

# Combining Cardiovascular and Pupil Features Using k-Nearest Neighbor Classifiers to Assess Task Demand, Social Context, and Sentence Accuracy During Listening

Trends in Hearing  
Volume 28: 1–22  
© The Author(s) 2024  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/23312165241232551  
journals.sagepub.com/home/tia



Bethany Plain<sup>1,2</sup> , Hidde Pielage<sup>1,2</sup>, Sophia E. Kramer<sup>1</sup> ,  
Michael Richter<sup>3</sup>, Gabrielle H. Saunders<sup>4</sup>, Niek J. Versfeld<sup>1</sup>,  
Adriana A. Zekveld<sup>1</sup> and Tanveer A. Bhuiyan<sup>5</sup>

## Abstract

In daily life, both acoustic factors and social context can affect listening effort investment. In laboratory settings, information about listening effort has been deduced from pupil and cardiovascular responses independently. The extent to which these measures can jointly predict listening-related factors is unknown. Here we combined pupil and cardiovascular features to predict acoustic and contextual aspects of speech perception. Data were collected from 29 adults (mean = 64.6 years, SD = 9.2) with hearing loss. Participants performed a speech perception task at two individualized signal-to-noise ratios (corresponding to 50% and 80% of sentences correct) and in two social contexts (the presence and absence of two observers). Seven features were extracted per trial: baseline pupil size, peak pupil dilation, mean pupil dilation, interbeat interval, blood volume pulse amplitude, pre-ejection period and pulse arrival time. These features were used to train k-nearest neighbor classifiers to predict task demand, social context and sentence accuracy. The k-fold cross validation on the group-level data revealed above-chance classification accuracies: task demand, 64.4%; social context, 78.3%; and sentence accuracy, 55.1%. However, classification accuracies diminished when the classifiers were trained and tested on data from different participants. Individually trained classifiers (one per participant) performed better than group-level classifiers: 71.7% (SD = 10.2) for task demand, 88.0% (SD = 7.5) for social context, and 60.0% (SD = 13.1) for sentence accuracy. We demonstrated that classifiers trained on group-level physiological data to predict aspects of speech perception generalized poorly to novel participants. Individually calibrated classifiers hold more promise for future applications.

## Keywords

listening effort, k-nearest neighbor, classification, physiological measures, social context

Received 5 March 2023; Revised 4 January 2024; accepted 25 January 2024

## Introduction

Hearing loss is a chronic condition associated with a myriad of negative consequences, including communication difficulties, stress, and the need for high listening effort (Canlon et al., 2013; Hasson et al., 2011; Héту et al., 1993; Holman et al., 2019; Pichora-Fuller et al., 2016). These issues are exacerbated in challenging acoustic conditions, such as when the signal-to-noise ratio (SNR) is poor (Picou et al., 2013). Beyond acoustic challenges, these listening situations are frequently underpinned by social contexts that may alter a person's listening behavior and experience (Matthen, 2016; Pichora-Fuller et al., 2016; Pielage et al., 2021). Current

<sup>1</sup>Otolaryngology Head and Neck Surgery, Ear & Hearing, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, the Netherlands

<sup>2</sup>Eriksholm Research Centre, Snekersten, Denmark

<sup>3</sup>School of Psychology, Liverpool John Moores University, Liverpool, UK

<sup>4</sup>Manchester Centre for Audiology and Deafness (ManCAD), University of Manchester, Manchester, UK

<sup>5</sup>Demant A/S, Smørum, Denmark

## Corresponding author:

Bethany Plain, Otolaryngology Head and Neck Surgery, Ear & Hearing, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam 1081 HZ, the Netherlands.  
Email: bethanyjoyplain@gmail.com



routine audiological tests are poor predictors of real-life hearing difficulties and do not account for the social context in which they occur (Keidser et al., 2020).

Predictive tools that account for social context might facilitate more complete diagnosis of listening difficulties and would provide audiologists with additional information to direct counseling and rehabilitation. Such diagnostic tools could incorporate physiological measures, which have been reported to reflect listening effort and stress (Diamond & Otter-Henderson, 2007; McGarrigle et al., 2014; Peelle, 2018; Pichora-Fuller et al., 2016). In this report, we examine whether a combination of features measured from the pupils and cardiovascular systems of listeners with hearing loss could be used to classify listening demand, social context, and sentence accuracy during a speech perception task.

### *Determinants and Consequences of Listening Effort*

Listening effort has been defined as “a specific form of mental effort occurring when a task involves listening” (Pichora-Fuller et al., 2016: 11S). In the case of speech, the underlying premise is that when information in the signal is degraded, for instance by hearing loss, and/or by the presence of interfering background noise, there is a greater reliance on cognitive resources to substitute missing information and understand the meaning of the speech (Peelle, 2018; Rönnerberg, 2003; Rönnerberg et al., 2008, 2013; Shinn-Cunningham & Best, 2008). Additionally, hearing loss can detrimentally affect selective attention, making it harder to inhibit and ignore surrounding sounds in the first place (Gatehouse & Akeroyd, 2009; Shinn-Cunningham & Best, 2008).

An important factor contributing to listening effort investment in everyday life is the social context in which listening takes place. Social context is thought to moderate listening effort by affecting the success importance of understanding and responding to speech (Hughes et al., 2018; Mackersie & Kearney, 2017; Pichora-Fuller, 2016). Further, the social connectedness achieved by communication may in itself be rewarding, motivating individuals to invest more effort (Hughes et al., 2018; Matthen, 2016). Hearing loss increases the effort needed to listen (Alhanbali et al., 2017) and can give rise to communication breakdown in social situations, causing embarrassment, stress, and even disengagement (Caduff et al., 2020; Mackersie & Kearney, 2017). Indeed, people with hearing loss report that they regularly withdraw from social situations (Holman et al., 2019) and are at higher risk of social isolation than those with normal hearing (Bott & Saunders, 2021; Shukla et al., 2020).

### *Physiological Responses During Listening*

Researchers have inferred information about cognitive processes, including effort and stress, from various physiological measures (Allen et al., 2014; McGarrigle et al., 2014;

Pichora-Fuller et al., 2016; Richter & Slade, 2017; Ziegler, 2012). These measures include neuroimaging techniques, such as electroencephalography (EEG) (Berger, 1929; Peelle, 2018), that derive information from the central nervous system, as well as measures of the peripheral nervous system, that is, the activity of the sympathetic and parasympathetic nervous systems (Kahneman, 1973). Effort investment during listening has been assessed as alpha power in the parietal lobe using EEG (Obleser & Weisz, 2012), as changes to the size of the pupils (Zekveld & Kramer, 2014), level of skin conductance (Mackersie et al., 2015; Mackersie & Cones, 2011), duration of the pre-ejection period (PEP) (Plain et al., 2020; Richter, 2016; Slade et al., 2021), and heart rate variability (HRV) (Mackersie et al., 2015; Mackersie & Calderon-Moultrie, 2016).

Individually, these different measures have been presented as correlates of listening effort, yet when applied concurrently, often show minimal agreement with one another (Alhanbali et al., 2019; Strand et al., 2018). For example, three studies measured pupillometry alongside EEG during different listening tasks. McMahon et al. (2016) demonstrated poor agreement between degree of pupil dilation and alpha power during a speech perception task. Ala et al. (2020) showed a similarly absent relationship between the same measures during longer stimuli (news clips). Finally, Alhanbali et al. (2019) measured pupil dilation, skin conductance, and alpha power simultaneously during a digit-repetition task and also found only weak associations between the measures. Physiological measures have also demonstrated limited agreement with behavioral (Seeman & Sims, 2015) and self-reported measures of listening effort (Wendt et al., 2016). Therefore, it has been suggested that different types of measures reflect different aspects of listening effort (Alhanbali et al., 2019; Strand et al., 2018, 2021).

To our knowledge, no studies have implemented both pupillometric and cardiovascular measures simultaneously during listening. This may be due to differences in timescales of traditional analyses using these measures. The pupil data are often extracted during the active listening part of each trial (and shortly after) only and subsequently an average is taken across the trials within a block (Winn et al., 2018). However, the cardiovascular data are typically extracted from a whole task block, including masking noise presentation, pauses between trials and the response time (Mackersie et al., 2015; Mackersie & Calderon-Moultrie, 2016; Plain et al., 2020; Seeman & Sims, 2015). In this work, we extracted trial-level responses from both the pupil and cardiovascular systems. Below we explain the measures that were analyzed in the current study.

**Pupil Features.** Listening to and repeating a short sentence elicits a transient pupil dilation, known as the task evoked pupil response (Zekveld et al., 2018). Utilizing this phenomenon, three measures are often extracted from the pupils

during speech perception tasks: the baseline pupil size (BPS), peak pupil dilation (PPD), and mean pupil dilation (MPD). BPS refers to the diameter of the pupil in the one-second period prior to the onset of the target speech. It provides information about the alertness of the individual (Granholm & Steinhauer, 2004) and anticipation of the upcoming task (Ayasse & Wingfield, 2020). PPD refers to the maximum pupil size elicited by the presentation of the target stimulus, in relation to BPS. It has been robustly demonstrated that as the difficulty of a listening task increases, so does PPD (Wendt et al., 2018; Zekveld et al., 2018). This relationship holds true until the task is deemed impossible or not worth the required effort and the participant disengages, at which point PPD reduces in magnitude (Ohlenforst et al., 2017). MPD refers to the mean of the pupil dilation response, from onset of the target stimulus to the response prompt, relative to BPS (Zekveld et al., 2010). Since pupil data are often noisy, MPD is thought to be a more robust measure than PPD for providing information about cognitive resource allocation (Ahern & Beatty, 1979; Verney et al., 2001; Zekveld et al., 2010), particularly at the trial level.

Of these three pupil diameter measures, PPD has been demonstrated to also respond to manipulation of social context during listening tasks. For instance, Zekveld et al. (2019) demonstrated an increase in PPD for both hearing impaired and normal hearing participants who were given evaluative feedback, in the form of verbal and visual input about their performance during a speech perception task, compared to those who were not given feedback. Similarly, Pielage et al. (2021) found an increase in PPD when normal hearing participants performed a speech perception task in tandem with another participant, compared to when the task was performed alone. In the co-present condition, participants took turns to repeat every other sentence. In both of these studies, the authors interpreted the increase in PPD as an increase in effort, related to increased success importance due to the social context manipulation. Interestingly, in Pielage et al.'s study, performance was unchanged by co-presence, whereas Zekveld et al.'s feedback manipulation improved performance at the easier (speech reception threshold 71%), but not the harder condition (speech reception threshold 50%).

**Cardiovascular Features.** Various cardiovascular measures reflecting autonomic nervous system (ANS) activity have also been applied to measure aspects of listening effort, including PEP (described in more detail below), HRV, heart rate, and blood pressure (Mackersie et al., 2015; Mackersie & Calderon-Moultrie, 2016; Mackersie & Cones, 2011; Plain et al., 2020; Richter, 2016; Slade et al., 2021). In addition, cardiovascular measures have been applied during nonlistening tasks to demonstrate the effect of the presence of observers (Bosch et al., 2009).

As described above, cardiovascular measures have generally been averaged across an entire task block (Jennings et al.,

1992). However, some researchers have analyzed shorter, transient cardiovascular responses during listening. For instance, Francis et al. (2016) measured pulse rate and pulse amplitude at the fingertip using photoplethysmography (PPG) during a sentence perception task (see Discussion section for more details). Pulse rate closely relates to heart rate (or interbeat interval, IBI, in our current study, being the inverse of pulse rate). Pulse rate is influenced by both sympathetic and parasympathetic nervous system activity. Whereas pulse amplitude (also known as blood volume pulse amplitude, BVPA) refers to the volume of blood in the capillaries during a heartbeat, and is related to sympathetic nervous system activity (Iani et al., 2004; Liu et al., 2021; Nitzan et al., 1998). An increase in sympathetic activity during mental effort leads to peripheral vasoconstriction and subsequently a decrease in the BVPA (Iani et al., 2004).

Other measures that may be of interest for trial-by-trial analyses include PEP and pulse arrival time (PAT). PEP is defined as the time interval between the start of the depolarization of the heart's left ventricle and opening of the aortic valve (Newlin & Levenson, 1979; Sherwood et al., 1986, 1990). Similar to BVPA, PEP is also recognized as an index of cardiac sympathetic nervous system activity (Ahmed et al., 1972; Newlin & Levenson, 1979). Though typically averaged across a block, it is possible to extract PEP during shorter time windows. For instance, Kuipers et al. (2017) extracted IBI and a PEP equivalent, referred to as RZ, at the trial level during a flanker task. Conflict trials in the flanker task led to cardiac deceleration and decreased RZ interval as compared to nonconflict trials. This effect could suggest increased effort investment, however the authors noted that changes in RZ did not occur within the physiologically expected time window, hindering the interpretation of this finding (Kuipers et al., 2017).

PAT is also sometimes referred to as pulse transit time (Chan et al., 2019). PAT consists of the time taken for the arterial pressure wave following a heartbeat (measured by the electrocardiogram) to travel to a more peripheral location, often the earlobe, toe, or fingertip (measured by PPG) (Block et al., 2020). PAT inversely relates to blood pressure (Block et al., 2020) and has been applied as a measure of stress (Hey et al., 2009). For instance, an increase in stress elicited by the Trier Social Stress Test (described below) has been demonstrated to correspond to a decrease in PAT, compared to baseline (Hey et al., 2009). To our knowledge, PAT has not been applied during any listening studies to date.

## Classification

Combining multiple physiological features obtained during listening requires an analysis tool that is able to evaluate the relationship between these features and the response variable. Classification, where algorithms called classifiers learn to categorize data into different classes, can be used to this end (Drummond, 2010). Classifiers trained on physiological

features have been applied within various fields, for example, to diagnose diseases (Sarkar & Leong, 2000), to differentiate emotions (Babiker et al., 2015), to detect attention deficit hyperactivity disorder (Das & Khanna, 2021), and to determine mental states, such as stress (Mozos et al., 2017; Rahman et al., 2015).

In this study, we trained and tested k-Nearest Neighbor (k-NN) classifiers. k-NN is a simple, nonparametric, supervised learning technique that assigns the label of an unlabeled data point based upon the majority vote of its neighbors (Hastie et al., 2009). This is achieved based on the distance between neighboring data points: those within close proximity are likely to be grouped together, whereas those separated by a large distance are not (Hastie et al., 2009). When training the classifiers, the optimal number of neighbors (k) must be selected, where  $k = 1$  means that the data point is labeled based upon its single closest neighbor alone.

## Aims

The main aim was to use a combination of pupil and cardiovascular features to predict acoustic and contextual aspects of a speech perception task. To this end, we trained k-NN classifiers using seven physiological features at the trial level: BPS, PPD, MPD, IBI, BVPA, PEP, and PAT. We trained the classifiers to predict: (1) the task demand level (i.e., the SNR corresponding to 50% versus 80% correct sentence repetition), (2) the social context (i.e., the presence vs. absence of two observers), or (3) sentence accuracy (correct vs. incorrect repetition). We anticipated that including a range of physiological features in our classifiers would provide superior prediction accuracy over individual measures. The rationale was that the features differ in their level of contribution from the sympathetic (SNS) and parasympathetic nervous system (PNS) branches. For example, PEP and BVPA are thought to reflect primarily sympathetic activity (Iani et al., 2004; Newlin & Levenson, 1979), whereas the other features (including the pupil features) are more mixed in ANS origin (Gordan et al., 2015; Steinhauer et al., 2004). The features also exhibit varied responsiveness to different stimuli and states. For instance, BPS is thought to reflect alertness and anticipation (Ayasse & Wingfield, 2020; Granholm & Steinhauer, 2004), whereas PPD and MPD are task evoked phenomena (Zekveld et al., 2010).

## Materials and Methods

### General Methods and Previous Analyses

The data were collected from hearing impaired participants during a speech perception task in a two (task demand) by two (social context) within-subject design. The task was conducted at two individually adapted SNRs corresponding to 50% and 80% correct (referred to here as SNR50% and SNR80%, respectively), and in the presence or absence of two observers. Pupillometric and cardiovascular measures

were recorded simultaneously throughout. The results from pupillometric and cardiovascular measures have previously been analyzed and presented separately (Pielage et al., 2023; Plain et al., 2021).

In Plain et al. (2021), cardiovascular parameters (HRV, PEP, blood pressure, and heart rate) were measured and averaged across blocks of sentences in relation to a baseline period. The main finding of the study was that baseline-corrected blood pressure change scores (systolic, diastolic, and mean arterial blood pressure) increased in the presence of observers. No cardiovascular measures were sensitive to the task demand manipulation. In Pielage et al. (2023), the main physiological outcome measures were PPD and BPS. BPS increased in the presence of the observers and PPD increased at the SNR50% compared to the SNR80% condition. The results of both taken together demonstrated increased physiological arousal or stress caused by the presence of the observers, and an increase in effort investment at the SNR50% condition compared to the SNR80% condition.

### Participants

The data were collected from 29 native Danish speaking, hearing-impaired participants (17 males; average age = 64.6 years,  $SD = 9.2$ ), who were recruited at Eriksholm Research Centre. Participants had symmetrical ( $<15$  dB difference between ears) sensorineural hearing losses (four frequency pure tone averages across 0.5, 1, 2, and 4 kHz were 50.2 dB HL [ $SD = 8.9$ ] for the right ear and 51.3 dB HL [ $SD = 8.7$ ] for the left ear). They were experienced users of Oticon hearing aids. For the purposes of the experiment, they were fitted with bilateral Oticon OPN1 hearing instruments with a double-layered dome attachment. The instruments were programmed to the manufacturer's first fit, microphone settings were omnidirectional; noise reduction, volume control and program functionality were disabled. Participants reported being free from psychiatric, neurological, ocular, or cardiovascular diseases. They provided written informed consent, and all procedures were approved by the Research Ethics Committees of the Capital Region of Denmark.

### Speech Perception Task

**Task Demand.** The task involved a speech perception task, using Danish hearing in noise test (HINT) sentences (Nielsen & Dau, 2011). The participant was required to repeat target sentences spoken by a female talker presented from a frontal loudspeaker in the presence of a four-talker babble masker. The four-talker masker was played from four loudspeakers positioned at 90°, 150°, 210°, and 270° and located 1.2 m away from the participant. Each loudspeaker played back a different recording of a newspaper article with silent gaps longer than 50 ms removed. Two were recordings of male voices and two of female voices; the speech was spectrally altered to match the long-term average speech spectrum

of the target speech. The loudspeaker position of each of the voices was randomized between blocks.

The individualized SNRs for the testing blocks (SNR50% and SNR80%) were determined by two adaptive procedures, one targeting 50% and the other targeting 80% correct. Physiological measures were collected during the adaptive procedures, but these data were not analyzed. The adaptive procedures used were described in detail by Plain et al. (2021). The masker was kept constant at 70 dB sound pressure level (SPL) in the adaptive blocks and task blocks. In the adaptive blocks, the target speech level was manipulated adaptively, whereas in the task blocks, the SNR remained constant throughout. Lists of sentences and sentence presentation order within each were counterbalanced across participants and conditions. All blocks consisted of 20 sentences, each of which had 3 s of babble masker preceding and following the target sentence, which lasted on average 1.5 s (range = 1.2–1.9 s). After the participant's verbal response, the experimenter scored the response and waited around 3 s prior to initiating the next trial. Word scoring was conducted live during the test session for which errors concerning verb tenses, single/plural nouns, definite/indefinite articles, changes to word order and omission or addition of phonemes were permitted (Plain et al., 2021). Word scoring was then converted to sentence scoring, such that the participant had to correctly repeat all words in a sentence to receive a "correct" score for that trial.

**Social Context.** The social context manipulation consisted of the above task blocks being completed either alone, referred to as the "alone" condition, or in the presence of two observers, referred to as the "observed" condition. When present, the observers were seated 1.2 m from the central point, facing inwards (toward the participant), at angles of 45° and 315° with respect to the participant. The participant was told to imagine that they were in a social situation with the observers and that the observers had spoken the target sentences. Observers were instructed to act in a neutral and nonthreatening manner and tasked to assess how good the participant would be as a communication partner in real life. The observers were not previously known to the participant however they were hearing-impaired individuals (recruited from the Eriksholm Research Centre database) of a similar age to the participants so that they could be likely social peers of the participant. The presentation order of experimental conditions was counterbalanced across participants, with the constraint that the two alone blocks occurred consecutively, as did the two observed blocks.

## Physiological Measures

### Pupil Features

**Equipment, Data Collection, and Processing.** Pupil size was measured during the task using a Tobii Spectrum eye

tracker (Tobii, Stockholm, Sweden), at a sampling frequency of 600 Hz. A constant ambient illumination of 200 lux was maintained in the test room for all experimental sessions. Participants were asked to focus on their gaze on the loudspeaker in front of them during stimulus presentation. Pupil data were measured continually during the task blocks and were separated into trials by use of a marker sent at each sentence onset. Sentences one to four were discarded, as the pupil data are considered to be relatively unstable at the start of a block (Winn et al., 2018). Please note that no averaging was conducted within a condition—each trial was considered separately, so that there were more data available with which to train and test the classifiers. Data from the right eye were analyzed unless the tracking of the left eye proved more successful.

The raw pupil data were preprocessed to remove noise and interpolate missing data (Pielage et al., 2023). Missing data were processed differently depending upon the duration of the gaps in the data. Due to an unknown issue, there were some very short gaps (around 5 ms) of missing data (<1% of the full trace). These were interpolated between the last present sample and the third sample after the gap by means of linear interpolation. Gaps longer than 25 ms in duration were not interpolated at this stage. After interpolation of the small data gaps, traces with more than 50% of missing data were excluded, and if more than five sentences in a condition were excluded, then this condition was removed from the analyses. In the accepted trials, any remaining missing samples (>25 ms) were presumed to be caused by blinks. A repeated measures analysis of variance revealed no significant effect of condition on the number of blinks recorded (task demand:  $F[1,26] = 0.01$ ,  $p = .92$ ,  $\eta^2 = .00$ ; social context:  $F[1,26] = 0.00$ ,  $p = .95$ ,  $\eta^2 = .00$ ; interaction:  $F[1,26] = 0.74$ ,  $p = .40$ ,  $\eta^2 = .03$ ). The data surrounding the blinks were removed—the data were cut from 50 samples (83 ms) prior to the first missing sample to the 80th present sample (133 ms) after the blink (Koelewijn et al., 2021). Subsequently, a 51-tap moving average filter that skipped over missing data was applied to smooth the trace and the removed samples were replaced through linear interpolation.

**Feature Extraction.** The following parameters were extracted per trial from the smoothed, processed pupil data: BPS, PPD, and MPD. The BPS corresponds to the average pupil size during the final second of the masking noise, prior to the onset of the target sentence. The BPS of each trial was subtracted from the other pupil data in the same trial, to baseline correct the values. The task evoked pupil response (TEPR) is the portion of the trace from the onset of the target sentence until the end of noise presentation, prior to the verbal response of the participant. During this window, the PPD refers to the maximum value during the TEPR relative to the BPS and finally, the MPD, which refers to the average of all values during the TEPR (also

relative to baseline). For the included sentences, a single value was extracted for each BPS, MPD, and PPD.

### Cardiovascular Features

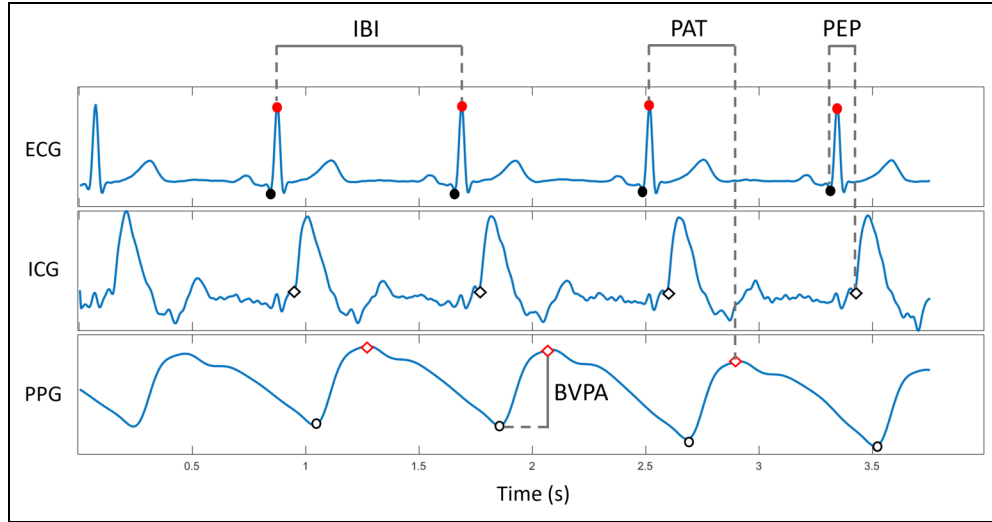
*Equipment, Data Collection, and Processing.* Cardiovascular measures were recorded by the Cardioscreen 2000 system (Medis, Ilmenau, Germany). The Cardioscreen 2000 measured three streams of continuous data throughout the baselines and task: an electrocardiogram (ECG), an impedance cardiogram (ICG) and a PPG. The equipment uses a standard ECG lead configuration for ECG assessment, lead II of Einthoven's triangle. The ECG and ICG were obtained via three disposable, solid gel, surface electrodes: one positioned on the left side of the neck, one beneath the left armpit at the level of the xiphoid process and one 10 cm lower than this. The PPG sensor consisted of a clip-on sensor, placed on the right earlobe of participants. It should be noted that PPG measured at the earlobe and fingertip have demonstrated mixed agreement—responses at the earlobe may be smaller in magnitude (Armañac-Julián et al., 2022; Awad et al., 2001; Fleischhauer et al., 2023; Hartmann et al., 2019; Kushki et al., 2011). ECG and ICG were measured at a sampling frequency of 1000 Hz, and PPG was measured at a sampling frequency of 200 Hz. Trigger signals were sent to the Cardioscreen 2000 at the onset of the babble masker of every trial, to allow the data to be divided into trials. The device also collected discrete blood pressure data, once per block, from a blood pressure cuff worn on participants' right arm. The blood pressure data will not be discussed further in this article, but can be found in Plain et al. (2021).

Separate MATLAB (ver. 2018b; MathWorks, Natick, MA) scripts were created for the cardiovascular data processing of each participant. This was preferable because parameters such as the amplitude of the ECG peak varied significantly between individuals. Firstly, the ECG, ICG, and PPG data were loaded into MATLAB. The PPG signal was up-sampled to 1000 Hz to match the ECG and ICG data. For many participants, the PPG signal was contaminated by an artifact at around 5.3 Hz. Where necessary, this was removed using a zero-phase 22nd order infinite impulse response band stop filter between 5.2 and 5.4 Hz. The PPG, ECG, and ICG signals were filtered using 6th order Butterworth bandpass filters. For the PPG signal, a passband of 0.1 to 8 Hz was applied, whereas for the ECG and ICG signals a passband of 1 to 30 Hz was applied (Raza et al., 1992). Data were subsequently divided into individual trials using the trigger information and the duration of each HINT sentence. For each trial, the cut cardiovascular data consisted of 3s of masking noise, the duration of the sentence, followed by 3s of masking noise. Data obtained during the verbal sentence repetition were discarded and sentences one to four for each condition were also excluded, in keeping with the pupil data.

*Feature Extraction.* Feature extraction was also undertaken in MATLAB (ver. 2018b). Several steps were required to obtain the features from the ECG, ICG, and PPG. The first step involved detecting relevant points in the different data signals (points labeled in Figure 1). For each trial, the R peaks in the ECG signal (indicated with red dots in Figure 1) were detected, representing each heartbeat. Visual inspection of peak detection was conducted to ensure all peaks were correctly detected. Then, for every heartbeat within a trial, semi-automatic procedures were used to detect two points: (1) the Q point of the PQRST complex of the ECG, which corresponds to the onset of left ventricular depolarization (indicated with black dots in Figure 1) and (2) the B point of the ICG, which corresponds to the opening of the aortic valve (indicated with white diamonds in Figure 1). The latter was detected using a tangent-based method. Tangent-based algorithms have been applied successfully to detect other cardiovascular parameters (Escobar-Restrepo et al., 2018; Hermans et al., 2017). The positioning of these points was confirmed to be correct by visual inspection. Where incorrect detection was demonstrated, the parameters of the algorithms were fine-tuned by the experimenter, such that the position was optimal. In addition, B points that were detected greater than two standard deviations from the position of the others within a trial were excluded.

Next, the PPG signal corresponding to each heart cycle was detected. Due to noise in the PPG signal, a signal quality index was calculated for each PPG segment to qualify it for feature extraction (Goldberger et al., 2000; Vest et al., 2018). Each PPG cycle was compared to a template PPG cycle for the same participant. Templates were created by the experimenter, by locating and averaging across an optimal section of the participant's PPG signal, without artifact or contamination. PPG cycles not meeting the quality threshold (those with <80% signal quality score) were discarded and not included in the analysis. The maximum point of the PPG cycle was labeled as the peak, and the minimum point was labeled as the trough (indicated with red diamonds and white circles, respectively, in Figure 1).

The data extracted as above were used to calculate four cardiovascular features in each trial: IBI, BVPA, PEP, and PAT. Figure 1 demonstrates how each of the features were extracted from the ECG, ICG, and PPG signals. IBI was calculated as the difference between each consecutive detected R peak. BVPA was calculated as the difference in amplitude at the foot of the PPG signal to the peak of the PPG signal. PEP was calculated by determining the time duration between the detected Q point of the ECG and B point of the ICG. Finally, PAT was calculated as the duration between the R peak of the ECG and the peak of the PPG signal. Depending on the participant's heart rate they might have around seven values for each feature per trial. The calculated values in each trial for each feature were averaged, such that there was a single number per feature per trial. This approach was selected, because trials contained



**Figure 1.** Extracting cardiovascular features from the ECG, ICG, and PPG signals. Q points (ECG) are denoted by black filled circles and R peaks with red filled circles. B points (ICG) are denoted by black unfilled diamonds. PPG peaks are denoted by red unfilled diamonds and the troughs are denoted by black unfilled circles. IBI is the interval between consecutive R peaks (ECG). PAT is the interval between the R peak (ECG) and the following PPG peak. PEP is the interval between the Q point (ECG) and the B point (ICG). BVPA is the difference in amplitude between the PPG peak and trough. The first cycle has not been annotated to show the morphology of the cycles clearly. BVPA = blood volume pulse amplitude; ECG = electrocardiogram; IBI = interbeat interval; ICG = impedance cardiogram; PAT = pulse arrival time; PEP = pre-ejection period; PPG = photoplethysmogram.

varying numbers of cycles (due to different heart rates) and to align the cardiovascular with the pupil data (i.e., a single PPD value per trial).

**Parameter Optimization.** For each trial of each condition for each participant there were three features from the pupil data (BPS, MPD, and PPD) and four features from the cardiovascular data (IBI, BVPA, PEP, and PAT). All seven features were imported into a table which had 1856 rows (all trials of all participants). Cardiovascular data for one participant and pupil data from two participants were excluded completely due to very poor signal quality. The total number of valid trials for each feature was as follows: BPS, 1665; MPD, 1665; PPD, 1665; IBI, 1721; BVPA, 1626; PEP, 1723; and PAT, 1598. To ensure outliers were removed, data that were more than three standard deviations away from the group mean for each feature was excluded. Ultimately, this resulted in no further trials being excluded for BPS, IBI, and PEP, whereas 21 additional trials were excluded for PPD, 17 for MPD, 32 for BVPA, and one for PAT. Finally, the features were standardized in preparation for their inclusion in the classifiers: the data were centered and scaled according to the feature means and standard deviations, respectively.

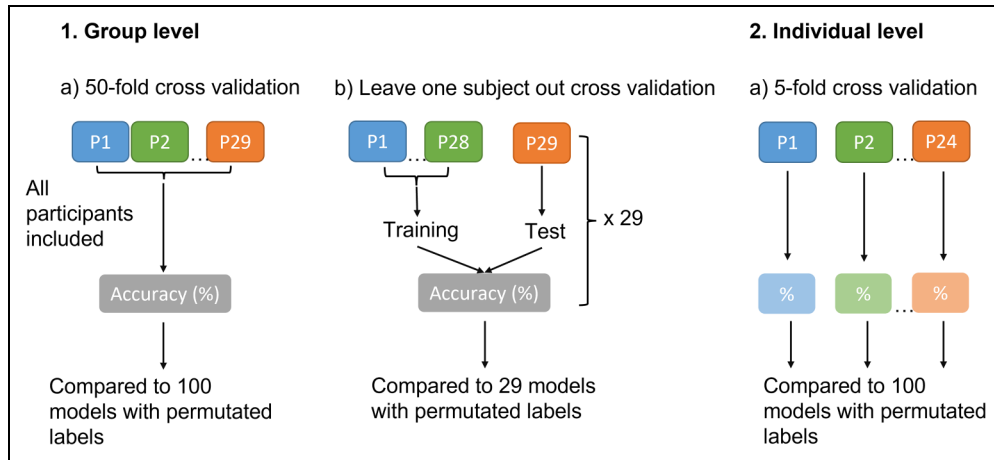
## Classification

In the present study, we trained and tested k-NN classifiers to predict three different aspects of the experiment: (1) the task

demand level (SNR50% or SNR80%), (2) the social context condition (presence or absence of two observers), and (3) the sentence accuracy (correct or incorrect repetition).

All k-NN algorithms were trained using the same physiological features but differed in the number of trials included. The first two classifiers, predicting task demand, and social context, included all data from suitable trials. For the third k-NN classifier predicting sentence accuracy, a subset of trials were included as k-NN classifiers are sensitive to imbalanced datasets (Wah et al., 2016). It contained all trials obtained during the SNR50% task demand condition, plus a subset of trials from the SNR80% condition: data obtained during all incorrect SNR80% trials and a subset of 25% of correct SNR80% trials. This resulted in a balanced dataset (50% correctly and 50% incorrectly repeated trials).

We trained classifiers on data at the group level and at the individual level (see summary of data pipeline in Figure 2). The group classifiers were first trained and tested by means of 50-fold cross validation (method 1a in Figure 2). The k-fold cross validation involves splitting the dataset randomly into k (in this case 50) groups of equal size. One group is held out as the test dataset, while the remaining data are used to train the classifier. When trained, the classifier is tested using the test dataset. This procedure is repeated k times. Though k-fold cross validation ensures that training and test data are separate, validating the classifier using this technique meant it was likely that different trials from the same participant were appearing in both the training and test datasets, which can falsely enhance the results of the



**Figure 2.** Schematic demonstrating the classification analyses conducted at the group and individual level. P = participant.

classifiers (Gholamiangonabadi et al., 2020; Miltiadous et al., 2021; Saeb et al., 2017).

To determine how generalizable the classifiers were to novel participants, we then conducted leave one subject out validation, where one participant's data were isolated from the training dataset and used exclusively for testing (method 1b in Figure 2). This was repeated 29 times, with each participant selected once to appear in the test dataset. This approach simulates a possible clinical application of these classifiers—if the group level classifiers performed well during leave one subject out validation, this suggests that physiological data from a novel, unseen individual could be provided to the classifiers to determine the person's task demand, social context or sentence accuracy while they performed the task.

Having trained k-NN classifiers on the group data, we then trained and tested k-NN classifiers on individual participants' data (method 2 in Figure 2), using five-fold cross validation. This allows classifiers to be calibrated for the individual, removing between subject variability. Training individualized classifiers has been done more often in speech tracking, for example, where the speech signal is encoded based upon the EEG (Fiedler et al., 2019; Jessen et al., 2019; O'Sullivan et al., 2017).

The trained k-NN classifiers were optimized by automatic tuning of hyperparameters in MATLAB (ver. 2018b). The function selected the most appropriate parameters for each classifier including the distance measure and the number of neighbors (k). The optimized classifier results (including distance measures and values of k) are reported here. All seven features were used as the input to each of our classifiers (i.e., no feature selection was undertaken). For each k-NN classifier, we present classifier accuracy, precision, sensitivity, specificity, and F1 score (a measure of the model's accuracy), defined by the formulae below, where TP = true positive, TN = true negative, FP = false positive, and FN = false negative. Classification accuracy (defined by equation

below) was our primary measure of classifier performance, and hence is the main focus of the Results and Discussion sections. The other result measures are only presented in tabular form.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

**Permutation Analysis.** To verify whether our classifiers were operating at above chance level, we conducted permutation analysis on the classifier accuracies. Permutation analysis is a simple, nonparametric technique that can be applied to assess the performance of classifiers (Golland & Fischl, 2003; Ojala & Garriga, 2010). The method involved repeatedly shuffling (or permuting) the data labels, such that they were misaligned with the dataset. For each permutation, a new classifier was trained and tested, allowing a null distribution of classifier accuracies to be estimated. A *p*-value was determined from this distribution, by calculating the proportion of all permutation accuracies that were equal to or that exceeded the original model's accuracy (Anderson & Ter Braak, 2010; Ojala & Garriga, 2010). The null hypothesis stated that there was no difference between the original classifier accuracy and the sampling distribution estimated with randomly shuffled labels. We conducted 100 permutations per original classifier (see Figure 2), allowing minimum *p*-values of 0.01 to be obtained.

**Neighborhood Component Analysis.** Finally, to determine the contribution of each of the seven features to the classifiers, neighborhood component analysis (NCA) was conducted in MATLAB (ver. 2018b) for each model. NCA is a nonparametric technique that enables feature selection with a view to maximizing the accuracy of classifiers (Yang et al., 2012). NCA is often applied with the aim of reducing the complexity and improving the efficiency of the model, by selecting significant features and reducing dimensionality. For example, NCA has been applied to EEG data to this end (Javaid et al., 2015; Raghu & Sriraam, 2018). In the present study, however, we had a relatively small pool of features (seven), and therefore opted to apply NCA to obtain information about the classifiers and the contribution of individual features without utilizing this information further to refine the classifiers. NCA was conducted on the classifiers obtained using k-fold cross validation: the group level and the individual classifiers. We opted not to perform NCA on the group level classifiers obtained using leave one subject out cross validation, as these performed close to chance level.

## Results

### Group Level Classifiers

**50-Fold Cross Validation.** Classifier parameters (number of neighbors and distance metric) and results (accuracy,

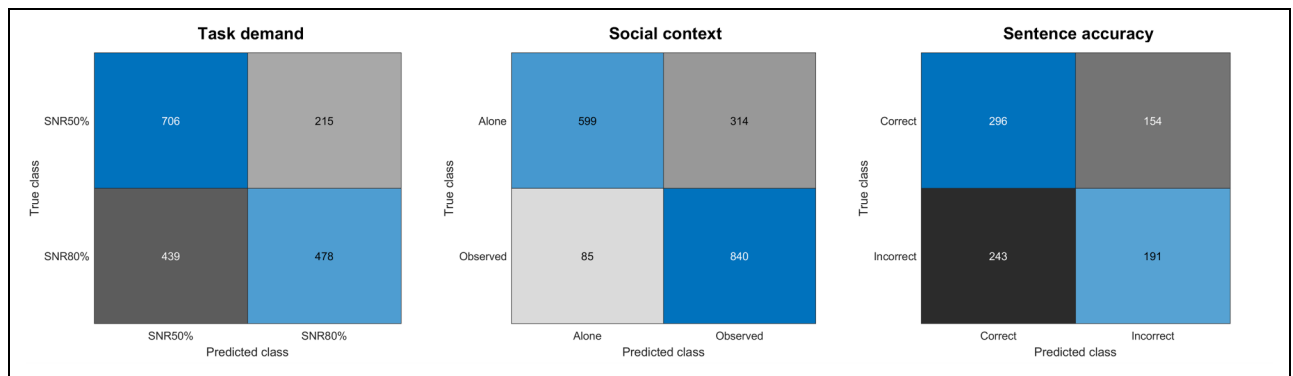
precision, sensitivity, specificity, and F1 scores) of the three classifiers are presented in Table 1. Confusion matrices are presented in Figure 3. Classification accuracy scores of 64.4%, 78.3%, and 55.1% were obtained for predicting task demand, social context, and sentence accuracy, respectively. These accuracy scores are demonstrated in Figure 4, denoted by vertical dashed lines. Figure 4 also demonstrates the distribution of accuracy scores of 100 classifiers with permuted labels (i.e., the distribution of chance level). The permutation analyses revealed that all three classifiers were operating at above chance accuracy levels ( $p < .01$  for task demand and social context,  $p = .04$  for sentence accuracy).

**Feature Importance.** Figure 5 shows the feature weights obtained by NCA for each of the classifiers. The three pupil features contributed to each classifier to a similar degree. The features contributing most to the prediction of task demand were BVPA and PEP, whereas the other cardiovascular features (IBI and PAT) made a negligible contribution to the model. The features contributing most to the prediction of social context were PEP, PAT and to a lesser extent IBI. Finally, for sentence accuracy, BVPA contributed most to the model, though this was not hugely pronounced in relation to the other features.

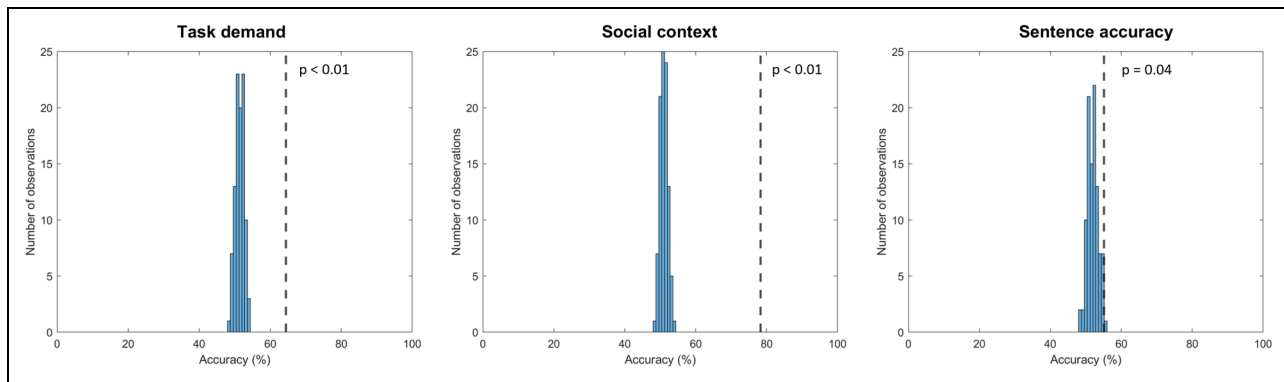
**Leave One Subject Out Cross Validation.** To ensure that the accuracy levels obtained by the classifiers were not a result

**Table 1.** Results of Group Level k-NN Classifiers Using 50-Fold Cross Validation.

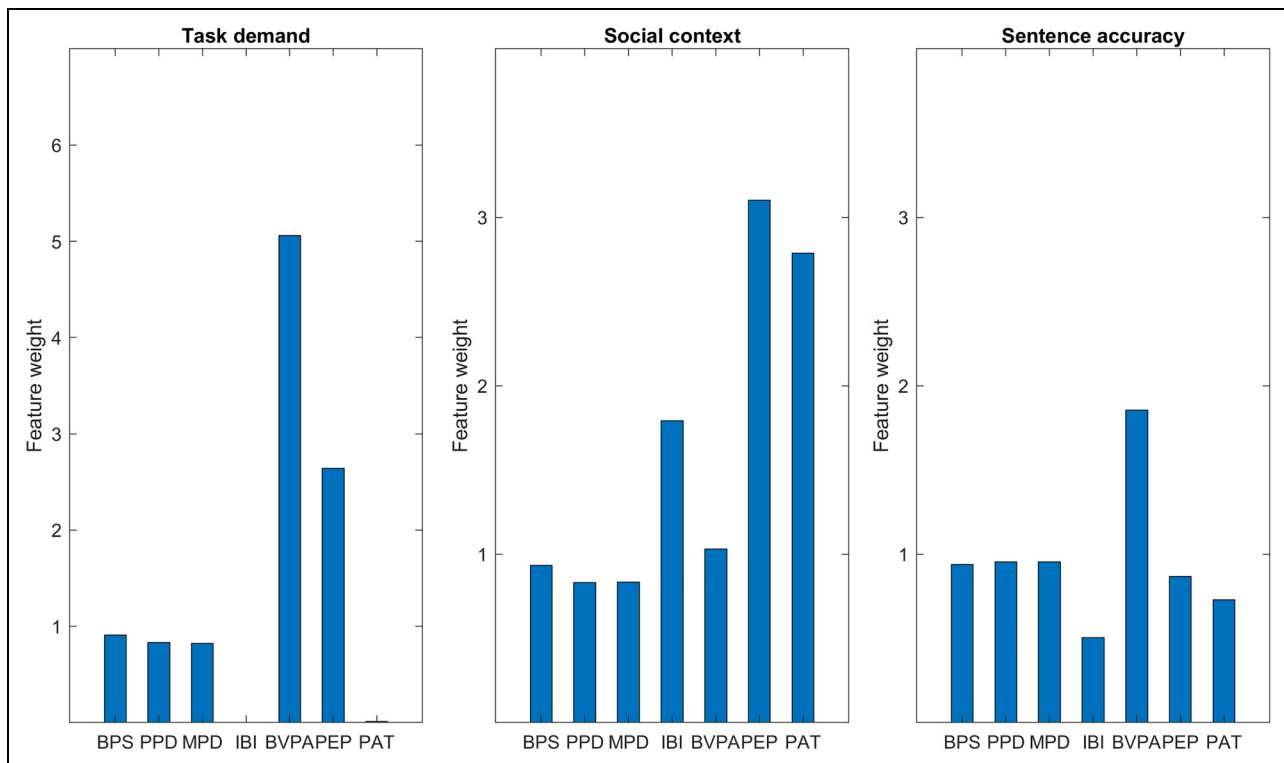
Model trained to predict	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1 score (%)	Number of neighbors	Distance metric
Task demand	64.4	61.7	76.7	52.1	68.3	1	City block
Social context	78.3	87.6	65.6	90.8	75.0	1	Standard Euclidean
Sentence accuracy	55.1	54.9	65.8	44.0	59.9	14	Spearman



**Figure 3.** Confusion matrices of 50-fold cross-validated group classifiers predicting task demand (left), social context (middle) and sentence accuracy (right). The confusion matrices have the following structure: the top left quadrant reflects true positive, the top right quadrant reflects false negative, the bottom left quadrant reflects false positive, and the bottom right quadrant reflects true negative. The shading of the quadrants corresponds to the proportion of the total number of trials; dark shading indicates a high proportion of trials in that quadrant, and light shading indicates a low proportion of trials in that quadrant.



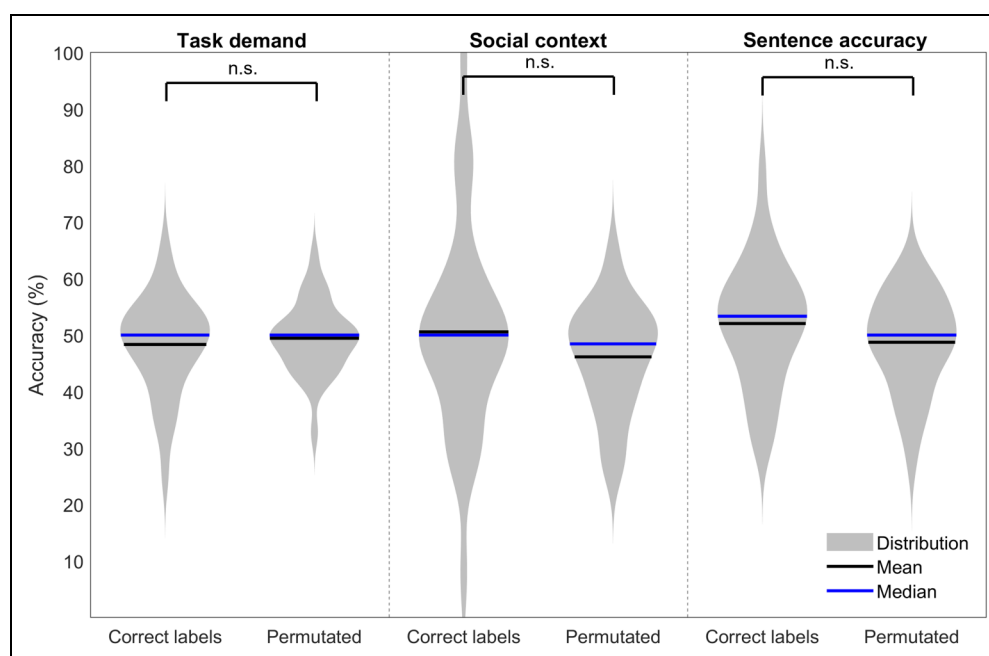
**Figure 4.** Histograms of permutation accuracy distributions for 50-fold cross-validated group classifiers predicting task demand (left), social context (middle), and sentence accuracy (right). The dashed line represents the performance of the model with correct labels. 100 permutations were conducted.



**Figure 5.** Feature weights obtained by NCA for the 50-fold cross-validated group classifiers predicting task demand (left), social context (middle), and sentence accuracy (right). A feature weight of zero demonstrates that the feature does not contribute, whereas features with a higher weight contribute more. Please note that the left-hand panel has a different y axis scale than the middle and righthand panels. BPS = baseline pupil size; BVPA = blood volume pulse amplitude; IBI = interbeat interval; MPD = mean pupil dilation; NCA = neighborhood component analysis; PAT = pulse arrival time; PEP = pre-ejection period; PPD = peak pupil dilation.

of bias in the classifier caused by k-fold cross validation, we conducted leave one subject out cross validation, where all trials from one participant were fully excluded from the training dataset and used exclusively for testing. We repeated this process systematically such that data from all participants were the test dataset on one occasion. Using this validation method, the average

performance of each of the classifiers dropped to around chance level (i.e., 50%): average accuracies were 48.3% (SD = 8.7), 50.6% (SD = 18.8), and 52.0% (SD = 11.0) for predicting task demand, social context, and sentence accuracy, respectively. Thus, it is evident that the k-fold cross validation procedure was inflating the performance accuracy of the classifiers, compared to



**Figure 6.** Violin plots demonstrating distributions of correct label and permuted label model accuracies for group classifiers validated with leave one subject out cross validation. Panels demonstrate task demand (left), social context (middle) and sentence accuracy (right). A single asterisk represents a significant difference between the correctly labeled and permutation accuracies at the level of  $p < .05$ . n.s. = not significant.

when testing and training were conducted on completely separate participants.

Permutation analyses were conducted on the classifiers validated using leave one subject out cross validation, to allow statistical comparison between the correctly labeled accuracies and permuted classifiers' accuracies. Figure 6 demonstrates the distributions of accuracy levels for both correctly labeled and permuted classifiers. Means and standard deviations for the permuted classifiers were as follows: 49.4% (SD = 6.3), 46.1% (SD = 9.6), and 48.7% (SD = 8.7) for predicting task demand, social context, and sentence accuracy, respectively. Levene's test revealed equal variances between the correct and permuted classifiers' accuracy for all three labels. Independent t-tests revealed no significant differences between the correct and permuted classifiers' accuracy levels for task demand ( $t(56) = -0.56$ ,  $p = .58$ ,  $d = -0.15$ ), social context ( $t(56) = 1.13$ ,  $p = .26$ ,  $d = 0.30$ ), or sentence accuracy ( $t(56) = 1.28$ ,  $p = .21$ ,  $d = 0.36$ ). This suggests that the correctly labeled classifiers have no predictive power.

### Individual Classifiers

Training separate classifiers for each participant allows classifiers to be personalized or calibrated to the individual. Five participants were excluded from the individual classification analysis due to missing data for one or more of the features. The results of the remaining 24 participants' individual

classifiers are presented in Tables 2, 3, and 4. Generally, the accuracy was higher for individual classifiers than for the group classification, which was expected. Average accuracy levels were 75.7% (SD = 9.9) for task demand, 89.7% (SD = 12.6) for social context and 68.2% (SD = 10.8) for sentence accuracy. Similar to the trend demonstrated by the k-fold cross-validated group level classifiers, average accuracy values for the individual classifiers were higher when predicting social context, compared to the task demand and sentence accuracy.

Permutation analysis was conducted to verify whether the individual classifiers were operating at above chance accuracy levels (demonstrated by asterisks ( $p < .01$ ) and plus signs ( $p < .05$ ) in Figure 7). The classifiers for 15, 21, or two individuals performed above chance in predicting task demand, social context, and sentence accuracy, respectively. For the social context classifiers especially, the high-performance suggests there is a true relationship between the social context and the physiological data in most individuals. In fact, two of the individual classifiers were able to predict social context with an accuracy of 100%. This suggests that the trained classifiers for these two participants were perfectly able to distinguish between the data in the alone and observed conditions. In contrast, one of the classifiers predicting sentence accuracy never predicted a positive outcome (correct response) and therefore had no sensitivity.

The results of NCA for the individual classifiers are demonstrated in Figure 8. Similar to the pattern observed for the

**Table 2.** Average Results of 29 Group k-NN Classifiers Using Leave One Subject Out Cross Validation (Separate Test and Training Participants).

	Average scores (%)				
	Accuracy (SD)	Precision (SD)	Sensitivity (SD)	Specificity (SD)	F1 score (SD)
<b>Correct labels</b>					
<i>Task demand</i>	48.3 (8.7)	48.1 (11.5)	52.5 (24.8)	44.0 (25.3)	48.1 (15.0)
<i>Social context</i>	50.6 (18.8)	46.9 (29.0)	35.4 (36.5)	66.5 (31.3)	33.7 (31.0)
<i>Sentence accuracy</i>	52.0 (11.0)	53.9 (19.0)	48.2 (31.3)	55.2 (28.6)	44.9 (25.2)
<b>Permuted labels</b>					
<i>Task demand</i>	49.4 (6.3)	49.7 (5.6)	69.8 (22.9)	28.9 (23.0)	56.4 (11.2)
<i>Social context</i>	46.1 (9.6)	45.9 (21.8)	44.5 (32.3)	48.6 (32.7)	39.1 (22.1)
<i>Sentence accuracy</i>	48.7 (8.7)	54.4 (20.4)	35.4 (28.8)	64.4 (29.5)	35.1 (23.3)

**Table 3.** Results of Individual k-NNs Predicting Task Demand.

ID	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1 score	Number of neighbors	Distance metric	Feature with highest weight
1	82.8	83.9	81.3	84.4	82.5	1	'cityblock'	'BVPA'
2	71.9	70.6	75.0	68.8	72.7	1	'correlation'	'BVPA'
3	78.1	75.0	84.4	71.9	79.4	1	'cosine'	'BVPA'
4	76.6	74.3	81.3	71.9	77.6	11	'cityblock'	'PAT'
5	82.8	86.2	78.1	87.5	82.0	4	'correlation'	'IBI'
6	65.6	65.6	65.6	65.6	65.6	2	'seuclidean'	'IBI'
7	59.4	56.5	81.3	37.5	66.7	2	'cosine'	'BVPA'
8	67.2	62.2	87.5	46.9	72.7	3	'cosine'	'BVPA'
9	73.4	74.2	71.9	75.0	73.0	1	'euclidean'	'PAT'
10	67.2	65.7	71.9	62.5	68.7	1	'cosine'	'BVPA'
11	73.4	65.3	100.0	46.9	79.0	1	'cosine'	'IBI'
12	62.5	61.8	65.6	59.4	63.6	32	'spearman'	'IBI'
13	89.1	90.3	87.5	90.6	88.9	7	'euclidean'	'BVPA'
14	76.6	79.3	71.9	81.3	75.4	1	'cosine'	'PAT'
15	54.7	53.2	78.1	31.3	63.3	1	'cityblock'	'BVPA'
16	59.4	56.0	87.5	31.3	68.3	1	'mahalanobis'	'BVPA'
17	59.4	58.3	65.6	53.1	61.8	9	'euclidean'	'PAT'
18	75.0	71.1	84.4	65.6	77.1	1	'cityblock'	'PAT'
19	89.1	85.7	93.8	84.4	89.6	4	'cosine'	'BVPA'
20	89.1	93.1	84.4	93.8	88.5	6	'seuclidean'	'BVPA'
21	73.4	74.2	71.9	75.0	73.0	1	'cityblock'	'BVPA'
22	62.5	60.0	75.0	50.0	66.7	5	'chebychev'	'BVPA'
23	59.4	57.1	75.0	43.8	64.9	6	'cityblock'	'BVPA'
24	73.4	68.3	87.5	59.4	76.7	1	'euclidean'	'BVPA'

BVPA = blood volume pulse amplitude; IBI = interbeat interval; PAT = pulse arrival time.

group level classifiers, the pupil features appear to have a small, but consistent contribution to the individual classifiers. Of the cardiovascular features, BVPA has a notable contribution to each of the different classifiers. This is also demonstrated in Tables 3, 4, and 5: BVPA was most frequently the greatest contributor to the classifiers. The other cardiovascular features demonstrate more variability in their contributions. For example, IBI was an important feature for the social context classifiers yet provided only a small contribution to the task demand classifiers.

## Discussion

The current study combined pupil and cardiovascular features from individuals with hearing loss to predict task demand, social context, and sentence accuracy during a speech perception task. In our view, this study boasts three novel aspects. The first pertains to the inclusion of group and individual classifiers, the use of which are relatively unexplored in the listening effort literature, the second, to the incorporation of a social context manipulation in our

**Table 4.** Results of Individual k-NNs Predicting Social Context.

ID	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1 score	Number of neighbors	Distance metric	Feature with highest weight
1	95.3	93.9	96.9	93.8	95.4	1	'mahalanobis'	'PAT'
2	98.4	97.0	100.0	96.9	98.5	1	'seuclidean'	'IBI'
3	100.0	100.0	100.0	100.0	100.0	3	'seuclidean'	'PAT'
4	95.3	93.9	96.9	93.8	95.4	4	'cityblock'	'IBI'
5	98.4	97.0	100.0	96.9	98.5	1	'correlation'	'IBI'
6	96.9	100.0	93.8	100.0	96.8	1	'minkowski'	'IBI'
7	50.0	50.0	100.0	0.0	66.7	2	'seuclidean'	'PPD'
8	93.8	88.9	100.0	87.5	94.1	1	'euclidean'	'BVPA'
9	89.1	90.3	87.5	90.6	88.9	32	'cityblock'	'BVPA'
10	95.3	91.4	100.0	90.6	95.5	4	'euclidean'	'BVPA'
11	84.4	76.2	100.0	68.8	86.5	7	'cityblock'	'IBI'
12	84.4	86.7	81.3	87.5	83.9	32	'spearman'	'BVPA'
13	90.6	84.2	100.0	81.3	91.4	3	'cosine'	'BVPA'
14	70.3	74.1	62.5	78.1	67.8	4	'euclidean'	'BVPA'
15	89.1	90.3	87.5	90.6	88.9	3	'euclidean'	'BVPA'
16	95.3	91.4	100.0	90.6	95.5	2	'seuclidean'	'BVPA'
17	64.1	69.6	50.0	78.1	58.2	5	'euclidean'	'IBI'
18	95.3	96.8	93.8	96.9	95.2	4	'euclidean'	'IBI'
19	62.5	59.1	81.3	43.8	68.4	10	'seuclidean'	'BVPA'
20	100.0	100.0	100.0	100.0	100.0	1	'chebychev'	'IBI'
21	90.6	90.6	90.6	90.6	90.6	1	'correlation'	'IBI'
22	95.3	100.0	90.6	100.0	95.1	5	'correlation'	'BVPA'
23	85.9	79.5	96.9	75.0	87.3	32	'spearman'	'IBI'
24	90.6	84.2	100.0	81.3	91.4	1	'seuclidean'	'IBI'

BVPA = blood volume pulse amplitude; IBI = interbeat interval; PAT = pulse arrival time; PPD = peak pupil dilation.

study design and the third, to the combination of pupil and cardiovascular features at the trial level.

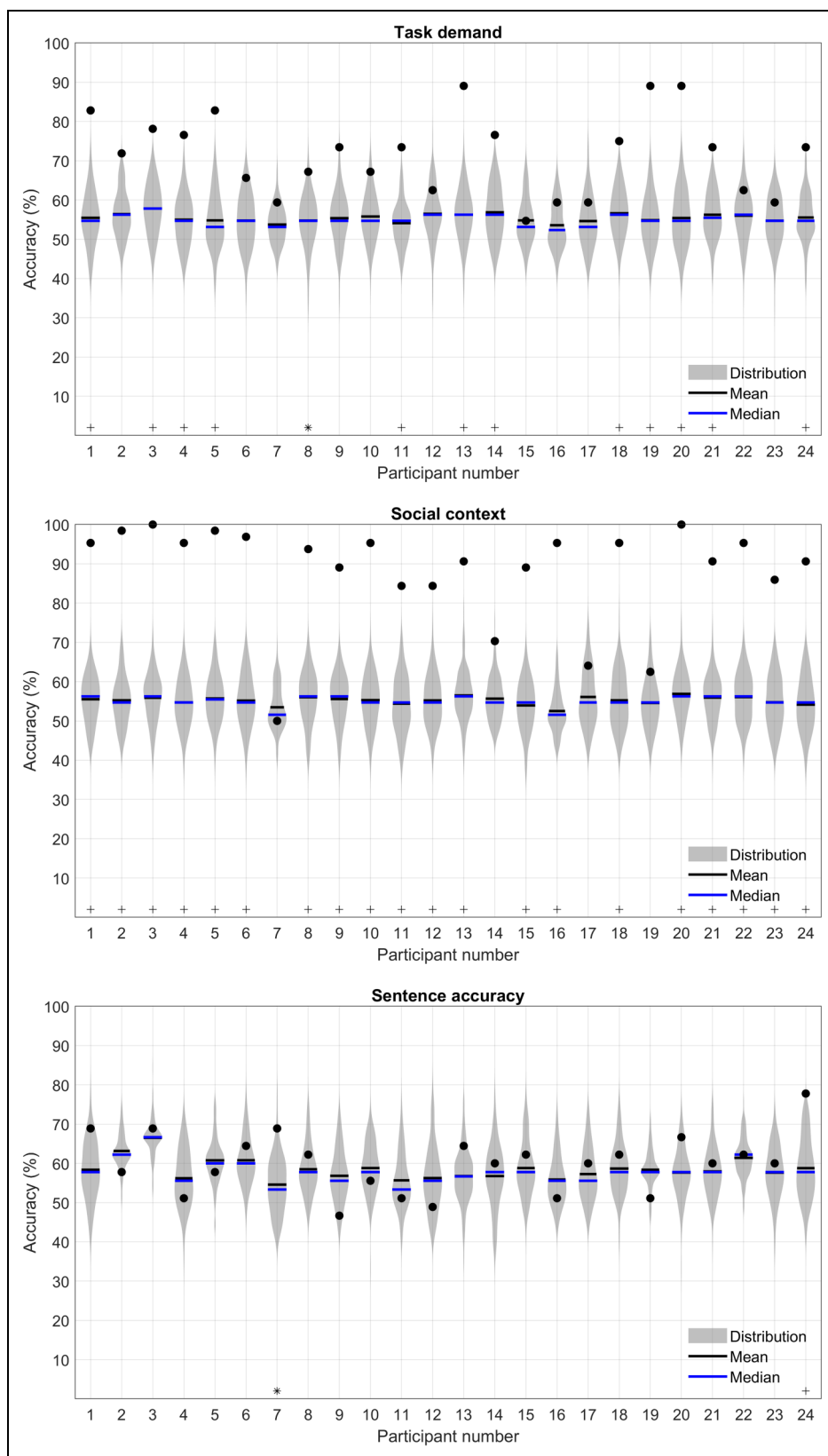
### Validation and Generalizability of Classifiers

An important finding from this work was the disparity between classifier performance depending on whether the classifiers were trained on group data or individual data. At the group level, k-fold cross validation (where data from the same participant may have occurred in both the training and test datasets) resulted in higher performance, whereas leave one subject out cross validation (where data from one participant was held out from training and used exclusively for testing) resulted in lower performance. This pattern suggests that the group classifiers generalized poorly to novel participants' data, a finding that is attributable to the nature of k-NN classifiers. The k-NN classifiers assign the label of a new data point based upon the neighbors nearest to it (in this case, a single nearest neighbor for the task demand and social context classifiers, and 14 nearest neighbors for the sentence accuracy classifier). Therefore, if the classifier has previously been exposed to data from a participant during training, it is better able to assign the same label to a similar data point during testing. Lower classification accuracies may also have resulted from sparsely sampled data in the classes.

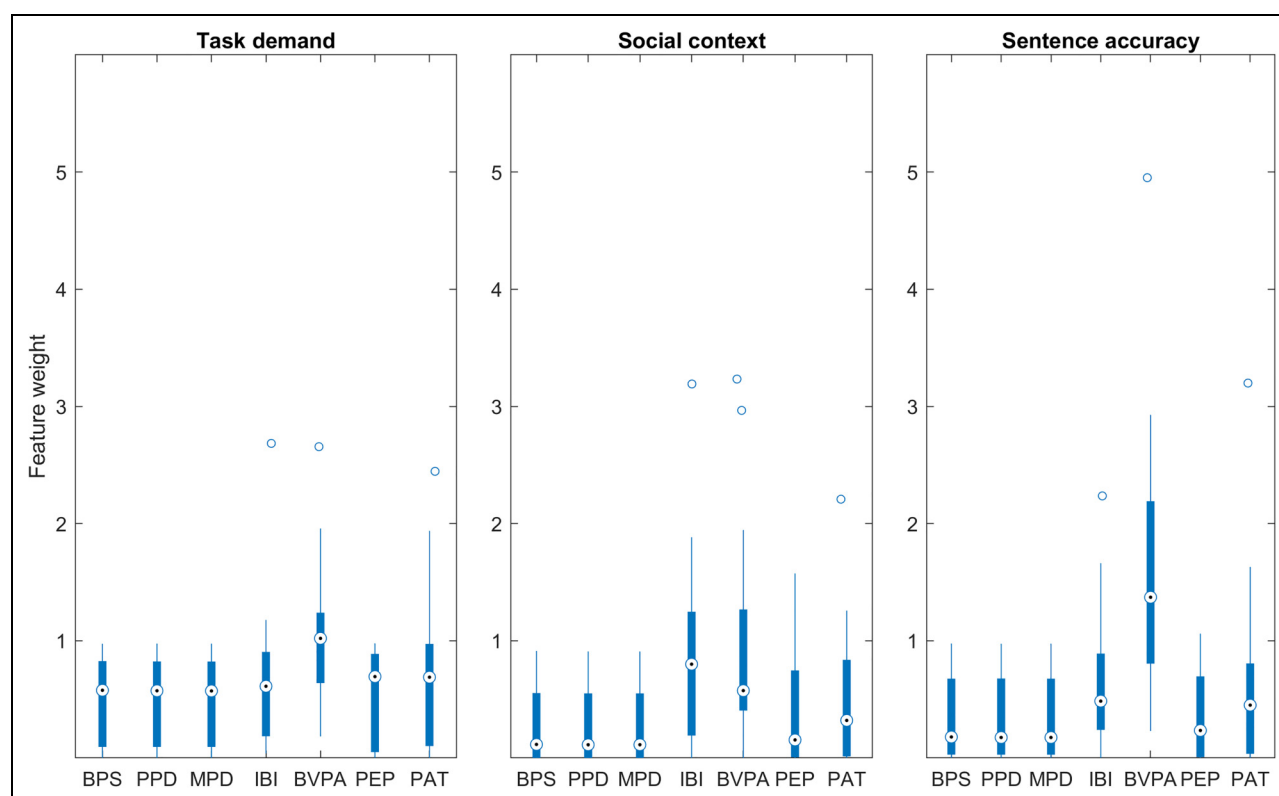
The poor generalizability of these group level classifiers suggests that the association between the to-be predicted variable and the physiological response differed between individuals. When classifiers were trained on the individual participants' data (i.e., within participant classifiers), performance improved (particularly for predicting task demand and social context). This suggests that within-subject variability was considerably smaller than between-subject variability. A similar pattern has also been demonstrated in other studies using physiological data to train classifiers to predict psychological states (Osotsi et al., 2020). The superior performance of individualized classifiers compared to group level classifiers ultimately has implications for future applications and technology incorporating these measures (discussed in more detail in the Future Applications section of the Discussion).

### Predicting Task Demand, Social Context, and Sentence Accuracy

The best performance levels were obtained by the k-NN classifiers trained to predict social context, that is, whether the trial reflected an alone or observed condition. The average performance of the individual classifiers was high (89.7%, SD = 12.6) and for all but three individuals, these classifiers were able to predict social context at an above chance



**Figure 7.** Violin plots demonstrating permutation accuracy distributions for individual classifiers. The upper panel reflects task demand accuracies, the middle panel reflects social context and the lower panel, sentence accuracy. Black filled dots represent the model accuracy using correct labeling (i.e., the original individual classifiers). Significant deviations of the correctly labeled classifier from the null distribution are demonstrated by asterisks at  $p < .05$  and plus signs at  $p < .01$ .



**Figure 8.** Box plots demonstrating feature weights obtained by NCA for individual classifiers trained to predict task demand (left), social context (middle), and sentence accuracy (right). BPS = baseline pupil size; BVPA = blood volume pulse amplitude; IBI = interbeat interval; MPD = mean pupil dilation; NCA = neighborhood component analysis; PAT = pulse arrival time; PEP = pre-ejection period; PPD = peak pupil dilation.

accuracy level. The next best performance levels were obtained by the classifiers predicting task demand, that is, whether the trial was presented at SNR50% or SNR80%. Average performance of the individual classifiers was 71.7% (SD = 10.2), with over half of the individual classifiers operating above chance level. Finally, the poorest performance was demonstrated by the sentence accuracy classifiers, which were trained to predict whether the trial was repeated correctly or incorrectly. The average individual classifier performance was 60.0% (SD = 13.1), and just two out of 24 of the individual classifiers were operating at an above chance level.

The timescale of the social context manipulation compared to that of task demand and sentence accuracy may have contributed to the superior classification performance: the observers were present continually throughout the whole block of sentences, which one might expect to produce a consistent physiological response. Whereas the task demand and sentence accuracy timescales were comparatively fluctuating in nature: the stimulus presentation itself only occurred for a short period (3 s of noise, then target sentence presentation plus noise, followed by 3 s of noise), before the participants repeated back what they heard (thus having a chance to score correctly or incorrectly).

Importantly, during the data window selected for our analysis, the participant had yet to repeat the sentence and may therefore not have been aware that they would make an error in sentence repetition. Also, physiological responses caused by sentence accuracy may have impacted the period after the window selected for analysis. For instance, Spruit et al. (2018) have shown decreases in IBI and PEP posterror compared to postcorrect Flanker and switch trials (Spruit et al., 2018). Indeed, the sentence accuracy may have even affected subsequent trials, rather than the present trial. This likely contributed to the relatively poor performance of the classifiers predicting sentence accuracy, compared to the other two classifiers.

Another contributing factor may be that for the purposes of this analysis we considered sentence accuracy as binary in nature (i.e., sentence repeated completely correctly or not), yet, in reality it is more continuous. We also assumed that sentence accuracy errors in the SNR50% condition and SNR80% conditions produced a similar response, where they may not have done. A final, potentially relevant factor is the smaller dataset used to predict sentence accuracy compared to the datasets used to predict social context and task demand.

**Table 5.** Results of Individual k-NNs Predicting Sentence Accuracy.

ID	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1 score	Number of neighbors	Distance metric	Feature with highest weight
1	68.9	61.5	80.0	60.0	69.6	20	'mahalanobis'	'BVPA'
2	57.8	25.0	5.9	89.3	9.5	10	'minkowski'	'BVPA'
3	68.9	100.0	6.7	100.0	12.5	2	'jaccard'	'BVPA'
4	51.1	40.0	9.5	87.5	15.4	23	'correlation'	'PAT'
5	57.8	N/A*	N/A*	96.3	N/A*	22	'correlation'	'BVPA'
6	64.4	56.3	50.0	74.1	52.9	2	'cityblock'	'PAT'
7	68.9	64.5	87.0	50.0	74.1	1	'mahalanobis'	'BVPA'
8	62.2	64.3	72.0	50.0	67.9	23	'cityblock'	'BVPA'
9	46.7	50.0	41.7	52.4	45.5	1	'cityblock'	'BVPA'
10	55.6	57.5	88.5	10.5	69.7	19	'correlation'	'BVPA'
11	51.1	53.1	70.8	28.6	60.7	8	'mahalanobis'	'IBI'
12	48.9	47.1	36.4	60.9	41.0	3	'jaccard'	'BVPA'
13	64.4	65.0	59.1	69.6	61.9	1	'spearman'	'BVPA'
14	60.0	57.5	95.8	19.0	71.9	6	'mahalanobis'	'BVPA'
15	62.2	59.5	100.0	15.0	74.6	2	'hamming'	'PAT'
16	51.1	51.9	60.9	40.9	56.0	2	'jaccard'	'BVPA'
17	60.0	57.5	95.8	19.0	71.9	11	'mahalanobis'	'BVPA'
18	62.2	61.8	84.0	35.0	71.2	8	'mahalanobis'	'BVPA'
19	51.1	56.7	65.4	31.6	60.7	1	'chebychev'	'BVPA'
20	66.7	69.2	72.0	60.0	70.6	12	'correlation'	'BVPA'
21	60.0	56.3	45.0	72.0	50.0	1	'mahalanobis'	'BVPA'
22	62.2	62.2	100.0	0.0	76.7	17	'hamming'	'BVPA'
23	60.0	62.5	25.0	88.0	35.7	10	'seuclidean'	'BVPA'
24	77.8	82.4	66.7	87.5	73.7	2	'chebychev'	'BVPA'

\*Note that precision, sensitivity, and F1 score are N/A for one participant because the model did not correctly predict any true cases (i.e., there were no true positives). BVPA = blood volume pulse amplitude; IBI = interbeat interval; N/A = not applicable; PAT = pulse arrival time.

### Feature Importance

The seven features included in our classifiers differ in their autonomic nervous system origins. Pupil size, IBI and PAT are thought to reflect mixed contributions from both the SNS and PNS system branches, whereas PEP and BVPA reflect mostly SNS activity (Ahmed et al., 1972; Czarnek et al., 2021; Iani et al., 2004; Malik et al., 1996; Newlin & Levenson, 1979; Nitzan et al., 1998; Zekveld et al., 2018). The features contributing most to our classifiers differed depending on which response variable was being predicted. The strongest contributors to the group level task demand classifier were BVPA, followed by PEP (both SNS), with no contribution from IBI and PAT (mixed). The importance of BVPA in this classifier corroborates the finding of Francis et al. (2016) who found that significant changes in BVPA, but not IBI, were elicited by varying the acoustic parameters (presenting two unmasked conditions, natural speech or synthetic speech, and two masked conditions, speech-shaped noise or two-talker babble). BVPA was also an important feature in the individual classifiers, contributing to classifiers predicting task demand, social context, and sentence accuracy (see Figure 8 and Tables 2, 3, and 4). The importance of this feature may be due to its SNS origins.

The strongest contributors to the social context classifier at the group level were PEP (SNS), PAT (mixed) and IBI (mixed

origin). This pattern was not directly reflected in the individual classifiers. Instead, BVPA (SNS) and IBI (mixed origin) were the most prominent features. Finally, the strongest contributor to the sentence accuracy classifier at the group and individual level was BVPA (SNS), with all other features contributing to a lesser and similar degree. The differences between the group level and individual level NCA reinforces the need for individual classifiers: a one size fits all approach is unsuitable as different people demonstrate different associations between the predicted variables and physiological responses during the same experiment.

When reviewing the classifiers' feature weights (Figures 5 and 8), one or more of the cardiovascular features generally outperformed the pupil features. The aforementioned variability in autonomic origins of the cardiovascular features may be a contributing factor. It is also likely that the pupil features contain more redundant information because they are closely related to one another, for example, the PPD and MPD both reflect the dilation of the pupil and are both normalized to the BPS (Winn et al., 2018). Correlation between features may have negatively impacted the performance of the classifiers, by increasing redundancy, computational cost and potentially affecting the distance metric (Alin, 2010; Hasan et al., 2021; Yigit, 2013). On the other hand, the cardiovascular measures are not baseline corrected and

reflect responses from different physiological systems—IBI and PEP are time intervals extracted from the heart cycle, PAT reflects a time interval that depends upon the heart cycle and the vasculature and finally, BVPA is an amplitude measure that reflects vascular changes. Though the range of the cardiovascular measures is likely to be greater than that of the pupil metrics, we applied standardization during the classification procedure to minimize this effect.

The predictive capacity of the pupil features may also have been reduced by a potential confound due to the timing of blinks with respect to sentence presentation. During preprocessing, it was confirmed that the frequency of blinks did not vary between experimental conditions. However, the timing of blinks with respect to sentence presentation was not explicitly explored during the analysis. The timing of blinks is important to consider as it has been demonstrated that blinks can have an impact on the pupil size during listening (Holtze et al., n.d.; Knapen et al., 2016; Yoo et al., 2021). Though a deeper analysis of the timing of blinks was beyond the scope of this work, it could be important to consider in future work.

To further explore if there was any benefit of including both cardiovascular and pupil features together, we trained k-NN classifiers (k-fold cross validation, method 1a in Figure 2) using the cardiovascular and pupil features separately. The accuracy obtained when including only the four cardiovascular features was similar to the original classifiers: 66.2% for predicting task demand, 82.3% for social context, 53.9% for sentence accuracy. In comparison, the accuracy obtained when including only the three pupil features was lower: 54.4% for predicting task demand, 57.6% for social context, and 53.9% for sentence accuracy. This suggests that there was no additional benefit of including the pupil measures in these classifiers.

## Future Applications

We have demonstrated that it is possible to combine different trial-level pupil and cardiovascular features to predict various aspects of a listening situation encountered in the laboratory. Of particular importance is the finding that individual classifiers demonstrated superior prediction abilities over group level classifiers. The real-life generalizability of our work is in part limited by the binary nature of the classifiers presented here. Indeed, in most situations, listening demand and performance are not “on” or “off” but instead operate on a scale. Similarly, social context is not a binary phenomenon in daily situations. Moreover, the present analyses considered each of these response variables independently of one another and did not account for potential interactions between them. Despite this, we believe this study provides a good first step in demonstrating that classification techniques may be useful in this context.

This study presents a step toward using physiological features to predict aspects of listening that are applicable to real

life, for example, different SNRs and social contexts. In the future, such a tool may prove beneficial in audiology clinics, in a diagnostic capacity or for testing hearing devices. It has been suggested that future ear level devices, like hearing aids, may even measure from the user’s physiology in situ (Goverdovsky et al., 2017). In this study, the predictive power of models trained on the individual was relatively high, which is promising for use in personalized hearing devices. With technology moving in such a direction, an important take home message from this study is that the physiological response to social context was predicted with higher accuracy levels than the task demand or sentence accuracy level. More work is needed to select which feature combinations are optimal, to reduce redundancy in the classifiers and ultimately find features that are better able to predict task demand and sentence accuracy.

## Acknowledgments

The authors are very grateful to Lorenz Fiedler and Tirdad Seifi Ala for interesting and helpful discussions regarding the analysis.

## Declaration of Conflicting Interests


The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


## Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: GS received support from NIHR Manchester Biomedical Research Centre, and the project received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie-Sklodowska-Curie grant agreement No 765329.

## ORCID iDs

Bethany Plain  <https://orcid.org/0000-0002-0852-4656>

Sophia E. Kramer  <https://orcid.org/0000-0002-0451-8179>

Adriana A. Zekveld  <https://orcid.org/0000-0003-1320-6908>

## References

- Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with scholastic aptitude test scores. *Science*, 205(4412), 1289–1292. <https://doi.org/10.1126/SCIENCE.472746>
- Ahmed, S. S., Levinson, G. E., Schwartz, C. J., & Ettinger, P. O. (1972). Systolic time intervals as measures of the contractile state of the left ventricular myocardium in man. *Circulation*, 46(3), 559–571. <https://doi.org/10.1161/01.CIR.46.3.559>
- Ala, T. S., Graversen, C., Wendt, D., Alickovic, E., Whitmer, W. M., & 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. <https://doi.org/10.1371/JOURNAL.PONE.0235782>
- Alhanbali, S., Dawes, P., Lloyd, S., & Munro, K. J. (2017). Self-reported listening-related effort and fatigue in hearing-impaired adults. *Ear and Hearing*, 38(1), e39–e48. <https://doi.org/10.1097/AUD.0000000000000361>

- Alhanbali, S., Dawes, P., Millman, R. E., & Munro, K. J. (2019). Measures of listening effort are multidimensional. *Ear and Hearing, 40*(5), 1084–1097. <https://doi.org/10.1097/AUD.0000000000000697>
- Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(3), 370–374. <https://doi.org/10.1002/WICS.84>
- Allen, A. P., Kennedy, P. J., Cryan, J. F., Dinan, T. G., & Clarke, G. (2014). Biological and psychological markers of stress in humans: Focus on the Trier Social Stress Test. *Neuroscience and Biobehavioral Reviews, 38*, 94–124. <https://doi.org/10.1016/j.neubiorev.2013.11.005>
- Anderson, M. J., & Ter Braak, C. J. F. (2010). Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation, 73*(2), 85–113. <https://doi.org/10.1080/00949650215733>
- Armañac-Julián, P., Kontaxis, S., Rapalis, A., Marozas, V., Laguna, P., Bailón, R., Gil, E., & Lázaro, J. (2022). Reliability of pulse photoplethysmography sensors: Coverage using different setups and body locations. *Frontiers in Electronics, 3*, 906324. <https://doi.org/10.3389/FELEC.2022.906324>
- Awad, A. A., Ghobashy, M. A. M., Ouda, W., Stout, R. G., Silverman, D. G., & Shelley, K. H. (2001). Different responses of ear and finger pulse oximeter wave form to cold pressor test. *Anesthesia and Analgesia, 92*(6), 1483–1486. <https://doi.org/10.1097/00000539-200106000-00026>
- Ayasse, N. D., & Wingfield, A. (2020). Anticipatory baseline pupil diameter is sensitive to differences in hearing thresholds. *Frontiers in Psychology, 10*, 2947. <https://doi.org/10.3389/FPSYG.2019.02947/BIBTEX>
- Babiker, A., Faye, I., Prehn, K., & Malik, A. (2015). Machine learning to differentiate between positive and negative emotions using pupil diameter. *Frontiers in Psychology, 6*(DEC). <https://doi.org/10.3389/FPSYG.2015.01921>
- Berger, H. (1929). Über das elektroencephalogramm des menschen. *Archiv Für Psychiatrie Und Nervenkrankheiten, 87*(1), 527–570. <https://doi.org/10.1007/BF01797193>
- Block, R. C., Yavarmanesh, M., Natarajan, K., Carek, A., Mousavi, A., Chandrasekhar, A., Kim, C. S., Zhu, J., Schifitto, G., Mestha, L. K., Inan, O. T., Hahn, J. O., & Mukkamala, R. (2020). Conventional pulse transit times as markers of blood pressure changes in humans. *Scientific Reports, 10*(1), 1–9. <https://doi.org/10.1038/s41598-020-73143-8>
- Bosch, J. A., De Geus, E. J. C., Carroll, D., Goedhart, A. D., Anane, L. A., Veldhuizen Van Zanten, J. J., Helmerhorst, E. J., & Edwards, K. M. (2009). A general enhancement of autonomic and cortisol responses during social evaluative threat. *Psychosomatic Medicine, 71*(8), 877–885. <https://doi.org/10.1097/PSY.0b013e3181baef05>
- Bott, A., & Saunders, G. (2021). A scoping review of studies investigating hearing loss, social isolation and/or loneliness in adults. *International Journal of Audiology, 60*(S2), 30–46. [https://doi.org/10.1080/14992027.2021.1915506/SUPPL\\_FILE/IJJA\\_A\\_1915506\\_SM3242.PDF](https://doi.org/10.1080/14992027.2021.1915506/SUPPL_FILE/IJJA_A_1915506_SM3242.PDF)
- Caduff, A., Feldman, Y., Ben Ishai, P., & Launer, S. (2020). Physiological monitoring and hearing loss: Toward a more integrated and ecologically validated health mapping. *Ear and Hearing, 41*, 120S–130S. <https://doi.org/10.1097/AUD.0000000000000960>
- Canlon, B., Theorell, T., & Hasson, D. (2013). Associations between stress and hearing problems in humans. *Hearing Research, 295*, 9–15. <https://doi.org/10.1016/J.HEARES.2012.08.015>
- Chan, G., Cooper, R., Hosanee, M., Welykholowa, K., Kyriacou, P. A., Zheng, D., Allen, J., Abbott, D., Lovell, N. H., Fletcher, R., & Elgendi, M. (2019). Multi-site photoplethysmography technology for blood pressure assessment: Challenges and recommendations. *Journal of Clinical Medicine, 8*(11), 1827. <https://doi.org/10.3390/JCM8111827>
- Czarnek, G., Richter, M., & Strojny, P. (2021). Cardiac sympathetic activity during recovery as an indicator of sympathetic activity during task performance. *Psychophysiology, 58*(2), e13724. <https://doi.org/10.1111/PSYP.13724>
- Das, W., & Khanna, S. (2021). A robust machine learning based framework for the automated detection of ADHD using pupilometric biomarkers and time series analysis. *Scientific Reports, 11*(1), 1–12. <https://doi.org/10.1038/s41598-021-95673-5>
- Diamond, L., & Otter-Henderson, K. (2007). Physiological measures. In R. W. Robins, C. R. Fraley, R. F. Krueger, L. Aiken, & M. Ashton (Eds.), *Handbook of research methods in personality psychology* ((1st ed., pp. 370–388). Guilford Press. <https://psycnet.apa.org/record/2007-11524-022>
- Drummond, C. (2010). Classification. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 171). Springer US. [https://doi.org/10.1007/978-0-387-30164-8\\_111](https://doi.org/10.1007/978-0-387-30164-8_111)
- Escobar-Restrepo, B., Torres-Villa, R., Kyriacou, P. A., Zheng, D., Chen, F., Mukkamala, R., & Pan, F. (2018). Evaluation of the linear relationship between pulse arrival time and blood pressure in ICU patients: Potential and limitations. *Frontiers in Physiology, 9*, 1848. <https://doi.org/10.3389/FPHYS.2018.01848/BIBTEX>
- Fiedler, L., Wöstmann, M., Herbst, S. K., & Obleser, J. (2019). Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *NeuroImage, 186*, 33–42. <https://doi.org/10.1016/J.NEUROIMAGE.2018.10.057>
- Fleischhauer, V., Bruhn, J., Rasche, S., & Zaunseder, S. (2023). Photoplethysmography upon cold stress—impact of measurement site and acquisition mode. *Frontiers in Physiology, 14*, 1127624. <https://doi.org/10.3389/FPHYS.2023.1127624/BIBTEX>
- Francis, A. L., MacPherson, M. K., Chandrasekaran, B., & Alvar, A. M. (2016). Autonomic nervous system responses during perception of masked speech may reflect constructs other than subjective listening effort. *Frontiers in Psychology, 7*(MAR), 263. <https://doi.org/10.3389/fpsyg.2016.00263>
- Gatehouse, S., & Akeroyd, M. (2009). Two-eared listening in dynamic situations. *Trends in Hearing, 45*(SUPPL. 1), 120–124. <https://doi.org/10.1080/14992020600783103>
- Gatehouse, S., & Akeroyd, M. (2009). Two-eared listening in dynamic situations. <http://dx.doi.org/10.1080/14992020600783103>, 45(SUPPL. 1), 120–124. <https://doi.org/10.1080/14992020600783103>
- Gholamiangonabadi, D., Kiselov, N., & Grolinger, K. (2020). Deep neural networks for human activity recognition with wearable sensors: Leave-one-subject-out cross-validation for model selection. *IEEE Access, 8*, 133982–133994. <https://doi.org/10.1109/ACCESS.2020.3010715>
- Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit,

- and PhysioNet. *Circulation*, 101(23). <https://doi.org/10.1161/01.CIR.101.23.E215>
- Golland, P., & Fischl, B. (2003). Permutation tests for classification: Towards statistical significance in image-based studies. Biennial International Conference on Information Processing in Medical Imaging.
- Gordan, R., Gwathmey, J. K., & Xie, L.-H. (2015). Autonomic and endocrine control of cardiovascular function. *World Journal of Cardiology*, 7(4), 204. <https://doi.org/10.4330/wjc.v7.i4.204>
- Goverdovsky, V., Von Rosenberg, W., Nakamura, T., Looney, D., Sharp, D. J., Papavassiliou, C., Morrell, M. J., & Mandic, D. P. (2017). Hearables: Multimodal physiological in-ear sensing. *Scientific Reports*, 7(1). <https://doi.org/10.1038/s41598-017-06925-2>
- Granholm, E., & Steinhauser, S. R. (2004). Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology*, 52(1), 1–6. <https://doi.org/10.1016/J.IPSYCHO.2003.12.001>
- Hartmann, V., Liu, H., Chen, F., Qiu, Q., Hughes, S., & Zheng, D. (2019). Quantitative comparison of photoplethysmographic waveform characteristics: Effect of measurement site. *Frontiers in Physiology*, 10(MAR), 431697. <https://doi.org/10.3389/FPHYS.2019.00198/BIBTEX>
- Hasan, M. J., Sohaib, M., & Kim, J. M. (2021). An explainable ai-based fault diagnosis model for bearings. *Sensors*, 21(12), 4070. <https://doi.org/10.3390/S21124070>
- Hasson, D., Theorell, T., Wallén, M. B., Leineweber, C., & Canlon, B. (2011). Stress and prevalence of hearing problems in the Swedish working population. *BMC Public Health*, 11(1), 1–12. <https://doi.org/10.1186/1471-2458-11-130>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.
- Hermans, B. J. M., Vink, A. S., Bennis, F. C., Filippini, L. H., Meijborg, V. M. F., Wilde, A. A. M., Pison, L., Postema, P. G., & Delhaas, T. (2017). The development and validation of an easy to use automatic QT-interval algorithm. *PLoS ONE*, 12, e0184352. <https://doi.org/10.1371/journal.pone.0184352>
- Hétu, R., Jones, L., & Getty, L. (1993). The impact of acquired hearing impairment on intimate relationships: Implications for rehabilitation. *International Journal of Audiology*, 32(6), 363–380. <https://doi.org/10.3109/00206099309071867>
- Hey, S., Walter, K., Gharbi, A., Von Haaren, B., König, N., & Löffler, S. (2009). *Continuous noninvasive pulse transit time measurement for psycho-physiological stress Monitoring*. <https://doi.org/10.1109/eTELEMED.2009.35>
- Holman, J. A., Drummond, A., Hughes, S. E., & Naylor, G. (2019). *International Journal of Audiology Hearing impairment and daily-life fatigue: A qualitative study Hearing impairment and daily-life fatigue: A qualitative study*. *International Journal of Audiology*, 58(7), 408–416. <https://doi.org/10.1080/14992027.2019.1597284>
- Holtze, B., Rosenkranz, M., Bleichner, M. G., Jaeger, M., & Debener, S. (n.d.). *Eye-blink patterns reflect attention to continuous speech*. <https://doi.org/10.31234/OSF.IO/N86YP>
- Hughes, S. E., Hutchings, H. A., Rapport, F. L., McMahon, C. M., & Boisvert, I. (2018). Social connectedness and perceived listening effort in adult cochlear implant users: A grounded theory to establish content validity for a new patient-reported outcome measure. *Ear and Hearing*, 39(5), 922–934. <https://doi.org/10.1097/AUD.0000000000000553>
- Iani, C., Gopher, D., & Lavie, P. (2004). Effects of task difficulty and invested mental effort on peripheral vasoconstriction. *Psychophysiology*, 41(5), 789–798. <https://doi.org/10.1111/J.1469-8986.2004.00200.X>
- Javaid, M., Yousaf, M., Sheikh, Q., Awais, M. M., Saleem, S., & Khalid, M. (2015). Real-time EEG-based human emotion recognition. *Springer*, 9492, 182–190. [https://doi.org/10.1007/978-3-319-26561-2\\_22](https://doi.org/10.1007/978-3-319-26561-2_22)
- Jennings, J. R., Kamarck, T., Stewart, C., Eddy, M., & Johnson, P. (1992). Alternate cardiovascular baseline assessment techniques: Vanilla or resting baseline. *Psychophysiology*, 29(6), 742–750. <https://doi.org/10.1111/j.1469-8986.1992.tb02052.x>
- Jessen, S., Fiedler, L., Münte, T. F., & Obleser, J. (2019). Quantifying the individual auditory and visual brain response in 7-month-old infants watching a brief cartoon movie. *NeuroImage*, 202, 116060. <https://doi.org/10.1016/j.neuroimage.2019.116060>
- Kahneman, D. (1973). *Attention and effort*. Prentice-Hall Inc.
- Keidser, G., Naylor, G., Brungart, D. S., Caduff, A., Campos, J., Carlile, S., Carpenter, M. G., Grimm, G., Hohmann, V., Holube, I., Launer, S., Lunner, T., Mehra, R., Rapport, F., Slaney, M., & Smeds, K. (2020). The quest for ecological validity in hearing science: What it is, why it matters, and how to advance it. *Ear & Hearing*, 41, 5S–19S. <https://doi.org/10.1097/AUD.0000000000000944>
- Knapen, T., De Gee, J. W., Brascamp, J., Nuiten, S., Hoppenbrouwers, S., & Theeuwes, J. (2016). Cognitive and ocular factors jointly determine pupil responses under equiluminance. *PLoS ONE*, 11(5), e0155574. <https://doi.org/10.1371/JOURNAL.PONE.0155574>
- Koelewijn, T., Zekveld, A. A., Lunner, T., & Kramer, S. E. (2021). The effect of monetary reward on listening effort and sentence recognition. *Hearing Research*, 406, 108255. <https://doi.org/10.1016/J.HEARES.2021.108255>
- Kuipers, M., Richter, M., Scheepers, D., Immink, M. A., Sjak-Shie, E., & van Steenbergen, H. (2017). How effortful is cognitive control? Insights from a novel method measuring single-trial evoked beta-adrenergic cardiac reactivity. *International Journal of Psychophysiology*, 119, 87–92. <https://doi.org/10.1016/j.ijpsycho.2016.10.007>
- Kushki, A., Fairley, J., Merja, S., King, G., & Chau, T. (2011). Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites. *Physiological Measurement*, 32(10), 1529–1539. <https://doi.org/10.1088/0967-3334/32/10/002>
- Liu, B., Zhang, Z., Di, X., Wang, X., Xie, L., Xie, W., & Zhang, J. (2021). The assessment of autonomic nervous system activity based on photoplethysmography in healthy young men. *Frontiers in Physiology*, 12, 1543. <https://doi.org/10.3389/FPHYS.2021.733264/BIBTEX>
- Mackersie, C. L., & Calderon-Moultrie, N. (2016). Autonomic nervous system reactivity during speech repetition tasks: Heart rate variability and skin conductance. *Ear and Hearing*, 37, 118S–125S. <https://doi.org/10.1097/AUD.0000000000000305>
- Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indexes of listening effort in a competing-talker task. *Journal of the American Academy of Audiology*, 22(2), 113–122. <https://doi.org/10.3766/jaaa.22.2.6>

- Mackersie, C. L., & Kearney, L. (2017). Autonomic nervous system responses to hearing-related demand and evaluative threat. *American Journal of Audiology*, 26(3S), 373–377. [https://doi.org/10.1044/2017\\_AJA-16-0133](https://doi.org/10.1044/2017_AJA-16-0133)
- Mackersie, C. L., Macphree, I. X., & Heldt, E. W. (2015). Effects of hearing loss on heart rate variability and skin conductance measured during sentence recognition in noise. *Ear and Hearing*, 36(1), 145–154. <https://doi.org/10.1097/AUD.0000000000000091>
- Malik, M., Camm, A. J., Bigger, J. T., Breithardt, G., Cerutti, S., Cohen, R. J., Coumel, P., Fallen, E. L., Kennedy, H. L., Kleiger, R. E., Lombardi, F., Malliani, A., Moss, A. J., Rottman, J. N., Schmidt, G., Schwartz, P. J., & Singer, D. H. (1996). Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 17(3), 354–381. <https://doi.org/10.1093/oxfordjournals.eurheartj.a014868>
- Matthen, M. (2016). Effort and displeasure in people who are hard of hearing. *Ear and Hearing*, 37, 28S–34S. <https://doi.org/10.1097/AUD.0000000000000292>
- McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group “white paper”. *International Journal of Audiology*, 53(7), 433–445. <https://doi.org/10.3109/14992027.2014.890296>
- McMahon, C. M., Boisvert, I., de Lissa, P., Granger, L., Ibrahim, R., Lo, C. Y., Miles, K., & Graham, P. L. (2016). Monitoring alpha oscillations and pupil dilation across a performance-intensity function. *Frontiers in Psychology*, 7(MAY), 745. <https://doi.org/10.3389/FPSYG.2016.00745/BIBTEX>
- Miltiadas, A., Tzamoura, K. D., Giannakeas, N., Tsiouras, M. G., Afrantou, T., Ioannidis, P., & Tzallas, A. T. (2021). Alzheimer’s disease and frontotemporal dementia: A robust classification method of EEG signals and a comparison of validation methods. *Diagnostics*, 11(8), 1437. <https://doi.org/10.3390/DIAGNOSTICS11081437>
- Mozos, O. M., Sandulescu, V., Andrews, S., Ellis, D., Bellotto, N., Dobrescu, R., & Ferrandez, J. M. (2017). Stress detection using wearable physiological and sociometric sensors. *International Journal of Neural Systems*, 27(2), 1650041. <https://doi.org/10.1142/S0129065716500416>
- Newlin, D. B., & Levenson, R. W. (1979). Pre-ejection period: Measuring beta-adrenergic influences upon the heart. *Psychophysiology*, 16(6), 546–552. <https://doi.org/10.1111/j.1469-8986.1979.tb01519.x>
- Nielsen, J. B., & Dau, T. (2011). The danish hearing in noise test. *International Journal of Audiology*, 50(3), 202–208. <https://doi.org/10.3109/14992027.2010.524254>
- Nitzan, M., Babchenko, A., Khanokh, B., & Landau, D. (1998). The variability of the photoplethysmographic signal—A potential method for the evaluation of the autonomic nervous system. *Physiological Measurement*, 19(1), 93–102. <https://doi.org/10.1088/0967-3334/19/1/008>
- Obleser, J., & Weisz, N. (2012). Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. *Cerebral Cortex*, 22(11), 2466–2477. <https://doi.org/10.1093/cercor/bhr325>
- Ohlenforst, B., Zekveld, A. A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., Versfeld, N. J., & Kramer, S. E. (2017). Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hearing Research*, 351, 68–79. <https://doi.org/10.1016/j.heares.2017.05.012>
- Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11, 1833–1863. <https://doi.org/10.1109/ICDM.2009.108>
- Osotsi, A., Oravec, Z., Li, Q., Smyth, J., & Brick, T. R. (2020). Individualized modeling to distinguish between high and low arousal states using physiological data. *Journal of Healthcare Informatics Research*, 4, 91–109. <https://doi.org/10.1007/s41666-019-00064-1>
- O’Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2017). Visual cortical entrainment to motion and categorical speech features during silent lipreading. *Frontiers in Human Neuroscience*, 10, 679. <https://doi.org/10.3389/FNHUM.2016.00679/BIBTEX>
- Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, 39(2), 204–214. <https://doi.org/10.1097/AUD.0000000000000494>
- Pichora-Fuller, K. (2016). How social psychological factors may modulate auditory and cognitive functioning during listening. *Ear and Hearing*, 37, 92S–100S. <https://doi.org/10.1097/AUD.0000000000000323>
- Pichora-Fuller, K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing*, 37, 5S–27S. <https://doi.org/10.1097/AUD.0000000000000312>
- Picou, E. M., Ricketts, T. A., & Hornsby, B. W. Y. (2013). How hearing aids, background noise, and visual cues influence objective listening effort. *Ear and Hearing*, 34(5), e52–e64. <https://doi.org/10.1097/AUD.0b013e31827f0431>
- Pielage, H., Plain, B. J., Saunders, G. H., Versfeld, N. J., Lunner, T., Kramer, S. E., & Zekveld, A. A. (2023). Copresence was found to be related to some pupil measures in persons with hearing loss while they performed a speech-in-noise task. *Ear and Hearing*, 44(5), 1190–1201. [https://journals.lww.com/ear-hearing/Fulltext/9900/Copresence\\_Was\\_Found\\_to\\_Be\\_Related\\_to\\_Some\\_Pupil.135.aspx](https://journals.lww.com/ear-hearing/Fulltext/9900/Copresence_Was_Found_to_Be_Related_to_Some_Pupil.135.aspx) <https://doi.org/10.1097/AUD.00000000000001361>
- Pielage, H., Zekveld, A. A., Saunders, G. H., Versfeld, N. J., Lunner, T., & Kramer, S. E. (2021). The presence of another individual influences listening effort, but not performance. *Ear & Hearing*, 42(6), 1577–1589. <https://doi.org/10.1097/aud.0000000000001046>
- Plain, B., Pielage, H., Richter, M., Bhuiyan, T. A., Lunner, T., Kramer, S. E., & Zekveld, A. A. (2021). Social observation increases the cardiovascular response of hearing-impaired listeners during a speech reception task. *Hearing Research*, 410, 108334. <https://doi.org/10.1016/J.HEARES.2021.108334>
- Plain, B., Richter, M., Zekveld, A. A., Lunner, T., Bhuiyan, T., & Kramer, S. E. (2020). Investigating the influences of task demand and reward on cardiac Pre-ejection period reactivity

- during a speech-in-noise task. *Ear & Hearing*, 42(3), 718–731. <https://doi.org/10.1097/aud.0000000000000971>
- Raghu, S., & Sriraam, N. (2018). Classification of focal and non-focal EEG signals using neighborhood component analysis and machine learning algorithms. *Expert Systems with Applications*, 113, 18–32. <https://doi.org/10.1016/j.eswa.2018.06.031>
- Rahman, T., Ghosh, A. K., Shuvo, M. H., & Rahman, M. (2015). Mental stress recognition using K-nearest neighbor (KNN) classifier on EEG signals [Paper presentation]. International conference on materials, electronics & information engineering, ICMEIE, 1–4.
- Raza, S. B., Patterson, R. P., & Wang, L. (1992). Filtering respiration and low-frequency movement artefacts from the cardiogenic electrical impedance signal. *Medical & Biological Engineering & Computing*, 30(5), 556–561. <https://doi.org/10.1007/BF02457837>
- Richter, M. (2016). The moderating effect of success importance on the relationship between listening demand and listening effort. *Ear and Hearing*, 37, 111S–117S. <https://doi.org/10.1097/AUD.0000000000000295>
- Richter, M., & Slade, K. (2017). Interpretation of physiological indicators of motivation: Caveats and recommendations. *International Journal of Psychophysiology*, 119, 4–10. <https://doi.org/10.1016/J.IJPSYCHO.2017.04.007>
- Rönnberg, J. (2003). Cognition in the hearing impaired and deaf as a bridge between signal and dialogue: A framework and a model. *International Journal of Audiology*, 42(SUPPL. 1), 68–76. <https://doi.org/10.3109/14992020309074626>
- Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, Ö., Signoret, C., Stenfelt, S., Pichora-Fuller, K., & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience*, 7. <https://doi.org/10.3389/fnsys.2013.00031>
- Rönnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: A working memory system for ease of language understanding (ELU). *International Journal of Audiology*, 47(SUPPL. 2), S99–S105. <https://doi.org/10.1080/14992020802301167>
- Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). The need to approximate the use-case in clinical machine learning. *GigaScience*, 6(5), 1–9. <https://doi.org/10.1093/GIGASCIENCE/GIX019>
- Sarkar, M., & Leong, T. Y. (2000). Application of K-nearest neighbors algorithm on breast cancer diagnosis problem [Paper presentation]. Proceedings / AMIA...Annual Symposium. AMIA Symposium, 759–763.
- Seeman, S., & Sims, R. (2015). Comparison of psychophysiological and dual-task measures of listening effort. *Journal of Speech, Language, and Hearing Research*, 58(6), 1781–1792. [https://doi.org/10.1044/2015\\_JSLHR-H-14-0180](https://doi.org/10.1044/2015_JSLHR-H-14-0180)
- Sherwood, A., Allen, M. T., Fahrenberg, J., Kelsey, R. M., Lovallo, W. R., & van Doornen, L. J. P. (1990). Methodological guidelines for impedance cardiography. *Psychophysiology*, 27(1), 1–23. <https://doi.org/10.1111/j.1469-8986.1990.tb02171.x>
- Sherwood, A., Allen, M. T., Obrist, P. A., & Langer, A. W. (1986). Evaluation of beta-adrenergic influences on cardiovascular and metabolic adjustments to physical and psychological stress. *Psychophysiology*, 23(1), 89–104. <https://doi.org/10.1111/j.1469-8986.1986.tb00602.x>
- Shinn-Cunningham, B. G., & Best, V. (2008). Selective attention in normal and impaired hearing. *Trends in Amplification*, 12(4), 283–299. <https://doi.org/10.1177/1084713808325306>
- Shukla, A., Harper, M., Pedersen, E., Goman, A., Suen, J. J., Price, C., Applebaum, J., Hoyer, M., Lin, F. R., & Reed, N. S. (2020). Hearing loss, loneliness, and social isolation: A systematic review. *Otolaryngology–Head and Neck Surgery*, 162(5), 622–633. <https://doi.org/10.1177/0194599820910377>
- Slade, K., Kramer, S. E., Fairclough, S., & Richter, M. (2021). Effortful listening: Sympathetic activity varies as a function of listening demand but parasympathetic activity does not. *Hearing Research*, 410, 108348. <https://doi.org/10.1016/j.heares.2021.108348>
- Spruit, I. M., Wilderjans, T. F., & van Steenbergen, H. (2018). Heart work after errors: Behavioral adjustment following error commission involves cardiac effort. *Cognitive, Affective and Behavioral Neuroscience*, 18(2), 375–388. <https://doi.org/10.3758/S13415-018-0576-6/TABLES/4>
- Steinhauer, S. R., Siegle, G. J., Condray, R., & Pless, M. (2004). Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. *International Journal of Psychophysiology*, 52(1), 77–86. <https://doi.org/10.1016/j.ijpsycho.2003.12.005>
- Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *Journal of Speech, Language, and Hearing Research*, 61(6), 1463–1486. [https://doi.org/10.1044/2018\\_JSLHR-H-17-0257](https://doi.org/10.1044/2018_JSLHR-H-17-0257)
- Strand, J. F., Ray, L., Dillman-Hasso, N. H., Villanueva, J., & Brown, V. A. (2021). Understanding speech amid the jingle and jangle: Recommendations for improving measurement practices in listening effort research. *Auditory Perception & Cognition*, 3, 169–188. <https://doi.org/10.1080/25742442.2021.1903293>
- Verney, S., Granholm, E., & Psychophysiology, D. D. (2001). Pupillary responses and processing resources on the visual backward masking task. *Cambridge.Org*.
- Vest, A. N., Da Poian, G., Li, Q., Liu, C., Nemati, S., Shah, A. J., & Clifford, G. D. (2018). An open source benchmarked toolbox for cardiovascular waveform and interval analysis. *Physiological Measurement*, 39(10), 105004. <https://doi.org/10.1088/1361-6579/AAE021>
- Wah, Y. B., Rahman, H. A. A., He, H., & Bulgiba, A. (2016). Handling imbalanced dataset using SVM and k-NN approach. *AIP Conference Proceedings*, 1750, 020023. <https://doi.org/10.1063/1.4954536>
- Wendt, D., Dau, T., & Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in Psychology*, 7(MAR). <https://doi.org/10.3389/FPSYG.2016.00345>
- Wendt, D., Koelewijn, T., Książek, P., Kramer, S. E., & Lunner, T. (2018). Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hearing Research*, 369, 67–78. <https://doi.org/10.1016/j.heares.2018.05.006>

- Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in Hearing*, 22, 2331216518800869. <https://doi.org/10.1177/2331216518800869>
- Yang, W., Wang, K., & Zuo, W. (2012). Neighborhood component feature selection for high-dimensional data. *Journal of Computers*, 7(1), 162–168. <https://doi.org/10.4304/jcp.7.1.161-168>
- Yigit, H. (2013). A weighting approach for KNN classifier [Paper presentation]. 2013 International conference on electronics, computer and computation, ICECCO 2013, 228–231. <https://doi.org/10.1109/ICECCO.2013.6718270>
- Yoo, K., Ahn, J., & Lee, S. H. (2021). The confounding effects of eye blinking on pupillometry, and their remedy. *PLoS ONE*, 16(12), e0261463. <https://doi.org/10.1371/JOURNAL.PONE.0261463>
- Zekveld, A. A., Koelewijn, T., & Kramer, S. E. (2018). The pupil dilation response to auditory stimuli: Current state of knowledge. *Trends in Hearing*, 22, 233121651877717. <https://doi.org/10.1177/2331216518777174>
- Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology*, 51(3), 277–284. <https://doi.org/10.1111/psyp.12151>
- Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, 31(4), 480–490. <https://doi.org/10.1097/AUD.0b013e3181d4f251>
- Zekveld, A. A., van Scheepen, J. A. M., Versfeld, N. J., Veerman, E. C. I., & Kramer, S. E. (2019). Please try harder! The influence of hearing status and evaluative feedback during listening on the pupil dilation response, saliva-cortisol and saliva alpha-amylase levels. *Hearing Research*, 381, 107768. <https://doi.org/10.1016/j.heares.2019.07.005>
- Ziegler, M. G. (2012). Psychological stress and the autonomic nervous system. In D. Robertson, I. Biaggioni, G. Burnstock, P. A. Low, & J. F. R. Paton (Eds.), *Primer on the autonomic nervous system* (pp. 291–293). Elsevier Inc. <https://doi.org/10.1016/B978-0-12-386525-0.00061-5>