features that are typically not addressed in literature. For example, the VA is sensitive to changes in task demand and requires straightforward instrumentation (a photoplethysmograph) and, as such, it can be used as a plausible alternative to the more extensively validated Pre-Ejection Period (Slade et al., 2021; Richter, 2016), a measure well-documented in literature but more intrusive

in terms of instrumentation. Moreover, features related to respiratory and cardiorespiratory coupling indicated a trend towards a slight increase in breath amplitude (as indicated by μRESP) and an increase in respiratory frequency (as indicated by $f_{\text{max}_{HF}}$) with increased task demand, suggesting that participants demonstrated a tendency to take larger and more closely spaced breaths in high task demand scenarios. In addition, the analysis of RSA using the bivariate point process framework indicated a trend towards a higher median value in the low task demand phase suggesting an augmented vagal activation (Bernston et al., 1997). Respiration is commonly overlooked as a biomarker for assessing listening effort. Among the few studies that do investigate the influence of breathing, significant outcomes are seldom achieved (Slade et al., 2021). Furthermore, the respiratory signal itself is frequently not directly examined. Instead, attention is directed towards the power of respiration within the RR signal spectrum, which has been demonstrated to be a pertinent feature reflecting parasympathetic nervous system withdrawal during increased speaking rates, albeit exclusively in subjects with hearing impairment (Mackersie et al., 2015; Mackersie and Calderon-Moultrie, 2016). Through the bivariate point process here used, it became feasible to sample the respiratory signal alongside the cardiac rhythm and perform real-time tracking of its amplitude and frequency in short time windows, highlighting the ability of the related features to assess LE as a function of task demand.

4.2.2. GSR and EEG

Concerning GSR, the literature suggests its utility as a measure of arousal, even at the auditory level, as features related to GSR amplitude tend to exhibit an increase with increasing task demand (Mackersie and Cones, 2011; Mackersie and Calderon-Moultrie, 2016). Nonetheless, inconsistent findings were reported in literature, for example variability in the behavior of GSR between sessions and limited sensitivity to changes in task demand were observed (Giuliani et al., 2020; Holube et al., 2016). In our study, the findings observed by analyzing the GSR signal diverge from those observed from the analysis of other signals, indicating less consistency in defining heightened sympathetic activation during high task demand phases. Interestingly, both Avg GSR and Env, (representing the average amplitude of the GSR signal and the envelope of the GSR phasic component, respectively), are significantly higher during the low task demand phase. This observation gains elucidation through the analysis of features derived from EEG. Despite the reported association of the alpha band with heightened LE (Obleser et al., 2012; Seifi Ala et al., 2020), EEG features proved unable to distinguish between the two test phases in this study. However, the spectral power of the more relevant frequency bands (i.e., α , β , θ) demonstrates a pronounced increase during the two task demand phases compared to the baseline, suggesting higher activation during the SNT task compared to the baseline. Nevertheless, as shown in Table 1 and Fig. 4, higher attention and engagement indices are observed during the low task demand phase. This physiological response pattern aligns with attention theory, particularly Lacey's theory (Lacey et al., 1963). According to this theory, tasks demanding focused attention typically result in a decrease in heart rate (as indicated by μRR in Table 1), often accompanied by an increase in GSR, as consistently observed in Table 1. This is evident, for instance, in features such as Avg Amplitude peaks, Avg GSR, and ENV, which exhibit higher values during periods of low task demand. This heart rate deceleration is a response to environmental stimuli demanding attention, such as the perception of visual or auditory stimuli (Lacey and Lacey, 1970). The elevated attention and engagement indices observed during the low task demand phase indicate heightened attentiveness and involvement. This may be related to a superior comprehension of stimuli compared to the high task demand phase, where understanding stimuli in noise becomes more challenging.

4.2.3. PUPIL

The pupillary signal stands out as one of the most accepted measures for studying LE, with pupillary dilation associated with task demand (Koelewyn et al., 2012; Haro et al., 2022). In this case our study did not yield statistically significant changes in features derived from the PUPIL signal, but revealed a gradual increase in median values with increasing task demand as shown in Table 1. Compared to the literature, this unexpected result can be explained considering that the pupillary data analysis employed a window-based approach, which included both the listening and response phases, rather than the trial-by-trial analysis commonly used in pupillary studies. Also, despite task-evoked changes in pupil size for intermediate tonic levels are independent of baseline pupil size, it remains standard practice to use a baseline-subtracted absolute pupil size as effort indicator, choosing as baseline a time period that goes from 100 ms to 2 s before the stimulus onset (Winn et al., 2018). These choices were made because the duration of SNT trials is insufficient to observe both a pupillary peak and a return to baseline, and to maintain consistency with the analysis of other physiological signals. However, this approach may limit the direct comparability of our pupillary findings with some previous studies.

In this particular context, the observation becomes relevant as we delve into the efficacy of less intrusive variables derived from BVP and GSR signals, which have contributed significantly to a more nuanced assessment of LE and can be measured with simpler sensors than more widely used measures found in literature, such as EEG and pupillary signals. Unfortunately, the intricate nature of LE in complex auditory tasks likely contributes to the absence of consensus in literature regarding the optimal selection of physiological measures. What sets this study apart is the exploration of two distinct levels of task demand, systematically monitoring a diverse array of physiological signals to discern their patterns. Nevertheless, the complex interplay between task demand, physiological responses, and cognitive engagement introduces an additional layer of complexity. While heightened task demand is conventionally associated with increased sympathetic activation, it may simultaneously lead to diminished attention and engagement, initiating a trade-off phase as delineated by motivation theory (Pichora-Fuller et al., 2016). Subjects in this phase strategically evaluate their cognitive resources to determine whether to escalate effort or concede due to the perceived demand of the task. This dynamic process may, in turn, result in reduced attention when subjects do not fully comprehend all presented stimuli.

4.3. Implications, limitations, and future research

BVP and GSR are recorded in continuous time and non invasively, therefore any setup using features extracted from these signals can be easily translated into simple setups (using, for example, wearable sensors) that assess and monitor LE as a function of task demand. The use of an adaptive protocol such as the one introduced here allows for LE monitoring in real time, so that task demand can be defined quickly and in a personalized way as a function of individual auditory ability. As such, the adaptive protocol here employed can help limiting the number of stimuli that become too easy or too difficult to recognize and ensures that the task remains challenging but not overly difficult, reducing the likelihood of participants becoming overwhelmed or disengaged. This can be particularly helpful when individuals with varying speech recognition performance are tested including, for example, individuals with hearing loss. Anchoring the physiological analysis windows to each subject's individual SRT allows for a fair comparison of low vs. high demand listening across conditions and individuals. This approach can be applied to compare different acoustic backgrounds or other experimental manipulations in future studies, as long as an adaptive procedure is used to estimate SRT in each condition. By comparing the physiological effects of high vs. low demand listening relative to the condition-specific SRT, researchers can investigate how

Table A.1

The Table reports for each subject, from 1 to 21, the number of trials in the low task demand window (trials_L) and high task demand window (trials_H), the mean dB SNR in the two windows (snr_L and snr_H , respectively), the average STOI value [%] of the stimuli in the two windows (STOI_L and STOI_H , respectively), and the individual SRT measured by the adaptive speech-in-noise test.

Subject	trials_L (n.)	trials_H (n.)	snr_L (dB)	snr_H (dB)	STOI_L (%)	STOI_H (%)	SRT (dB)
1	24	28	-4.18 (6.43)	-17.16 (1.68)	78.24 (11.29)	49.87 (3.02)	-17.25
2	17	17	-10.26 (1.55)	-16.09 (1.73)	65.09 (3.98)	53.12 (3.29)	-15.25
3	21	18	-2.99 (5.98)	-12.31 (1.88)	80.48 (10.22)	60.99 (5.18)	-11.25
4	24	24	-4.18 (6.43)	-18.72 (2.11)	78.24 (11.29)	47.42 (3.74)	-16.00
5	17	17	-2.57 (5.49)	-11.29 (0.92)	81.56 (9.06)	63.75 (3.69)	-11.00
6	25	24	-11.19 (2.50)	-19.02 (1.19)	64.28 (5.34)	46.86 (2.59)	-18.75
7	24	18	-4.18 (6.43)	-19.28 (2.80)	78.24 (11.29)	46.88 (5.42)	-17.25
8	12	12	0.83 (5.16)	-8.48 (1.58)	87.60 (7.33)	73.36 (3.92)	-8.00
9	31	28	-5.64 (6.29)	-17.56 (2.23)	75.28 (11.61)	50.03 (4.04)	-14.75
10	15	15	-9.03 (1.99)	-16.27 (1.13)	68.68 (5.15)	53.00 (2.80)	-14.75
11	24	23	-4.18 (6.43)	-16.52 (1.08)	78.24 (11.29)	51.56 (2.75)	-16.25
12	21	20	-2.99 (5.98)	-15.79 (1.64)	80.48 (10.22)	54.39 (4.16)	-14.50
13	24	20	-4.18 (6.43)	-18.31 (2.05)	78.24 (11.29)	47.46 (3.86)	-15.25
14	24	23	-4.18 (6.43)	-19.49 (2.03)	78.24 (11.29)	46.18 (3.56)	-18.00
15	24	24	-4.18 (6.43)	-17.39 (1.24)	78.24 (11.29)	50.61 (2.80)	-16.50
16	13	12	-3.63 (2.43)	-10.90 (1.49)	81.59 (5.43)	67.06 (3.92)	-9.50
17	20	19	-6.46 (4.05)	-17.09 (1.66)	74.67 (8.40)	50.93 (3.22)	-16.50
18	18	17	-12.50 (3.49)	-21.90 (2.24)	59.75 (7.86)	42.70 (3.54)	-19.25
19	21	20	-9.18 (2.59)	-16.55 (1.47)	68.79 (6.37)	51.96 (3.43)	-14.75
20	33	33	-7.11 (3.52)	-17.31 (0.98)	74.34 (7.11)	50.58 (3.08)	-17.25
21	24	22	-4.18 (6.43)	-19.00 (1.56)	78.24 (11.29)	47.07 (3.46)	-17.75

different conditions influence physiological processes associated with LE while accounting for individual differences in performance. However, it is important to acknowledge that the observed physiological responses may vary depending on the specific manipulations employed, and future research should explore the generalizability of this protocol across different experimental contexts.

Notwithstanding the encouraging results, this study has some limitations. For example, in this study the slope of the speech recognition function was not considered when defining task demand conditions due to difficulties in estimating slope from short adaptive procedures such as the one here used. However, slope is a factor that potentially influences LE. While this likely does not impact comparisons among normal-hearing participants, it limits our ability to generalize findings across studies and prevents a full assessment of how auditory stimuli located different points along the psychometric curve may affect LE. This gap could be explored in future studies through the analysis of physiological responses on multiple SNR levels that could allow a reliable computation of slope and a deeper investigation on its influence on the relationship between task demand and LE. Moreover, only individuals with normal hearing were involved. Further research involving a larger sample of participants, including both individuals with normal hearing and with varying degrees of hearing impairment, would be important to address more specifically the ability of the proposed protocol to capture different levels of task demand on an individual basis, further validating the features here identified. Moreover, in this study LE was assessed only during an adaptive SNT using VCV stimuli, and task demand was defined based on SNR criteria only, thus lacking to simulate an ecological condition e.g., conversational stimuli, real-world LE conditions that would be more appropriate for LE characterization. Specifically, other strategies for manipulating LE, including both task demand manipulation (e.g., speech distortions, competing speech, simulated three-dimensional challenging listening conditions) and motivation (e.g., inclusion of reward mechanisms) need to be further investigated in order to more accurately assess the ability of features extracted from physiological signals to characterize LE in varying listening conditions and in individuals with varying hearing abilities. Also, different kinds of stimuli, such as sentences or words, should be used in future studies to acquire responses that better reflect ecological, real-world speech processing. In addition, in this study the EEG signal was analyzed only using spectral features summarizing brain activity in the whole temporal window. Further investigation of the EEG signal, using features able to capture the time-varying nature and the spatial distribution of cortical responses, may help define further,

more specific, real-time measures of LE, potentially reaching a lower temporal resolution and thus potentially addressing the responses to single stimulus presentation trials. Regarding pupillary response, future research could explore alternative analysis approaches that combine trial-by-trial pupillary analysis with window-based analysis of slower physiological signals to further refine our understanding of listening effort.

Moreover, the current protocol involves a degree of “a posteriori” processing, as the physiological comparison windows are selected after the individual’s SRT is measured and the adaptive procedure is completed. This introduces a slight delay compared to a fully real-time system. Future work should strive towards developing real-time processing methods that can assess listening effort on a moment-by-moment basis, without the need for post-hoc analysis. Such advancements would enable researchers and clinicians to monitor and respond to changes in listening effort more quickly and effectively. Nevertheless, the key principles demonstrated in this study, such as using individualized performance metrics and strategic window selection, provide valuable insights that could inform the development of real-time systems in the near future. In addition, the ability of exploring physiological responses in short time windows, aided by smart advanced algorithms like point process modeling, enables the application of adaptive machine learning models that can differentiate LE levels, aligning with other assessments used to investigate stress (Xu et al., 2015), emotions (Polo et al., 2024a), and cognitive load (Liu et al., 2023). Overall, this fusion of signals from wearable sensors with machine learning holds significant promise for scientific advancements in practical applications, for example for future real-time optimization of human-machine interaction or for devising novel strategies for adapting the hearing aid to the individual LE, in addition to the individual auditory profile. Towards the development of systems for physiologically-driven control of human-machine interfaces, further analysis of rapid responses (e.g., EEG, pupillary response) on a trial-by-trial basis will be important to identify measures able to monitor LE in real time.

Another peculiar aspect of the study lies in the non-randomized nature of the task demand phases due to the adaptive SNT design, with the low task demand phase typically preceding the high task demand phase. This choice was aimed at limiting the possible influence of increased arousal in the high difficulty phase on physiological responses in low difficulty phases. The adaptive and continuous nature of the test, designed to be as short as possible, does not include a baseline phase separating the two difficulty windows and the same baseline was used to identify changes in physiological features in the low and

Table A.2

Median and median absolute deviation of the statistically significant features extracted from physiological signals in the Baseline (B), low (L), and high (H) task demand windows. Statistical significance is determined at the 0.05 level. Statistically significant differences are reported in the third column. The fourth column shows the p-values related to the comparisons, and the fifth describes the test performed: 0 if ANOVA and 1 if the Friedman's test.

	B	L	H	*	p	ANOVA/Friedman
ECG						
μ_{RR} [s] $\cdot 10^{-2}$	84.60 (9.44)	83.60 (8.81)	81.80 (8.53)	L-H	0.03	0
RRTOT [s ²] $\cdot 10^{-2}$	0.46 (0.63)	0.36 (0.19)	0.21 (0.17)	B-H	0.01	1
RRVLF [s ²] $\cdot 10^{-2}$	0.13 (0.56)	0.15 (0.14)	0.07 (0.09)	B-H	0.02	1
EEG						
PSD α F $\cdot 10^{-2}$	2.07 (2.35)	4.97 (3.21)	4.25 (2.81)	B-L,B-H	e-4,0.04	1
PSD β F $\cdot 10^{-2}$	2.03 (3.45)	14.07 (7.92)	7.55 (7.84)	B-L,B-H	e-4,0.001	1
PSD θ F $\cdot 10^{-2}$	7.80 (3.32)	11.40 (2.84)	11.04 (3.57)	B-L	0.03	0
PSD α P $\cdot 10^{-2}$	2.48 (3.27)	6.41 (4.28)	4.84 (4.09)	B-L,B-H	0.002,0.02	0
PSD β P $\cdot 10^{-2}$	2.48 (4.39)	11.98 (7.54)	10.99 (7.53)	B-L,B-H	0.001,0.03	1
β/θ F	0.11 (0.74)	1.01 (0.81)	0.97 (1.51)	B-L	0.01	1
β/θ P	0.20 (0.48)	0.94 (0.60)	0.50 (0.69)	B-L,B-H	0.002,0.006	1
β/α F	0.59 (1.79)	1.75 (1.24)	1.56 (2.70)	B-L,B-H	0.03,0.01	1
β/α P	0.56 (0.88)	1.56 (0.56)	1.41 (0.86)	B-L,B-H	0.01,0.04	0
GSR						
Avg GSR [μ S]	1.81 (2.13)	2.12 (2.22)	1.40 (2.32)	B-L,L-H	0.01,0.003	1
Env [μ S] $\cdot 10^{-2}$	1.88 (3.78)	2.27 (4.25)	0.67 (4.97)	L-H	0.009	1
BVP						
VA [a.u.]	4.88 (1.89)	5.82 (2.11)	3.99 (1.73)	B-H,L-H	0.01,e-4	0
PAT [s] $\cdot 10^{-2}$	29.96 (1.93)	30.21 (1.95)	29.13 (1.93)	L-H	0.003	0
RESP						
μ_{RESP} [a.u.]	31.07 (3.25)	31.30 (3.39)	31.39 (3.54)	B-H,L-H	e-5,0.01	1
PUPIL						
SDD [mm]	0.22 (0.10)	0.17 (0.03)	0.17 (0.07)	B-L,B-H	e-4,0.02	0
DLF [mm ²]	0.92 (3.33)	0.41 (0.23)	0.28 (0.10)	B-L,B-H	0.003,0.003	1
DVHF [mm ²]	0.49 (0.50)	0.28 (0.18)	0.29 (0.34)	B-L	0.03	0
DLFtoHF	0.98 (1.20)	0.63 (0.48)	0.58 (0.29)	B-L,B-H	0.003,e-4	0

high task demand windows. As such, the current design helped limit the effect of the test sequence on physiological responses by presenting the low task demand phase first, as the experience of struggling with speech understanding is undoubtedly stimulating and would likely influence the subsequent low demand phase if presented in reverse order. Future investigations should explore randomized task demand conditions, including baseline phases between the two difficulty levels, and assess the potential impact of the fixed sequence on specific physiological features to ensure the robustness and generalizability of the results. It should be clarified that the protocol's design, specifically its focus on task demands and VCV stimuli, resulted in the exclusion of certain factors from the experimental evaluation, such as cognitive load and motivation. These factors are important for a comprehensive understanding of listening effort responses and should be addressed in future research.

5. Conclusion

The present study delves into six distinct autonomous and central physiological responses whose information is carried by specific signals such as EEG, GSR, ECG, RESP, PUPIL, and BVP. The objective is to construct a physiological framework tied to variations in task demand within an adaptive speech-in-noise test. The fast execution and simplicity of the test sought to alleviate the onset of fatigue and focus solely on the examination of the task demand component. Notably, there was a marked consistency among physiological signals indicating an increase in sympathetic response during the high task demand phase evident, for example, in heightened heart rate, blood pressure, and breath amplitude. In summary, BVP volume amplitude and point process-derived breath amplitude, emerge as the most discriminative features between the two test phases, whereas the GSR signal, typically linked to arousal, points at heightened attention during the low task demand phase, complemented by increased engagement, also supported by the

analysis of spectral features extracted from EEG signals. As a final overarching message, this investigation highlights the centrality of having to assess a person's attentive state in order to understand physiological responses, particularly when exploring effort exertion during listening.

Funding sources

This work was supported in part by the Italian Ministry of Universities and Research, under the complementary actions to the NRRP "Fit4MedRob-Fit for Medical Robotics" Grant (#PNC0000007). Speech-in-noise test development was supported in part by the Capita Foundation through Project WHISPER, Widespread Hearing Impairment Screening and PrEvention of Risk (2022 and 2020 Capita Foundation Auditory Research Grants).

CRedit authorship contribution statement

Edoardo Maria Polo: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Davide Simeone:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation. **Maximiliano Molura:** Writing – review & editing, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Alessia Paglialonga:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Riccardo Barbieri:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

See Tables A.1 and A.2.

Data availability

Data will be made available on request.

References

- Alhanbali, S., Dawes, P., Millman, R., Munro, K., 2019. Measures of listening effort are multidimensional. *Ear Hear.* 40 (5), 1084–1097. <http://dx.doi.org/10.1097/AUD.0000000000000697>.
- Bakker, J., Pechenizkiy, M., Sidorova, N., 2011. What's your current stress level? Detection of stress patterns from GSR sensor data. In: 2011 IEEE 11th International Conference on Data Mining Workshops. IEEE, <http://dx.doi.org/10.1109/icdmw.2011.178>.
- Barbieri, R., Matten, E., Alabi, A., Brown, E., 2005. A point-process model of human heartbeat intervals: new definitions of heart rate and heart rate variability. *Am. J. Physiol.-Heart Circ. Physiol.* 288 (1), H424–H435. <http://dx.doi.org/10.1152/ajpheart.00482.2003>.
- Beatty, J., 1982. Task-evoked pupillary responses processing load and the structure of processing resources. *Psychol. Bull.* 91 (2), 276.
- Berka, C., Levendowski, D., Lumicao, M., et al., 2007. Eeg correlates of task engagement and mental workload in vigilance learning and memory tasks. *Aviat. Space Environ. Med.* 78 (5), B231–B244.
- Bernston, G., et al., 1997. Heart rate variability: Origins methods and interpretive caveats. *Psychophysiology* 34 (6), 623–648. <http://dx.doi.org/10.1111/j.1469-8986.1997.tb02140.x>.
- Boucsein, W., 2012. *Electrodermal Activity*. Springer Science & Business Media.
- Braithwaite, J., et al., 2013. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology* 49 (1), 1017–1034.
- Brehm, J., Self, E., 1989. The intensity of motivation. *Annu. Rev. Psychol.* 40 (1), 109–131. <http://dx.doi.org/10.1146/annurev.ps.40.020189.000545>.
- Byrne, D., et al., 1994. An international comparison of long-term average speech spectra. *J. Acoust. Soc. Am.* 96 (4), 2108–2120. <http://dx.doi.org/10.1121/1.410152>.
- Cassani, C., et al., 2022. Selecting a pre-processing pipeline for the analysis of EEG event-related rhythms modulation. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society. EMBC, IEEE, <http://dx.doi.org/10.1109/embc48229.2022.9871394>.
- Chakraborty, A., Sadhukhan, A., Pal, D., Mitra, M., 2024. Automated detection of mental stress using multimodal characterization of PPG signal for AI based healthcare applications. *SN Comput. Sci.* 5 (6), 736.
- Chen, Z., Brown, E., Barbieri, R., 2009. Assessment of autonomic control and respiratory sinus arrhythmia using point process models of human heart beat dynamics. *IEEE Trans. Biomed. Eng.* 56 (7), 1791–1802. <http://dx.doi.org/10.1109/tbme.2009.2016349>.
- Chen, Z., et al., 2010. Dynamic assessment of baroreflex control of heart rate during induction of propofol anesthesia using a point process method. *Ann. Biomed. Eng.* 39 (1), 260–276. <http://dx.doi.org/10.1007/s10439-010-0179-z>.
- Clayton, M., Yeung, N., Kadosh, R., 2015. The roles of cortical oscillations in sustained attention. *Trends in Cognitive Sciences* 19 (4), 188–195.
- Coelli, S., Barbieri, R., Reni, G., Zucca, C., Bianchi, A., 2017. Eeg indices correlate with sustained attention performance in patients affected by diffuse axonal injury. *Med. Biol. Eng. Comput.* 56 (6), 991–1001. <http://dx.doi.org/10.1007/s11517-017-1744-5>.
- Cómez, C., Vázquez, M., Vaquero, E., López-Mendoza, D., Cardoso, M., 1998. Frequency analysis of the EEG during spatial selective attention. *Int. J. Neurosci.* 95 (1–2), 17–32. <http://dx.doi.org/10.3109/00207459809000646>.
- Cooke, M., Lecumberri, M., Scharenborg, O., van Dommelen, W., 2010. Language-independent processing in speech perception: Identification of english intervocalic consonants by speakers of eight European languages. *Speech Commun.* 52 (11–12), 954–967. <http://dx.doi.org/10.1016/j.specom.2010.04.004>.
- Critchley, H., 2002. Electrodermal responses: what happens in the brain. *Neuroscientist* 8 (2), 132–142.
- Cvijanović, N., Kechichian, P., Janse, K., Kohlrausch, A., 2017. Effects of noise on arousal in a speech communication setting. *Speech Commun.* 88, 127–136. <http://dx.doi.org/10.1016/j.specom.2017.02.001>.
- Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134 (1), 9–21. <http://dx.doi.org/10.1016/j.jneumeth.2003.10.009>.
- Delorme, A., Palmer, J., Onton, J., Oostenveld, R., Makeig, S., 2012. Independent EEG sources are dipolar. *PLoS One* 7 (2), e30135. <http://dx.doi.org/10.1371/journal.pone.0030135>.
- Duchowski, A., Krejtz, K., Krejtz, I., et al., 2018. The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–13.
- Farabbi, A., Mainardi, L., 2022. Eeg analysis of selective attention during error potential BCI experiments. In: 2022 IEEE 21st Mediterranean Electrotechnical Conference. MELECON, IEEE, <http://dx.doi.org/10.1109/melecon53508.2022.9842955>.
- Fleureau, J., Guillotel, P., Huynh-Thu, Q., 2012. Physiological-based affect event detector for entertainment video applications. *IEEE Trans. Affect. Comput.* 3 (3), 379–385. <http://dx.doi.org/10.1109/t-affc.2012.2>.
- Francis, A., Bent, T., Schumaker, J., Love, J., Silbert, N., 2021. Listener characteristics differentially affect self-reported and physiological measures of effort associated with two challenging listening conditions. *Atten. Percept. Psychophys.* 83 (4), 1818–1841. <http://dx.doi.org/10.3758/s13414-020-02195-9>.
- Frantzidis, C., et al., 2010. On the classification of emotional biosignals evoked while viewing affective pictures: An integrated data-mining-based approach for healthcare applications. *IEEE Trans. Inf. Technol. Biomed.* 14 (2), 309–318. <http://dx.doi.org/10.1109/titb.2009.2038481>.
- García-Pérez, M., 1998. Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vis. Res.* 38 (12), 1861–1881. [http://dx.doi.org/10.1016/s0042-6989\(97\)00340-4](http://dx.doi.org/10.1016/s0042-6989(97)00340-4).
- Giuliani, N., Brown, C., Wu, Y.-H., 2020. Comparisons of the sensitivity and reliability of multiple measures of listening effort. *Ear Hear.* 42 (2), 465–474. <http://dx.doi.org/10.1097/aud.0000000000000950>.
- Greco, A., Valenza, G., Scilingo, E., 2016. *Advances in Electrodermal Activity Processing with Applications for Mental Health*. Springer International Publishing, <http://dx.doi.org/10.1007/978-3-319-46705-4>.
- Haro, S., Rao, H., Quatieri, T., Smalt, C., 2022. Eeg alpha and pupil diameter reflect endogenous auditory attention switching and listening effort. *Eur. J. Neurosci.* 55 (5), 1262–1277. <http://dx.doi.org/10.1111/ejn.15616>.
- Hétu, R., Riverin, L., Lalande, N., Getty, L., St-Cyr, C., 1988. Qualitative analysis of the handicap associated with occupational hearing loss. *Br. J. Audiol.* 22 (4), 251–264. <http://dx.doi.org/10.3109/03005368809076462>.
- Holube, I., Haeder, K., Imbery, C., Weber, R., 2016. Subjective listening effort and electrodermal activity in listening situations with reverberation and noise. *Trends Hear.* 20, 233121651666773. <http://dx.doi.org/10.1177/2331216516667734>.
- Hornsby, B., Naylor, G., Bess, F., 2016. A taxonomy of fatigue concepts and their relation to hearing loss. *Ear Hear.* 37 (1), 136S–144S. <http://dx.doi.org/10.1097/aud.0000000000000289>.
- Houtveen, J., Rietveld, S., De Geus, E., 2002. Contribution of tonic vagal modulation of heart rate central respiratory drive respiratory depth and respiratory frequency to respiratory sinus arrhythmia during mental stress and physical exercise. *Psychophysiology* 39 (4), 427–436.
- Kim, J., Andre, E., 2008. Emotion recognition based on physiological changes in music listening. *IEEE Trans. Pattern Anal. Mach. Intell.* 30 (12), 2067–2083. <http://dx.doi.org/10.1109/tpami.2008.26>, Institute of Electrical and Electronics Engineers (IEEE).
- Klimesch, W., 1999. Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Rev.* 29 (2–3), 169–195.
- Koelewijn, T., Zekveld, A., Festen, J., Kramer, S., 2012. Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear Hear.* 33 (2), 291–300. <http://dx.doi.org/10.1097/aud.0b013e3182310019>.
- Lacey, J., Kagan, J., Lacey, B., Moss, H., 1963. The visceral level: Situational determinants and behavioral correlates of autonomic response patterns. In: Knapp, P. (Ed.), *Expression of the Emotions in Man*. New York International University Press, pp. 161–196.
- Lacey, J., Lacey, B., 1970. Some autonomic-central nervous system interrelationships. In: Black, P. (Ed.), *Physiological Correlates of Emotion*. Academic Press, New York, pp. 205–228.
- Laeng, B., Endestad, T., 2012. Bright illusions reduce the eye's pupil. *Proc. Natl. Acad. Sci.* 109 (6), 2162–2167.
- Leek, M., 2001. Adaptive procedures in psychophysical research. *Percept. Psychophys.* 63 (8), 1279–1292. <http://dx.doi.org/10.3758/bf03194543>.
- Leensen, M., de Laat, J., Snik, A., Dreschler, W., 2011. Speech-in-noise screening tests by internet part 2: improving test sensitivity for noise-induced hearing loss. *Int. J. Audiol.* 50 (11), 835–848. <http://dx.doi.org/10.3109/14992027.2011.595017>.
- Lisetti, C., Nasoz, F., 2004. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP J. Adv. Signal Process.* 2004 (11), <http://dx.doi.org/10.1155/s1108657044006192>, Springer Science and Business Media LLC.
- Liu, Y., et al., 2023. Cognitive load prediction from multimodal physiological signals using multiview learning. *IEEE J. Biomed. Health Inform. Inst. Electr. Electron. Eng. (IEEE)* 1–11. <http://dx.doi.org/10.1109/jbhi.2023.3346205>.
- Mackersie, C., Calderon-Moultrie, N., 2016. Autonomic nervous system reactivity during speech repetition tasks: Heart rate variability and skin conductance. *Ear Hear.* 37 (1), 118S–125S. <http://dx.doi.org/10.1097/aud.0000000000000305>.
- Mackersie, C., Cones, H., 2011. Subjective and psychophysiological indexes of listening effort in a competing-talker task. *J. Am. Acad. Audiol.* 22 (2), 113–122. <http://dx.doi.org/10.3766/jaaa.22.2.6>.
- Mackersie, C., MacPhee, I., Heldt, E., 2015. Effects of hearing loss on heart rate variability and skin conductance measured during sentence recognition in noise. *Ear Hear.* 36 (1), 145–154. <http://dx.doi.org/10.1097/aud.0000000000000091>.

- Mattys, S., Brooks, J., Cooke, M., 2009. Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cogn. Psychol.* 59 (3), 203–243. <http://dx.doi.org/10.1016/j.cogpsych.2009.04.001>.
- McMahon, C., et al., 2016. Monitoring alpha oscillations and pupil dilation across a performance-intensity function. In: *Frontiers in Psychology*. Vol. 7, Frontiers Media SA, <http://dx.doi.org/10.3389/fpsyg.2016.00745>.
- McShefferty, D., Whitmer, W., Akeroyd, M., 2015. The just-noticeable difference in speech-to-noise ratio. *In: Trends in Hearing*. Vol. 19, SAGE Publications, 233121651557231. <http://dx.doi.org/10.1177/2331216515572316>.
- Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., Maess, B., 2012. Adverse listening conditions and memory load drive a common alpha oscillatory network. *J. Neurosci.* 32 (36), 12376–12383. <http://dx.doi.org/10.1523/jneurosci.4908-11.2012>.
- Paglalonga, A., Grandori, F., Tognola, G., 2013. Using the speech understanding in noise (SUN) test for adult hearing screening. *Am. J. Audiol.* 22 (1), 171–174. [http://dx.doi.org/10.1044/1059-0889\(2012\)12-0055](http://dx.doi.org/10.1044/1059-0889(2012)12-0055).
- Paglalonga, A., Polo, E., Lenatti, M., Mollura, M., Barbieri, R., 2023. A screening platform for hearing loss and cognitive decline: WHISPER (widespread hearing impairment screening and Prevention of risk). *Stud. Health Technol. Inform.* 309, 170–174. <http://dx.doi.org/10.3233/shti230768>.
- Paglalonga, A., Polo, E., Zanet, M., Rocco, G., Van Waterschoot, T., Barbieri, R., 2020. An automated speech-in-noise test for remote testing: Development and preliminary evaluation. *Am. J. Audiol.* 29 (3S), 564–576. http://dx.doi.org/10.1044/2020_AJA-19-00071.
- Paglalonga, A., Tognola, G., Grandori, F., 2014. A user-operated test of suprathreshold acuity in noise for adult hearing screening: The sun (Speech Understanding in Noise) test. *Comput. Biol. Med.* 52, 66–72. <http://dx.doi.org/10.1016/j.combiomed.2014.06.012>, Elsevier BV.
- Parreira, J., Chalumuri, Y., Mousavi, A., Modak, M., Zhou, Y., Sanchez-Perez, J., Hahn, J., 2023. A proof-of-concept investigation of multi-modal physiological signal responses to acute mental stress. *Biomed. Signal Process. Control* 85, 105001.
- Partala, T., Surakka, V., 2003. Pupil size variation as an indication of affective processing. *Int. J. Hum.-Comput. Stud.* 59 (1–2), 185–198. [http://dx.doi.org/10.1016/S1071-5819\(03\)00017-X](http://dx.doi.org/10.1016/S1071-5819(03)00017-X).
- Peelle, J., 2018. Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear Hear.* 39 (2), 204–214. <http://dx.doi.org/10.1097/aud.0000000000000494>.
- Petersen, E., Wöstmann, M., Obleser, J., Stenfelt, S., Lunner, T., 2015. Hearing loss impacts neural alpha oscillations under adverse listening conditions. *Front Psychol.* 6, 177. <http://dx.doi.org/10.3389/fpsyg.2015.00177>.
- Picard, R., Vyzas, E., Healey, J., 2001. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Trans. Pattern Anal. Mach. Intell.* 23 (10), 1175–1191. <http://dx.doi.org/10.1109/34.954607>.
- Pichora-Fuller, M., et al., 2016. Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear Hear.* 37 (1), 5S–27S. <http://dx.doi.org/10.1097/aud.0000000000000312>.
- Pion-Tonachini, L., Kreutz-Delgado, K., Makeig, S., 2019. Iclabel: An automated electroencephalographic independent component classifier dataset and website. *Neuroimage* 198, 181–197. <http://dx.doi.org/10.1016/j.neuroimage.2019.05.026>.
- Plain, B., Richter, M., Zekveld, A., Lunner, T., Bhuiyan, T., Kramer, S., 2020. Investigating the influences of task demand and reward on cardiac pre-ejection period reactivity during a speech-in-noise task. *Ear Hear.* 42 (3), 718–731. <http://dx.doi.org/10.1097/aud.0000000000000971>.
- Polo, E., Farabbi, A., Mollura, M., Mainardi, L., Barbieri, R., 2024a. Understanding the role of emotion in decision making process: using machine learning to analyze physiological responses to visual auditory and combined stimulation. In: *Frontiers in Human Neuroscience*. 17, 1286621. <http://dx.doi.org/10.3389/fnhum.2023.1286621>.
- Polo, E., Farabbi, A., Mollura, M., Paglalonga, A., Mainardi, L., Barbieri, R., 2024b. Comparative assessment of physiological responses to emotional elicitation by auditory and visual stimuli. *IEEE J. Transl. Eng. Health Med.* 12, 171–181. <http://dx.doi.org/10.1109/jtehm.2023.3324249>.
- Polo, E., Mollura, M., Barbieri, R., Paglalonga, A., 2023. Multivariate Classification of Mild and Moderate Hearing Loss using a Speech-in-Noise Test for Hearing Screening At a Distance. In: *Lecture Notes of the Institute for Computer Sciences Social Informatics and Telecommunications Engineering*. Springer Nature Switzerland, pp. 81–92. http://dx.doi.org/10.1007/978-3-031-28663-6_7.
- Polo, E., Mollura, M., Paglalonga, A., Barbieri, R., 2022. Listening effort: Cardiovascular investigation through the point process. In: *2022 Computing in Cardiology*. CinC, Vol. 498, IEEE, pp. 1–4. <http://dx.doi.org/10.22489/CinC.2022.211>.
- Pong, M., Fuchs, A., 2000. Characteristics of the pupillary light reflex in the macaque monkey: Discharge patterns of pretectal neurons. *J. Neurophysiol.* 84 (2), 964–974. <http://dx.doi.org/10.1152/jn.2000.84.2.964>.
- Rabiner, L., Gold, B., Yuen, C., 1978. Theory and application of digital signal processing. *IEEE Trans. Syst. Man Cybern.* 8 (2), <http://dx.doi.org/10.1109/tsmc.1978.4309918>, 146–146.
- Richter, M., 2016. The moderating effect of success importance on the relationship between listening demand and listening effort. *Ear Hear.* 37 (1), 111S–117S. <http://dx.doi.org/10.1097/aud.0000000000000295>.
- Rocco, G., Bernardi, G., Ali, R., van Waterschoot, T., Polo, E., Barbieri, R., Paglalonga, A., 2023. Characterization of the intelligibility of vowel-consonant-Vowel (VCV) recordings in five languages for application in speech-in-noise screening in multilingual settings. *Appl. Sci.* 13 (9), 5344. <http://dx.doi.org/10.3390/app13095344>.
- Roup, C., Green, D., Debacker, J., 2020. The impact of speech recognition testing on state anxiety in Young, middle-age, and older adults. *J. Speech Lang. Hear. Res.* 63, 1–12. http://dx.doi.org/10.1044/2020_JSLHR-19-00246.
- Sassi, R., et al., 2015. Advances in heart rate variability signal analysis: joint position statement by the e-cardiology ESC working group and the European heart rhythm association co-endorsed by the Asia Pacific heart rhythm society. *Europace* 17 (9), 1341–1353. <http://dx.doi.org/10.1093/europace/euv015>.
- Schlauch, R., Rose, R., 1990. Two- three- and four-interval forced-choice staircase procedures: Estimator bias and efficiency. *J. Acoust. Soc. Am.* 88 (2), 732–740. <http://dx.doi.org/10.1121/1.399776>.
- Sedghamiz, H., 2014. Matlab implementation of pan tompkins ECG QRS detector. <http://dx.doi.org/10.13140/RG.2.2.14202.59841>, Unpublished.
- Seeman, S., Sims, R., 2015. Comparison of psychophysiological and dual-task measures of listening effort. *J. Speech Lang. Hear. Res.* 58 (6), 1781–1792. http://dx.doi.org/10.1044/2015_jslhr-h-14-0180.
- Seifi Ala, T., Graversen, C., Wendt, D., Alickovic, E., Whitmer, W., Lunner, T., 2020. An exploratory study of EEG alpha oscillation and pupil dilation in hearing-aid users during effortful listening to continuous speech. *PLoS One* 15 (7), e0235782. <http://dx.doi.org/10.1371/journal.pone.0235782>.
- Shelton, B., Scarrow, I., 1984. Two-alternative versus three-alternative procedures for threshold estimation. *Percept. Psychophys.* 35 (4), 385–392. <http://dx.doi.org/10.3758/bf03206343>.
- Shields, C., Sladen, M., Bruce, I., Kluk, K., Nichani, J., 2023. Exploring the correlations between measures of listening effort in adults and children: A systematic review with narrative synthesis. *Trends Hear.* 27, 233121652211371. <http://dx.doi.org/10.1177/23312165221137116>.
- Slade, K., Kramer, S., Fairclough, S., Richter, M., 2021. Effortful listening: Sympathetic activity varies as a function of listening demand but parasympathetic activity does not. *Hear. Res.* 410, 108348. <http://dx.doi.org/10.1016/j.heares.2021.108348>.
- Strand, J., Brown, V., Merchant, M., Brown, H., Smith, J., 2018. Measuring listening effort: Convergent validity sensitivity and links with cognitive and personality measures. *J. Speech Lang. Hear. Res.* 61 (6), 1463–1486. http://dx.doi.org/10.1044/2018_jslhr-h-17-0257.
- Taal, C., Hendriks, R., Heusdens, R., Jensen, J., 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. pp. 4214–4217. <http://dx.doi.org/10.1109/ICASSP.2010.5495701>.
- Taylor, J., Carr, D., Myers, C., Eckberg, D., 1998. Mechanisms underlying very-low-frequency RR-interval oscillations in humans. *Circulation* 98 (6), 547–555. <http://dx.doi.org/10.1161/01.cir.98.6.547>.
- Tusman, G., et al., 2018. Photoplethysmographic characterization of vascular tone mediated changes in arterial pressure: an observational study. *J. Clin. Monit. Comput.* 33 (5), 815–824. <http://dx.doi.org/10.1007/s10877-018-0235-z>.
- Vaez, N., Desgualdo-Pereira, L., Paglalonga, A., 2014. Development of a test of suprathreshold acuity in noise in Brazilian portuguese: A new method for hearing screening and surveillance. *BioMed Research International* 2014, 652838. <http://dx.doi.org/10.1155/2014/652838>.
- Winn, M., Wendt, D., Koelewijn, T., Kuchinsky, S., 2018. Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends Hear.* 22, 2331216518800869. <http://dx.doi.org/10.1177/2331216518800869>.
- Xu, Q., Nwe, T., Guan, C., 2015. Cluster-based analysis for personalized stress evaluation using physiological signals. *IEEE J. Biomed. Health Inf.* 19 (1), 275–281. <http://dx.doi.org/10.1109/jbhi.2014.2311044>.
- Zanet, M., Polo, E., Lenatti, M., van Waterschoot, T., Mongelli, M., Barbieri, R., 2021. 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. <http://dx.doi.org/10.1109/jbhi.2021.3100368>.
- Zanet, M., Polo, E., Rocco, G., Paglalonga, A., Barbieri, R., 2019. Development and preliminary evaluation of a novel adaptive staircase procedure for automated speech-in-noise testing. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC, IEEE*, <http://dx.doi.org/10.1109/embc.2019.8857492>.
- Zekveld, A., Festen, J., Kramer, S., 2013. Task difficulty differentially affects two measures of processing load: The pupil response during sentence processing and delayed cued recall of the sentences. *J. Speech Lang. Hear. Res.* 56 (4), 1156–1165. [http://dx.doi.org/10.1044/1092-4388\(2012\)12-0058](http://dx.doi.org/10.1044/1092-4388(2012)12-0058).
- Zekveld, A., Koelewijn, T., Kramer, S., 2018. The pupil dilation response to auditory stimuli: Current state of knowledge. *Trends Hear.* 22, 2331216518777174.