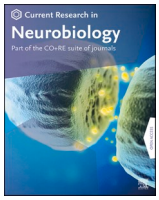
Current Research in Neurobiology 3 (2022) 100052

Contents lists available at ScienceDirect

Current Research in Neurobiology

journal homepage: www.sciencedirect.com/journal/current-research-in-neurobiology



Investigating effortful speech perception using fNIRS and pupillometry measures

Xin Zhou [a,*], Emily Burg [a,b], Alan Kan [c], Ruth Y. Litovsky [a,b]

[a] Waisman Center, University of Wisconsin Madison, WI, USA
[b] Department of Communication Science and Disorders, University of Wisconsin Madison, WI, USA
[c] School of Engineering, Macquarie University, Sydney, NSW, Australia

ARTICLE INFO

Keywords:
Functional near-infrared spectroscopy
Left lateral frontal cortex
Left auditory cortex
Pupillometry
Mental effort

ABSTRACT

The current study examined the neural mechanisms for mental effort and its correlation to speech perception using functional near-infrared spectroscopy (fNIRS) in listeners with normal hearing (NH). Data were collected while participants listened and responded to unprocessed and degraded sentences, where words were presented in grammatically correct or shuffled order. Effortful listening and task difficulty due to stimulus manipulations was confirmed using a subjective questionnaire and a well-established objective measure of mental effort – pupillometry. fNIRS measures focused on cortical responses in two *a priori* regions of interest, the left auditory cortex (AC) and lateral frontal cortex (LFC), which are closely related to auditory speech perception and listening effort, respectively. We examined the relations between the two objective measures and behavioral measures of speech perception (task performance) and task difficulty.

Results: demonstrated that changes in pupil dilation were positively correlated with the self-reported task difficulty levels and negatively correlated with the task performance scores. A significant and negative correlation between the two behavioral measures was also found. That is, as perceived task demands increased and task performance scores decreased, pupils dilated more. fNIRS measures (cerebral oxygenation) in the left AC and LFC were both negatively correlated with the self-reported task difficulty levels and positively correlated with task performance scores. These results suggest that pupillometry measures can indicate task demands and listening effort; whereas, fNIRS measures using a similar paradigm seem to reflect speech processing, but not effort.

1. Introduction

Listening effort is defined as "*the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out listening tasks*" (Pichora-Fuller et al., 2016). Long-term elevated listening effort has been related to fatigue, social isolation, and decreased quality of life (Alhanbali et al., 2017; Bess and Hornsby, 2014; Hughes et al., 2018; McGarrigle et al., 2014). Listeners affected by hearing loss typically report having to expend greater listening effort to understand speech in noisy environments (Alhanbali et al., 2017; Krueger et al., 2017). However, there is a dearth of knowledge regarding the neural mechanisms involved in effortful speech perception in listeners who are hearing impaired such as those who use cochlear implants. Functional near-infrared spectroscopy (fNIRS) is a promising technology for understanding effortful listening in a wide range of listeners and is compatible with cochlear implants (see perspectives in Bortfeld, 2019; for reviews see Butler et al., 2020). This study aimed to examine the viability of using fNIRS to identify neural markers for effortful speech perception in normal hearing (NH) listeners to serve as a baseline for future studies in hearing-impaired populations. fNIRS responses were collected using a newly designed stimulus paradigm that systematically controlled task difficulty. To assess listening effort exerted for this paradigm, task-evoked pupillometry, which is a well-established technique for assessing listening effort (see reviews by Laeng et al., 2012; Zekveld et al., 2018), was measured and compared with speech perception (task performance) and self-reported task difficulty. The same stimulus paradigm, once validated, was then used for fNIRS data collection to test the hypothesis that fNIRS can also provide a robust objective measure of listening effort.

1.1. Listening effort and pupillometry

Based on empirical data and a model of attention and effort by Kahneman (1973), Pichora-Fuller et al. (2016) proposed that listening

* Corresponding author.

https://doi.org/10.1016/j.crneur.2022.100052
Received 27 December 2021; Received in revised form 12 May 2022; Accepted 12 August 2022
Available online 20 August 2022

effort involves a combination of cognitive capacity demands needed to perform a task as well as listener's motivation. When the task demands are low, increasing motivation does not lead to increased listening effort. Similarly, when listeners' motivation to succeed in the task is low, the amount of listening effort expended is low regardless of the task demands. On the other hand, when listeners' motivation is high, increasing task demand results in greater listening effort being expended. However, if task demands become too high, listeners may lose motivation and listening effort subsides. Thus, listening effort is dependent upon a critical combination of task demands and situation that engages listeners' motivation.

Task-evoked pupillometry has been used to quantify effortful cognitive processing for more than 50 years (see review by Laeng et al., 2012). Pupillometry refers to the science of measuring pupillary changes during a task, resulting from a neural inhibitory mechanism acting on the parasympathetic oculomotor complex, or Edinger–Westphal nucleus, by the noradrenergic system's locus coeruleus (Beatty and Lucero-Wagoner, 2000). Prior research has shown that changes in pupil dilation correlate with task engagement and listening effort in response to distorted speech or speech in noise (Engelhardt et al., 2010; Kuchinsky et al., 2013; Piquado et al., 2010; Winn et al., 2015; Zekveld et al., 2014; Zekveld et al., 2010, 2011), or when listening to speech with low and high context (Winn and Moore, 2018). In line with the listening effort model proposed by Pichora-Fuller et al. (2016), when the task is too difficult and speech intelligibility is at floor level, pupil dilation decreases (Zekveld and Kramer, 2014), suggesting possible disengagement from the task and a decrease in effort being exerted. In other words, listeners simply "give up" on the idea that they might be successful on the task and stop paying attention to the content of the stimuli. These studies have established pupillometry as a reliable objective measure of listening effort and task demands as long as conditions are controlled to maintain listeners' motivation.

The first goal of the current study was to establish whether a newly designed paradigm would provide a systematic change in listening effort with varying task demands. For this paradigm, speech materials were spectrally degraded, i.e., vocoded which simulated the input from cochlear implants (Shannon et al., 1995) in three conditions, with two of them being further shuffled or interrupted. Listeners were also tested in a condition with clear (non-vocoded) sentences that were shuffled. Task demands were evaluated based on self-reported task difficulty levels and percent correct performance. The latter also confirmed listeners' motivation (engagement in task), as it was assumed that when task demand is high, high motivation was required to achieve good performance. Listening effort was assessed using pupillometry. Based on prior literature, we expected that pupil dilation would be greater in conditions when the task was reported harder and performance was poorer, thereby indicating higher effortful listening. Once these relationships were validated for the manipulations in the current study, fNIRS measures of cortical activity were collected in the same individuals with the same stimuli to investigate the abilities of fNIRS to quantify effortful speech perception.

1.2. fNIRS measures and regions of interest (ROIs)

fNIRS measures concentration changes of hemoglobin in the cerebral blood flow, which indirectly reveals task-evoked changes in local neuronal activity through neurovascular coupling (León-Carrión and León-Domínguez, 2012). fNIRS measures are highly correlated with the measure of blood-oxygen-level-dependent signals using functional magnetic resonance imaging (fMRI), which is the current gold standard of brain imaging in research and clinical assessment (Duan et al., 2012; Noah et al., 2015; Steinbrink et al., 2006). Because of the utilization of near-infrared light, fNIRS has the advantages over fMRI of being quiet, more tolerable with study designs involving noise, compatible with ferrous materials, and more tolerable with motion artifacts. Hence, fNIRS has been implemented to examine auditory perception and cognitive functions in populations that are challenging for fMRI such as children and infants (Cabrera and Gervain, 2020; Cristia et al., 2014; Lloyd-Fox et al., 2014, 2019; Mao et al., 2021), or patients implanted with magnetic devices such as cochlear implants (Anderson et al., 2016; Chen et al., 2017; Chen et al., 2016; McKay et al., 2016; Saliba et al., 2016; van de Rijt et al., 2016; Zhou et al., 2018). In the present study, we used fNIRS measures to examine cortical ROIs where neural markers of listening effort and speech intelligibility have been identified in NH listeners in previous studies using both fMRI and fNIRS.

The first cortical ROI was the left lateral frontal cortex (LFC). A meta-analysis of 485 neuroimaging studies by Liakakis et al. (2011) identified three main distinct clusters of ROIs in the left inferior frontal gyrus (IFG) within the LFC, with one ROI for language processing including semantic and phonological processing, one ROI for working memory, and one ROI for inhibitory response and executive function. As speech comprehension involves processing language at phonological, semantic, and syntactic levels, and short-term memory to retrieve one's knowledge of language, the ROIs in the left IFG are of specific interest when investigating a neural marker of effortful speech perception. Further, NH adults have been reported to show greater cortical activity in the left IFG when listening to spectrally degraded but intelligible speech (Lawrence et al., 2018; Wijayasiri et al., 2017; Wild et al., 2012b) or speech in noise (Alain et al., 2018), compared to unprocessed speech, suggesting a relation between cortical activity in the left IFG and listening effort. Due to the limited spatial resolution of fNIRS, this study aimed to measure cortical responses to effortful speech processing from the left LFC that covers the IFG.

The second cortical ROI was the left auditory cortex (AC). Results from meta-analyses of 128 neuroimaging studies (Vigneau et al., 2006, 2011) suggest that comprehension of speech includes sound perception in the primary AC and phonological processing in the middle and posterior superior temporal sulcus (STS). Arguments have long existed about which aspects of speech processing are left-lateralized, and which involve both hemispheres (Poeppel, 2014). Nonetheless, it is agreeable that, while the right AC might also be a host of lexical and context processing, it may not be specifically involved in phonological representation or working memory. Using positron emission tomography (PET), Scott et al. (2000) found that cortical activity in the left superior temporal gyrus (STG), but not in the right, was sensitive to passive listening of speech. However, when participants were required to actively listen and provide responses, intelligible speech produced greater cortical activity compared to unintelligible speech in the bilateral posterior STSs or STGs (Okada et al., 2010; Wild et al., 2012a). The above-mentioned studies support the idea that the left AC is sensitive to the intelligibility of speech stimuli. Inconsistencies regarding activity in the right AC could reflect differences between passive and active listening, and the variances in acoustics used across studies. That is, attention and active responses during speech perception might activate the right AC to promote clarity of auditory input.

The left AC and the left LFC are closely connected for auditory perception and semantic processing (see meta-analyses by Binder et al., 2009; see review by Hickok and Poeppel, 2007). An fNIRS study by Lawrence et al. (2018) examined the cortical responses in a group of NH listeners for spectrally degraded speech, and found that, as the intelligibility increased from 25% correct to 100% correct, responses in the ACs increased and responses in the left LFC (in the IFG) decreased. As speech intelligibility decreased (but remained above zero) and the task demands increased, greater effort was exerted to understand the speech compared to when speech was more intelligible. The opposite patterns of responses between AC and left IFG suggest that they contribute differently to the effortful perception of spectrally degraded speech.

The ultimate goal of the current study was to establish if fNIRS can provide a measure of effortful speech perception, by comparing fNIRS data with pupillometry and behavioral measures. We first examined the relations between pupillometry measures and the behavioral assessments of task difficulty and task performance (i.e., speech perception) to

validate our stimulus paradigm. Using the same stimulus paradigm, we then examined fNIRS measures in the left LFC and AC. As cortical activity in the left LFC has been associated with listening effort, and activity in the left AC has been associated with speech intelligibility, we hypothesized that fNIRS measures in these two regions would reveal similar relations with behavioral measures of effortful listening, in correspondence with pupillometry measures. Specifically, we predicted that fNIRS responses in the left LFC would increase as task difficulty was elevated, while activity in the left AC would decrease as task performance was reduced. We also considered the possibility that fNIRS measures in the left AC and LFC and pupillometry measures might yield different outcomes related to our stimulus paradigm. There were some unavoidable methodological differences between the two objective measures. If listening effort measures are particularly sensitive to these methods, then the associations between fNIRS and pupillometry might be weakened.

2. General methods

2.1. Participants

Twenty-eight (17 females, 25 right-handed) NH listeners, 18–27 years old, with a mean age of 21.6 years, and standard deviation (SD) of 2.3 years, were recruited for this study. This sample size was determined based on the sample sizes reported in previous studies also using fNIRS. For all listeners, hearing within normal limits was verified at octave frequencies between 125 Hz and 8000 Hz, with no more than 20 dB HL audiometric thresholds in each ear and with less than 10 dB difference between the two ears at any frequency. All listeners were native English-speaking students from the University of Wisconsin-Madison and were paid for their time. Experimental protocols were within standards set by the National Institutes of Health and approved by the University of Wisconsin–Madison's Health Sciences Institutional Review Board. All participants gave written consent. Data for all listeners were collected in two separate sessions a few days apart, with one session for study 1 (pupillometry) and one session for study 2 (fNIRS). The order of the two sessions was counterbalanced across participants.

2.2. Stimuli and conditions

Stimuli consisted of a subset of AuSTIN sentences (Dawson et al., 2013) with five or six words, with 3–4 keywords each, recorded by an American female speaker. AuSTIN sentences are modeled based on the simple and short Bamford-Kowal-Bench (BKB) sentences (Bench et al., 1979), and are suitable to test speech intelligibility in hearing-impaired children. An example AuSTIN sentence is 'He LOCKED the CAR DOOR', with the keywords in upper case. NH listeners were tested with sentences in quiet in four listening conditions that varied in task difficulty: vocoded (V), shuffled (S), shuffled-vocoded (VS), or vocoded-interrupted (VI). In the vocoded condition, the sentences were processed in AngelSim™ (TigerCIS) software using a white-noise carrier whereby the spectrum was divided into eight frequency bands between 200 Hz and 7000 Hz, with filters based on Greenwood functions (Shannon et al., 1995). The vocoded sentences were to simulate the spectrally degraded input from cochlear implants, with the envelope information being transmitted but temporal fine information being compromised. For the vocoded-interrupted condition, 31.25 ms silence periods replaced speech segments every 62.5 ms. The sentences were interrupted to further reduce the temporal information contained in speech, compared to the vocoded condition. In the two shuffled conditions, the last three words of the sentence were changed to produce a grammatically incorrect sentence. For instance, participants might hear 'He LOCKED CAR the DOOR' instead of the original sentence 'He LOCKED the CAR DOOR'. The sentences were shuffled for two reasons. First, listening to natural AuSTIN sentences in quiet is effortless for NH hearing adults, hence resulting in ceiling performance and minimal pupil dilation (Zekveld and Kramer, 2014). Second, sentences were shuffle-vocoded at the word level to simulate the scenario in which hearing-impaired listeners are around multiple persons, and they may confuse words from different people but have to fill the gap to follow the conversations. Speech stimuli were presented through a loudspeaker positioned at 0° azimuth and a distance of 1.5 m in front of participants.

Task difficulty for each stimulus condition was reported by participants using a difficulty scale between 0 and 10 [0 (effortless), 1 (extremely easy), 2 (very easy), 3 (easy), 4 (moderate), 5 (somewhat hard), 6 (moderately hard), 7 (hard), 8 (very hard), 9 (very, very hard), or 10 (extremely hard)]. The difficulty scale was shown on a monitor in front of the listener.

3. Study 1 – validating the designed paradigm

The goal of the first study was to establish whether our designed stimulus paradigm could be used to measure systematic changes in listening effort and task difficulty. Task-evoked pupillometry responses were measured and compared with self-reported task difficulty and task performance. Agreement across measures would validate this stimulus paradigm designed for the fNIRS component of the study.

3.1. Methods

3.1.1. Data collection

The pupillometry session was conducted in a sound booth (International Acoustics Company; IAC) with lighting in the room calibrated to provide a luminance of 46 lux at participants' eye position. An eye tracker (Eyelink 1000, SR research) was used for collecting pupil dilation data at a sampling frequency of 1000 Hz. After being seated in a chair with their chin and forehead supported by a headrest, a series of eye tracker calibration procedures were conducted to optimize the pupillary response measures. Pupillary data were collected from the left eye of all participants, except for one participant who reported having a dominant right eye. The dynamic range of pupil size for each participant was measured by having them look at a cross at the center of a screen that varied from dim (1.3 lux, Fig. 1A) to bright (137.5 lux, Fig. 1B) luminance for 11 s, which caused pupil dilation and constriction, respectively (Piquado et al., 2010). Both before and after the speech perception task, five trials of pupillary response in the same luminance condition (dim or bright), were collected in a row with a 10–20 s break between trials. The order of luminance conditions was randomized.

Speech-related pupillary data were collected in separate trials that started with a 3 s silent baseline period before the presentation of a 2-s sentence, followed by a 1.5 s waiting period (Fig. 1C). Then an audio prompt (a 200 ms beep) was used to indicate to the participant to verbally repeat back the sentence they heard. Participants had a maximum of 6.5 s window to respond after the prompt. In the shuffled conditions, as mentioned above, participants had to mentally reorder the sentences and report a grammatically correct sentence before responding. During the trial, a fixation marker (cross) was shown on a gray background on a monitor that was mounted above the eye tracker in front of participants. The luminance of the monitor measured at participants' eye position was 50.6 lux. The cross changed color to provide a visual indication of each phase of the trial.

Trials for each stimulus condition were grouped into blocks of five sentences, and conditions were presented in a random order that was counterbalanced across participants. In total, five blocks (25 trials total) of data were collected for each listening condition. The pupillometry session took about 1.5–2 h per person.

3.1.2. Assessments of self-reported task difficulty and task performance

Task performance scores in the pupillometry session were calculated for each condition as the number of *whole* sentences correctly repeated back divided by the total number of trials (n = 25). Whenever participants incorrectly repeated back any single word in the sentence, the trial

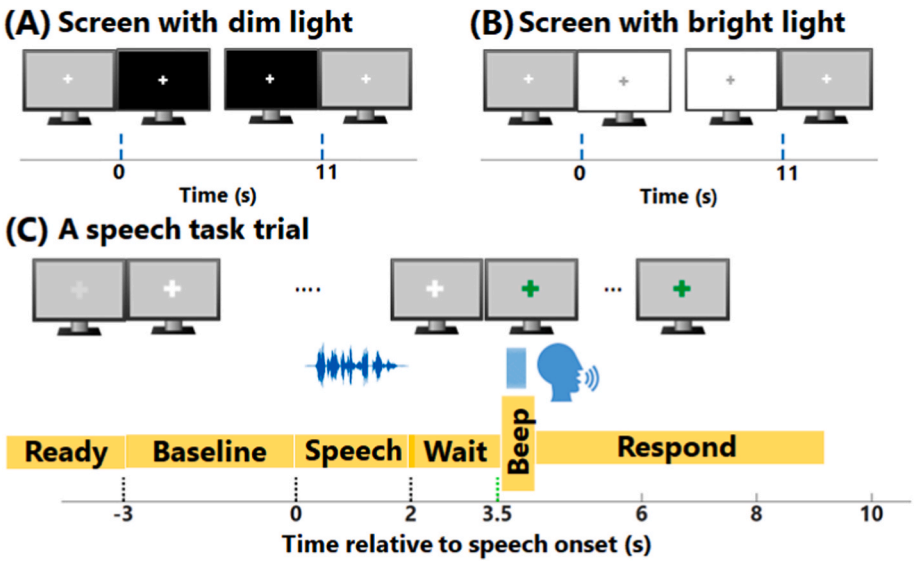


Fig. 1. Schematic diagram of pupillometry data collection. The figure shows the stimulus timeline for measuring an individual's pupil dynamic range when exposed to (A) dim and (B) bright screens, respectively, and (C) shows the timeline of a trial of the speech perception task. Each trial started with a baseline period, followed by a speech sentence, then a waiting period, and finally a response period when participants verbally repeated back the sentence they heard.

was counted as wrong. In the shuffled conditions, responses were counted as correct when participants successfully reordered the words and repeated back a coherent sentence. We used the term 'task performance' instead of 'speech intelligibility' as we believe that in the shuffled condition, sentences out of order were still intelligible but incorrect. Task performance scores were then transformed using a rationalized arcsine transform to alleviate ceiling effects (Studebaker, 1985). After each block of five sentences, participants were asked to report the task difficulty level between 0 and 10. The self-reported task difficulty level per condition was the average across five blocks for each stimulus condition.

3.1.3. *Pupillometry data analysis*

Pupillometry data analyses were conducted in MATLAB (R2017, The MathWorks, Inc.) with functions from CHAP toolbox used to de-blink and exclude trials of poor data quality (Hershman et al., 2019). The pupillometry data were pre-processed to account for eye blinks. Blinks in the pupil data were first identified by detecting pupil measurements (<100 ms) that were outside 3.5 SD of the mean pupil dilation in a trial. Then, to de-blink, a linear interpolation was applied based on the neighbor samples. Trials with more than 30% eye blinks were considered too noisy and were excluded from further analysis (Winn et al., 2018). Trials that had gaps (>100 ms) when the stimulus was being presented (likely due to participants closing their eyes or looking away from the monitor) were also excluded from further analyses. The pupil data were then down-sampled from 1000 Hz to 100 Hz and smoothed using a locally weighted linear regression MATLAB function (LOWESS, Burkey, 2013). This smoothing method was proposed in the studies by Cleveland (1979, 1981). Further, for each participant, trials that were outside 1.5 SD of the grand average of pupil responses across all 25 trials within 1–5 s after stimulus onset were considered outliers and were excluded from further analysis. The group mean ± SD of trials that were excluded was 2.10 ± 1.47, 2.45 ± 1.40, 2.17 ± 1.44, and 2.42 ± 1.24 in the shuffled, vocoded, shuffled-vocoded, and vocoded-interrupted

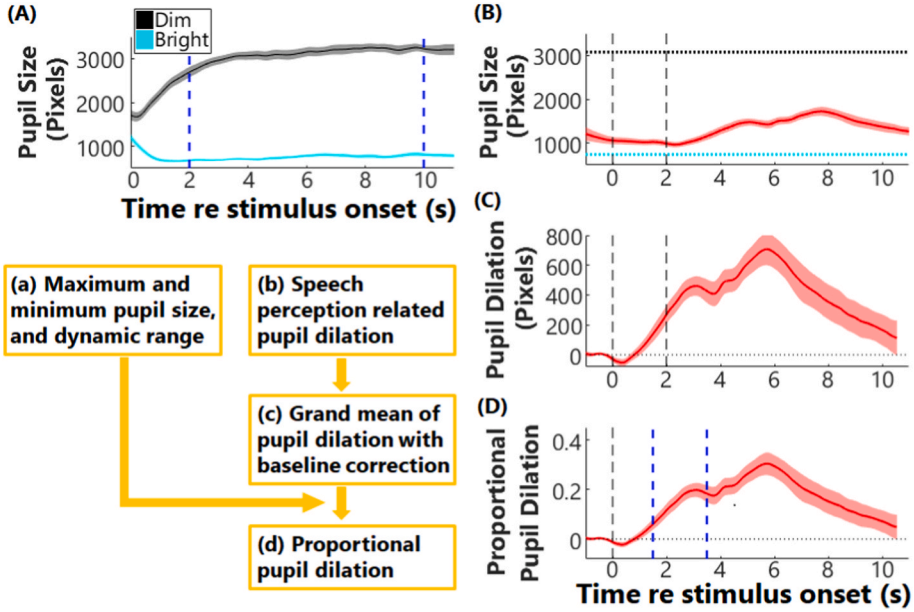


Fig. 2. Diagram of pupillometry data analysis. Panel (A) shows the pupil dilation to dim (black) and bright light (cyan). Solid lines and areas are for the grand means and standard error of the means (SEM). The two blue vertical dash lines plot the time window (2–10 s) that the pupil sizes were averaged to calculate the maximum and minimum of pupil sizes in response to light, i.e., dynamic range. Panel (B) plots an example of pupillometry response (mean ± SEM) in a speech perception task, with two black vertical dash lines showing the onset and offset of the speech stimulus. The black and cyan lines are the calculated maximum and minimum of pupil sizes in response to light, respectively, as shown in panel (a). Panel (C) plots the pupil dilation after baseline subtraction. Panel (D) plots the proportional pupil dilation relative to the dynamic range of pupil sizes. The two blue vertical dash lines plot the time window that the maximum of pupil dilation was identified. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

conditions, respectively.

The dynamic range of pupil sizes for individual participants was estimated by averaging the pupillary data across 10 repetitions in the dim and bright conditions, separately, as shown in Fig. 2A. From these averaged responses, the maximum and minimum pupil sizes were calculated as the mean pupil size within 2–10 s of the dim and bright conditions, respectively. The dynamic range of pupil size was then calculated as the difference between the maximum and minimum of pupil sizes (Fig. 2B, between the two horizontal dash lines).

To calculate speech-related pupil dilation, the baseline pupil sizes within 1 s preceding speech onset of each trial were averaged and then subtracted from the whole trial of pupillary data. After baseline correction, the pupil responses per listening condition were averaged across trials (Fig. 2C). As pupillary responses are the results of changes in luminance, (para)sympathetic nervous activity, arousal, attention, and effort (Laeng et al., 2012; Pichora-Fuller et al., 2016; Stanners et al., 1979), it is important to control the confounds in the experimental setup. To normalize the confounding effects on non-task-evoked pupillary responses in individuals, the proportional pupil dilation (Fig. 2D) was calculated as the change in pupil size divided by individual participants' dynamic range (Fig. 2B) for each listening condition. This is to account for the fact that the same change in pupil size might indicate different amounts of effort across individuals. For someone who has a small dynamic range of pupil sizes, a small task-evoked change in pupil size can be a large portion of his/her own dynamic range, indicating a large change in effort. Whereas for someone who has a large dynamic range, a small change in pupil size likely indicates little effort being exerted.

The proportional change in pupil size during the post-stimulus silent period of each trial (within t = 1.5–3.5 s relative to speech onset, Fig. 2D) was considered of interest because pupil dilation continues to change as listeners recall the speech information, conceptualize, and formulate a response. The magnitude of the first peak of the proportional change in pupil size within this window (t = 1.5 s–3.5 s) was calculated as the amount of pupil dilation per condition for individuals and used as a proxy measure of listening effort (see reviews by Winn et al., 2018).

3.1.4. Statistical analyses

Statistical analyses were carried out using R (version 3.6.0, R Core Team, 2019). To address our first goal of validating our designed stimulus paradigm, we conducted aligned rank transform (ART) analyses of variance (ANOVA) using ARTool (Wobbrock et al., 2011). ART tests were conducted on 1) the self-reported difficulty levels, 2) the task performance scores in the pupillometry session, and 3) the proportional peak amplitudes of pupil dilation. ART tests, which are nonparametric factorial analyses, were conducted for two reasons: (1) the self-reported task difficulty levels were ordinal measures; and (2) pupillometry data were not normally distributed, nor were their variances spherical. For follow-up pairwise comparisons, estimated marginal means (emmeans package) were examined and were Tukey's HSD corrected. Repeated measures correlations (Bakdash and Marusich, 2017) were calculated using rmcorr package to examine the relations between objective measures (peak pupil dilation) and behavioral measures (self-reported difficulty levels or task performance scores) of effortful speech perception. Rmcorr reveals the common regression slope, the association shared among individuals, without the violation of the independence of observations.

4. Results

Fig. 3 shows the self-reported task difficulty levels (Fig. 3A) and speech intelligibility scores (Fig. 3 B) collected during the pupillometry session for all participants. Listeners' report of task difficulty from the easiest to the most difficult was: shuffled, vocoded, shuffled-vocoded, and vocoded-interrupted (Fig. 3A). As self-reported difficulty increased, task performance in the corresponding listening condition

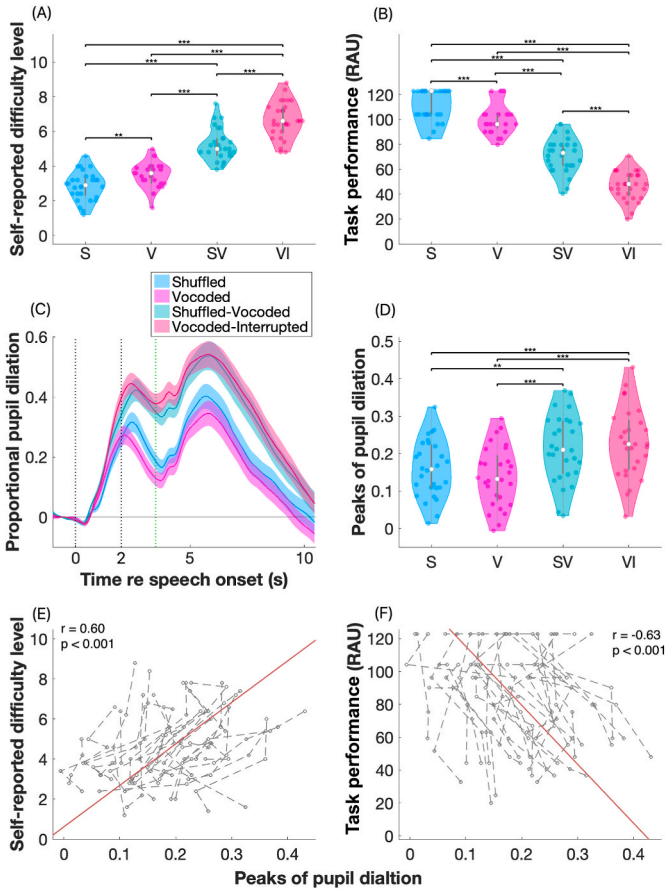


Fig. 3. Results from study 1 (pupillometry session). Violin plots show the self-reported task difficulty levels in panel (A) and percent correct task performance scores (RAU) in panel (B). The white dots indicate the group medians for each condition. Between 0 and 10, larger numbers indicate greater task difficulty levels. '*' with connection lines indicate the significance of pairwise comparison results with '*' for $p < 0.05$, '**' for $p < 0.01$, and '***' for $p < 0.001$. Panel (C) plots the group means (lines) and standard error of the means (SEMs, shaded areas) of proportional pupil dilation. The two blue lines indicate the window within which the peaks of proportional pupil dilation were identified. Panel (D) plots the peak of proportional pupil dilation for individuals (dots) in the shuffled (S), vocoded (V), shuffled-vocoded (SV), and vocoded-interrupted (VI) conditions; the white dots indicate the median results for each condition. The repeated measure correlations between the peaks of pupil dilation and self-reported task difficulty levels and task performance scores in individuals are shown in panels (E) and (F), respectively. Gray dots and lines show results in individuals in four conditions. The orange lines in panels (E) and (F) indicate the common association between pupillometry measures and behavioral measures among individuals. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

decreased (Fig. 3B), with a significant and negative correlation between the two measures ($r = -0.0.90$, $p < 0.001$). Detailed statistical results are reported in Table 1. The grand mean and standard error of mean (SEM) of proportional pupil dilation across trials are shown in Fig. 3C. After stimulus onset (t = 0 s), pupil dilation increased, peaking for the first time during the waiting period (the second black vertical dash line and the green dash line) and for the second time when participants verbally repeated the sentence they heard. The median of peaks of the proportional pupil dilation is shown as the white dot in Fig. 3D. Results from the ART test found significant differences across conditions, with greater proportional peaks of pupil dilation in vocoded-interrupted and shuffled-vocoded compared to in the shuffled and vocoded conditions (see Table 1). These results indicate that the shuffled-vocoded and vocoded-interrupted conditions required higher effort compared to the

Table 1
Statistical results for the NH group for the behavioral and objective measures. For each measure, results were compared between four listening conditions, i.e., vocoded (V), shuffled (S), shuffled-vocoded (SV), and vocoded-interrupted (VI).

Measures		Differences across conditions (results from ART test)	*Post hoc* results (Tukey's HSD corrected)
Self-reported task difficulty (pupillometry session)		$F(3, 81) = 171{,}63$, $p < 0.001$	**(S** and **V) < (SV** and **VI)**; $p < 0.001$ S < V; $p < 0.01$
Self-reported task difficulty (fNIRS session)		$F(3, 78) = 91{,}89$, $p < 0.001$	**(S** and **V) < (SV** and **VI)**; $p < 0.001$
Speech intelligibility scores (pupillometry session)		$F(3, 81) = 288.08$, $p < 0.001$	**S > V > SV > VI**; $p < 0.001$
Pupil dilation		$F(3, 81) = 34.49$, $p < 0.001$	**S < SV**; $t(81) = -5.52$, $p < 0.001$ **S < VI**; $t(81) = -654$, $p < 0.001$ **V < SV**; $t(81) = -7.22$, $p < 0.001$ **V < VI**; $t(81) = -8.24$, $p < 0.001$
fNIRS measures	**left LFC**	$F(3, 81) = 5.18$, $p = 0.003$	**S > VI**; $t(81) = 3.45$, $p = 0.005$ **SV > VI**; $t(81) = 3.28$, $p = 0.008$ **V > VI**; $t(81) = 2.78$, $p = 0.034$
	left AC	$F(3, 81) = 2.96$, $p = 0.037$	**S > VI**; $t(81) = 2.84$, $p = 0.029$

vocoded sentence and shuffled conditions. The repeated measures correlation results between the peaks of pupil dilation and the two behavioral measures are shown in Fig. 3 (E and F). Results found that the pupillometry measures were significantly positively correlated with self-reported difficulty levels were (Fig. 3E, $r = 0.60$, $p < 0.001$) and negatively correlated with the task performance scores (Fig. 3F, $r = -0.63$, $p < 0.001$).

To summarize, the behavioral measures revealed significantly different task difficulty levels and task performance scores between the four conditions, with the self-reported task difficulty increased and task performance decreased in the order of, shuffled, vocoded, shuffled-vocoded and vocoded-interrupted conditions. In line with the behavioral measures, we found greater pupil dilation in the shuffled-vocoded and vocoded-interrupted conditions, compared to in the shuffled and vocoded conditions, but there were no significant differences between the former or the latter two conditions. Further, results from repeated measure correlations found that as the perceived task demands increased and task performance decreased, changes in pupil dilation were greater. Overall, our pupil results suggest that the designed stimulus paradigm was able to elicit varying levels of listening effort in a group of NH listeners.

5. Study 2 - fNIRS

The second study tested the hypothesis that fNIRS could provide a robust objective measure of listening effort by comparing fNIRS measures from the left AC and LFC with behavioral measures of task difficulty and speech intelligibility scores.

5.1. Methods

5.1.1. Data collection

The fNIRS data collection was designed to be as similar to the pupillometry data collection as possible but with the constraints that listeners could not verbally repeat back sentences as it would disturb the fNIRS measurement and that hemoglobin response changes are typically slower than that of pupil dilation. The fNIRS session was conducted in a second standard IAC sound booth equipped with a NIRScout (NIRx Medical Technologies, LLC) system for data collection. fNIRS uses near-infrared light to measure concentration changes of oxygenated and deoxygenated hemoglobin (here called ΔHbO and ΔHbR, respectively) in multiple capillary beds in extracerebral and cerebral brain tissues (Jobsis, 1977; Villringer et al., 1993). The NIRScout system had 16 LED light sources that emitted near-infrared light at wavelengths of 760 and 850 nm, 16 avalanche photodiode (APD) detectors, and a bundle of 8 short channels. Fig. 4A shows the connection between light sources (red), detectors (blue), and short channels (green circles) on the left hemisphere. A symmetric montage (not shown) was placed on the right hemisphere. The optodes were held in place by a NIRScap matched to the head circumference for each participant, and were located based on the standardized 10-10 system (Acharya et al., 2016). As the distances between some source-detector pairs could be above the optimal distance (30 mm) for fNIRS data recording, plastic spacers were used to keep the channel distances at 30 mm. To center the NIRScap and correctly posit the optodes on the head, the Cz was positioned halfway between the Nasion and Inion, and halfway between the two pre-auricular points. Further, the frontal location Fp1 was positioned at 10% of the Nasion-Inion distance (a few centimeters above the eyebrows). The cap was attached to a chest wrap for fixation. Then, gains for all the channels set by the NIRScout system were checked. Gains between 4 and 7 suggest good light intensity, as a maximum gain of 8 indicates that detectors receive little or no light from the source. Whereas gains between 0 and 3 indicate that the light detected may have taken a more direct route from the source to detector rather than being scattered through layers of the cortex as required. When channels did not show good intensity, it was most likely the optodes were not perpendicular to the skin or due to hair artifacts. In this case, the optodes were taken out and the hair underneath was pushed away before replacing the optodes. The procedure was repeated until most of the channels had gains between 4 and 7 (good light intensity).

A pseudo-random block design was used for fNIRS data collection. Five 2-s long sentences in the same listening condition were grouped in a 12.6 s long block, with a 0.65 s interval between each sentence (Fig. 4C). At the end of each block, a sentence was shown on the monitor (no audio), which may or may not have been one of the previous five sentences. Across all the blocks, half of the sentences shown on the monitor were not presented in the preceding block of five sentences. In the shuffled conditions, the sentence on the monitor was shown unshuffled (grammatically correct). Participants were given 7 s to indicate whether the sentence on the monitor was one of the sentences presented in the preceding block by clicking a mouse button; yes (left click) or no (right click). Immediately after a mouse click, or until the end of 7 s when no response was made, a baseline period (ranging from 25 to 35 s) started. Blocks of different listening conditions were presented in a random order and counterbalanced across listeners. After each testing period, participants were required to report the task difficulty level of each testing condition using the same scale as the pupillometry task. To help with recalling the difficulty and stimuli, participants listened to an additional block of sentences for each condition before responding. A total of ten blocks per listening condition were collected across four testing periods within 1–1.5 h.

5.1.2. Assessments of self-reported task difficulty and response accuracy

In the fNIRS session, the self-reported task difficulty level in each listening condition was calculated as the mean across the 4 testing periods. Participants' response accuracy in each condition was scored as the number of correct mouse clicks in response to the sentences on the monitor divided by the total number of blocks (n = 10), which was only used to indicate whether participants were engaged during the task. One-sample t-tests (one-way) were conducted on the response accuracies in the fNIRS session versus the chance level (50% correct) for each

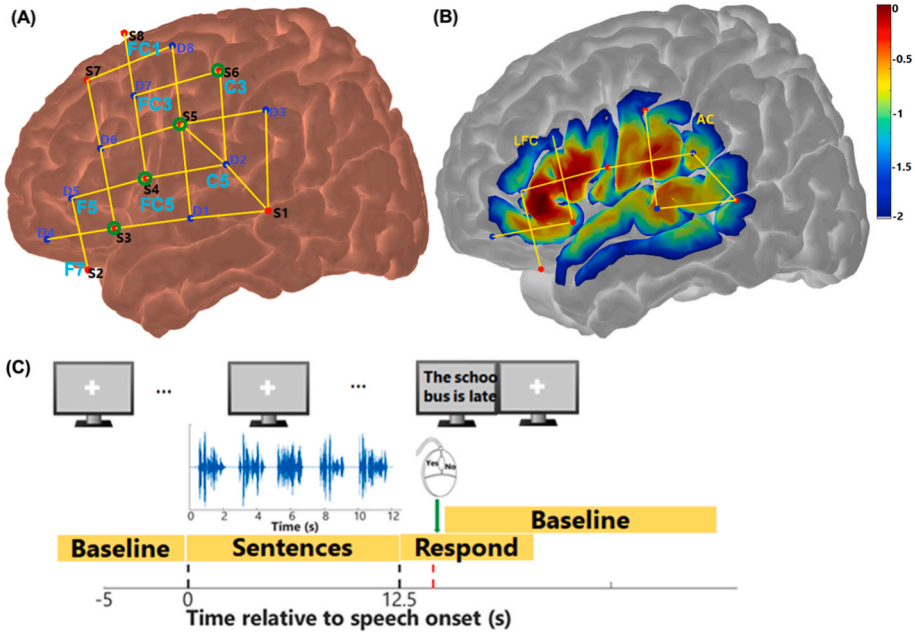


Fig. 4. fNIRS montage and schematic diagram of data collection. Panel (A) plots the anatomical locations of light sources (S1 – S8, red dots) and detectors (D1 - D8, blue dots) and the connections between them that provide fNIRS channels. The locations of sources and detectors on the 10-10 system were labeled in light blue. Only the left hemisphere is shown here, with a symmetric montage on the right. Green circles show the locations where short channels (8 mm) are located. Panel (B) shows the sensitivity map (with a log10 unit of mm^{-1}) of near-infrared light in measuring local chromophores in two regions of interest (ROIs), i.e., the lateral frontal cortex (LFC), and auditory cortex (AC). The more reddish color refers to the better sensitivity; the map was generated using AtlasViewer toolbox (Aasted et al., 2015). Panel (C) shows the timeline of a speech perception block in the fNIRS session, which started with a baseline, followed by a block of 5 sentences, and then a response period when participants responded to the sentence on the monitor by clicking mouse buttons. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

condition.

5.1.3. fNIRS data analysis

The fNIRS signals recorded by the NIRScout system were imported into MATLAB, with scripts written by the authors to pre-process data and exclude channels of poor data quality, and scripts from Homer2 software (Huppert et al., 2009) for computing ΔHbO and ΔHbR. The differences between ΔHbO and ΔHbR responses, i.e., ΔHbO - ΔHbR, were calculated, which indicates cerebral oxygenation, here called ΔHbC (Izzetoglu et al., 2004). Further statistics were conducted on ΔHbC amplitudes for two reasons. First, ΔHbC amplitudes combined information from both ΔHbO and ΔHbR measures. Running statistics on one (ΔHbC) not only revealed information from both measures but also reduced the complexity of reporting results from both measures, separately. Second, ΔHbC responses have revealed changes in neuronal activity in the prefrontal cortex related to mental effort (Ayaz et al., 2012; Liang et al., 2016; Nazeer et al., 2020; Rovetti et al., 2019). To pre-process the fNIRS data, step-like noises in each channel were identified as gaps in the data that were 2 SDs above or lower than the mean in a trial and were removed. These step-like noises can be caused by a sudden loss of contact between optodes and the skin, or the interposition of hair. To exclude fNIRS data of poor quality, the heartbeat signals in all channels were examined. As heartbeats are a salient signal in the fNIRS measurements, channels that fail to record a heartbeat signal is also unlikely to record other physiological or neural responses (Pollonini et al., 2014). fNIRS channels that showed correlations poorer than 0.35 between the heartbeat signal in the light intensity data of two different NIR wavelengths were excluded from further analysis. A lower cut-off threshold was chosen here compared to the recommendation of 0.75 in Pollonini et al. (2014), for three reasons. First, a cut-off threshold of 0.35 ensured at least 4 short channels were included for the GLM-PCA method, as recommended in Sato et al. (2016), which can provide a robust estimation of cerebral activity after denoising. Second, in a previous study (Zhou et al., 2020), a lower cut-off threshold (e.g., 0.15) yielded similar statistical conclusions compared to a cut-off threshold of 0.75. Third, across 28 participants in the current study, most of the participants showed good data quality. The medians (p_{50}) of SCI values were generally above 0.8, and the 25 percentile (p_{25}) of SCI values were above 0.6, except for two participants (Subj7 and Subj25). A cut-off at SCI = 0.35 would exclude no more than 25% channels per person per session. Please see supplementary materials. For LFC and AC across both hemispheres, with a total of 8 channels for each, the mean ± SD numbers of regular channels included across participants were 7.54 ± 0.79 and 7.72 ± 0.62, respectively. The mean ± SD numbers of short channels (with a total of 8) included for further analysis were 7.29 ± 1.23. Using scripts from Homer2, the NIR light intensity data for the two wavelengths in each channel were divided by the mean intensity value per wavelength, and then log-transformed, to compute optical density. To correct for motion artifacts in the optical density data, a wavelet analysis was then performed based on the method proposed in Molavi and Dumont (2010). Coefficients from wavelet analysis that were above 0.1 interquartile range indicated noise in the data and were set to zero. Finally, the concentration changes in ΔHbO and ΔHbR responses were calculated from the optical density data using the modified Beer-Lambert law (Delpy et al., 1988), with the effect of age and wavelengths of near-infrared light on the calculation of differential pathlength factor adjusted (Scholkmann and Wolf, 2013).

To reduce the systemic responses in the extracerebral tissue in the fNIRS data, a short-channel subtraction method was performed using a principal component analysis (PCA) method on short channels using scripts written by the authors. The first two principal components (PCs) that contributed the most to responses in short channels were assumed to be the 'global' systemic response component that also existed in the regular fNIRS channels. The first two PCs were treated as regressors for a general linear model to fit ΔHbO or ΔHbR signal in each channel. The product of the first two PCs and the corresponding coefficient from linear regression were then subtracted from ΔHbO or ΔHbR, separately, for each channel (Noah et al., 2021; Zhou et al., 2020). A third-order Butterworth band-pass filter (cut-off frequency at 0.01–0.5 Hz) was applied to remove the high-frequency physiological signals and low frequency drifts in the ΔHbO, ΔHbR, and ΔHbC (Yucel et al., 2021).

Finally, the block-average responses were calculated after subtracting the average of 5 s baseline before stimulus onset for each block of ΔHbO, ΔHbR and ΔHbC, separately. Individual blocks with values above or below the mean ± 2.5 SDs across ten blocks were excluded. For LFC and AC on two hemispheres, with a total of 8 channels for each, the mean ± SD numbers of regular channels included across participants were 7.54 ± 0.79 and 7.72 ± 0.62, respectively. The mean ± SD numbers of short channels (with a total of 8) included for further analysis were 7.29 ± 1.23. The block-average responses for each ROI were

calculated by averaging responses across the channels within that ROI. To quantify individuals' fNIRS responses, the amplitudes of ΔHbC responses were calculated by first identifying the peak of response 5–20 s after stimulus onset and then calculating the average of responses within a 5 s window centered at the identified peak.

5.1.4. Statistical analyses

To address our second goal of establishing fNIRS as a measure of effortful speech perception, we also conducted ART tests on fNIRS data (ΔHbC amplitudes), in the four conditions for all *a priori* ROIs on both hemispheres. ART tests were conducted because fNIRS measures were not normally distributed, nor were their variances spherical. As significant differences were found between ROIs and among conditions, with no significant differences between the two hemispheres, ART tests were conducted on the two *a priori* ROIs – left LFC and AC separately to examine the differences between conditions. For follow-up pairwise comparisons, estimated marginal means were examined and were Tukey's HSD corrected to account for multiple comparisons. For ROIs that showed a significant effect of condition, repeated measures correlations were calculated to examine the relations between ΔHbC amplitudes and the two behavioral measures (self-reported difficulty levels or task performance scores recorded from the pupillometry session). The Holm-Bonferroni method was used for multiple comparison corrections.

5.2. Results in study 2

Fig. 5 shows the group means (lines) and SEMs (shaded areas) of block-averaged fNIRS responses in the LFC (panel A) and AC (panel B) on the left and right hemispheres. After stimulus onset (t = 0 s), ΔHbO responses (red) slowly increased and ΔHbR responses (blue) decreased, with both returning to baseline (horizontal dash lines) after stimulus offset. The result that ΔHbO and ΔHbR changed in opposite directions, i.e., anti-correlated, was consistent with the profile of changes in hemoglobin related to neuronal activity. The cerebral oxygenation (ΔHbC) is also plotted (green). The group means ± SEMs of ΔHbC amplitudes in four conditions for each *a priori* ROI are shown in Fig. 5C. Results from ART tests on the ΔHbC amplitudes found a significant effect of ROI ($F(1, 405) = 28.51$, $p < 0.001$) and significant differences between conditions ($F(3, 405) = 4.95$, $p = 0.002$) but no significant effect of hemisphere ($F(1, 405) = 0.27$, $p = 0.61$), with no significant interaction between any of them.

For our *a priori* analysis, we conducted ART tests for fNIRS measures in the left LFC and AC to further examine the effect of effortful speech perception. Results from ART tests found a significant effect of condition on the ΔHbC amplitudes in the left LFC (Table 1). Interestingly, ΔHbC amplitudes in the left LFC were the smallest in the vocoded-interrupted (Fig. 5C), with no significant difference between the vocoded versus the two shuffled conditions. The group mean response accuracies were above the chance level of 50% correct ($p < 0.001$, Fig. 6B) in all conditions indicating that, on average, listeners remained engaged throughout the task in the fNIRS session. Although, a few participants showed lower than chance level response accuracy (below 50% correct) in the vocoded-interrupted condition. Further, participants reported the task difficulty in four conditions from the easiest to the hardiest in the order of shuffled, vocoded, shuffled-vocoded, and vocoded-interrupted (Fig. 6A) in the fNIRS session. The order of self-reported difficulty has a similar trend to that in the pupillometry session (Fig. 3A), except that there was no significant difference between the shuffled and vocoded conditions in the fNIRS session. The AC on both sides showed a trend of reduced ΔHbC amplitudes in the following order: shuffled, vocoded, shuffled-vocoded, vocoded-interrupted conditions (Fig. 5D). This trend corresponded to the decreasing task performance scores measured in the pupillometry session (Fig. 3B). Results from ART tests found a significant effect of condition on the ΔHbC amplitudes in the left AC, with

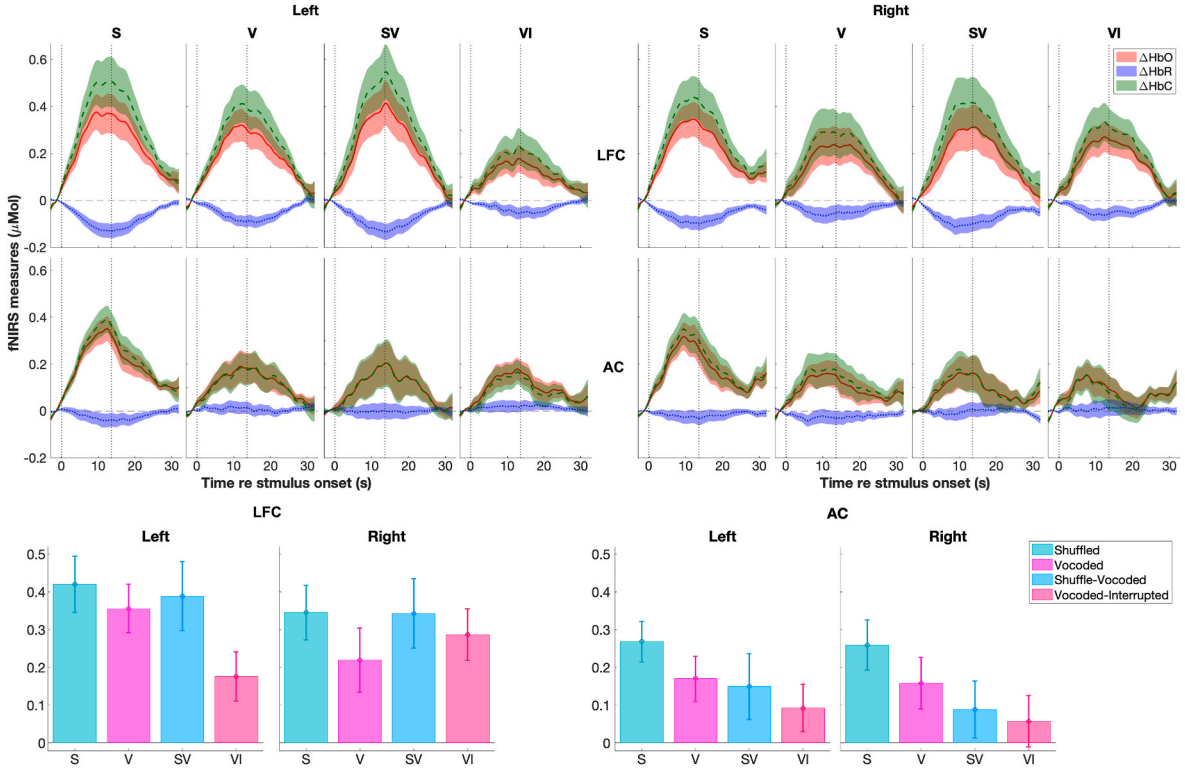


Fig. 5. Block-averaged fNIRS responses from study 2. Panels (A) and (B) plot the group means (lines) and standard error of the means (SEMs, shaded areas) of the ΔHbO (red, solid lines), ΔHbR responses (blue, dot lines), and the cerebral oxygenation (yellow, dash lines), i.e., ΔHbC, in the shuffled (S), vocoded (V), shuffled-vocoded (SV), and vocoded-interrupted (VI) conditions in two ROIs on the left and right hemispheres, respectively. In each subpanel, the two vertical dash lines plot the onset and offset of stimulation. Panel (C) plots the group mean (bars) and SEM (error bars) of the ΔHbC amplitudes in three ROIs on the left and right hemispheres. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

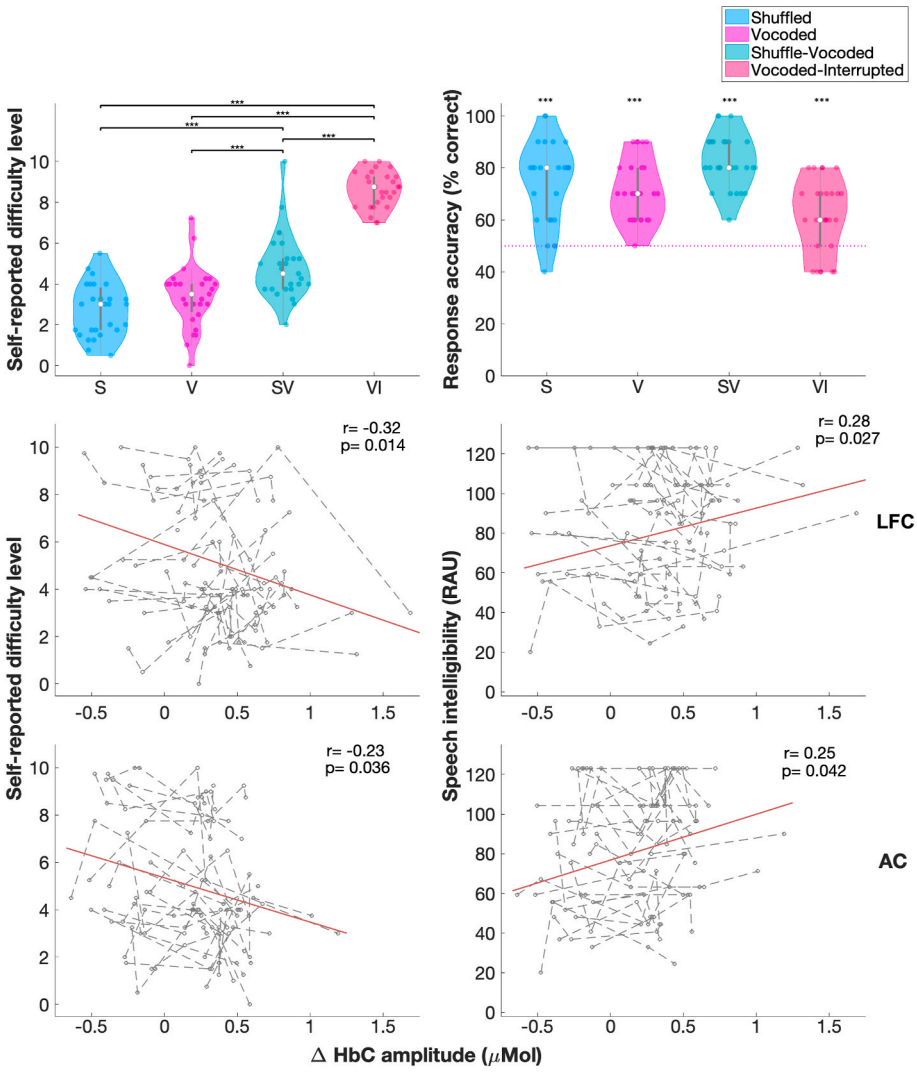


Fig. 6. Results from study 2 (fNIRS session). Panel (A) plots the self-reported task difficulty levels in the shuffled (S), vocoded (V), shuffled-vocoded (SV), and vocoded-interrupted (VI) conditions in the fNIRS session. Panel (B) plots the response accuracy in the fNIRS session; '***' indicates the significance of results ($p < 0.001$) from one-sample tests on the response accuracies versus 50%, i.e., chance level, (black horizontal lines). Panel (C–F) show the repeated measures correlations (rmcorr) between ΔHbC amplitudes in two ROIs and the self-reported task difficulty levels (from study 2) and task performance scores (from study 1). Gray dots and lines show results in individuals in four conditions. The orange lines in panels (C–F) indicate the common association between fNIRS measures in two ROIs and behavioral measures among individuals. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

greater amplitudes in the shuffled compared to in the vocoded-interrupted conditions (Fig. 5D).

As both the left LFC and AC showed a significant effect of listening condition, repeated measures correlations were examined for ΔHbC amplitudes in the two ROIs with behavioral measures. The ΔHbC amplitudes in both the left LFC (Fig. 6C) and AC (Fig. 6E) were significantly and *negatively* correlated with task difficulty levels recorded in the fNIRS session, and *positively* correlated with task performance scores measured in the pupillometry session (Fig. 6, D and F), with p-values Holm-Bonferroni corrected. That is, *greater* change in cerebral oxygenation in the left LFC and AC predicted *lower* self-reported task difficulty and *better* task performance. These results were surprising to us as we predicted opposite fNIRS response patterns between the left LFC and AC. Specifically, we predicted that response in the left LFC would increase with the task difficulty level, whereas response in the AC would increase with the speech intelligibility score as tasks became easier. The results in the LFC were driven by greater responses in the two shuffled conditions compared to the unshuffled and vocoded-interrupted conditions, suggesting that responses in the LFC were *not* related to task demands, but related to other perspectives of speech processing such as syntactic processing.

We also explored the relations between fNIRS responses in the left LFC and AC, respectively, with peak pupil dilation, as both fNIRS and pupillometry measures were correlated with behavioral measures of task difficulty and speech intelligibility. However, pupillometry measures were not correlated with fNIRS measures in the left LFC ($r = -0.05$, $p = 0.681$) or AC ($r = -0.10$, $p = 0.386$).

6. Discussion

We examined effortful speech perception using fNIRS and pupillometry to address two research goals in two separate studies. Our goal for study 1 was to validate our study-specific paradigm, which was designed to vary listening effort in a group of young NH listeners, measured using pupillometry, by manipulating the task difficulty and task performance (sentence understanding). In line with our expectations, as listeners reported higher task difficulty, task performance

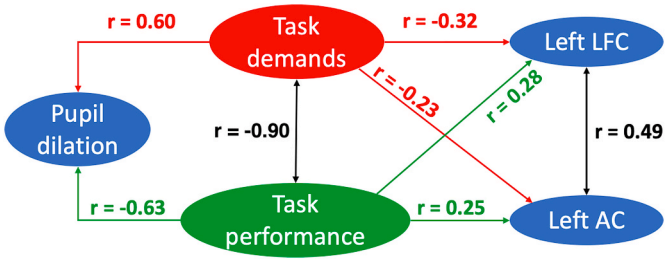


Fig. 7. Summary of relations between objective and behavioral measures of effortful speech perception.

scores decreased, and peaks of proportional pupil dilation increased (Fig. 7). Using the same stimuli and similar study design for the same individuals as in study 1, our goal for study 2 was to understand whether fNIRS could provide robust measures of listening effort that correlated with pupillometry and behavioral measures. Our results found that the cerebral oxygenation (ΔHbC) amplitudes in the left LFC and AC were both significantly and *negatively* correlated with self-reported task difficulty levels, and *positively* correlated with task performance scores (Fig. 7). However, fNIRS measures in the left LFC and AC were not related to pupillometry measures. These results suggest that fNIRS measures in the left LFC and AC were related to task demands and task performance, but might not reveal changes in listening effort specifically.

6.1. Pupillometry measures of effortful speech perception

Given the limited information about fNIRS measures of cognitive load, this study investigated whether fNIRS might provide a reasonable objective measure of listening effort, by comparing results from fNIRS with that from pupillometry. Pupillometry provides a continuous measure of pupil size, has been reported to reveal cognitive load (Beatty, 1982) and has been utilized extensively for speech perception tasks in previous studies (e.g., Koelewijn et al., 2012; Winn et al., 2015; Zekveld et al., 2010). To examine listening effort, the current study manipulated speech sentences, by degrading the speech spectrum, shuffling the order of words in the sentences, or periodically interrupting the speech segments to vary task demands. While pupillometry has been utilized to investigate listening effort for decades, the current study was, to our knowledge, the first one to use this approach to examine the impact of shuffled sentences or interrupted speech.

Interrupting speech by keeping segments on and off (silent) has been reported to decrease speech intelligibility depending on the interrupting rate (Cherry and Taylor, 1954; Miller and Licklider, 1950), as listeners may miss essential parts of some phonemes and syllable hence hard to identify the words they hear. When speech was vocoded to simulate the input from a cochlear implant, the loss of temporal-spectral information can further reduce the speech intelligibility of the interrupted speech. In a previous study, Bhargava et al. (2016) interrupted speech at varying rates between 1.5 and 24 Hz and presented to listeners with NH and cochlear implants. They found that NH listeners outperformed cochlear implant listeners at each single interrupting rates. Whereas when listening to interrupted speech that were eight-channel vocoded, the performance of NH listeners deteriorated and was quite similar that of cochlear implant listeners listening to interrupted speech at the same rates. Consistent with those results, NH listeners in the current study had low performance in the vocoded-interrupted condition with a median task performance score of 48 RAU versus ceiling scores when speech clean but shuffled (Fig. 3A). Coupled with the poorer task performance, NH listeners also reported the vocoded-interrupted condition the hardest among four conditions and showed the greatest peaks of proportional pupil dilation. These results confirmed that listening to vocoded-interrupted speech was demanding and NH listeners exerted increased listening effort for speech perception.

Our pupillometry measures found that degrading the temporal (comparing the vocoded-interrupted versus vocoded conditions) or spectral (comparing shuffle-vocoded versus shuffled conditions), information, both increased task difficulty, decreased behavioral performance, and resulted in more effort being exerted, manifested as greater pupil dilation (Fig. 3D). Though, our result showed significant differences in the task difficulty levels and task performances (Fig. 3A and B) between the shuffled and vocoded, and between shuffled-vocoded and vocoded-interrupted conditions, with no significant differences in the pupil dilation. In fact, NH listeners had a trend of greater pupil dilation in the shuffled versus vocoded conditions, i.e., opposite to the behavioral measures. These results could be due to that in the shuffled versus unshuffled conditions, listeners needed to reorder the words in the sentences before responding. The extra processing needed to reorganize a sentence may not have been perceived as having greater difficulty despite needing greater mental resources, and potentially long time and greater latency to peak when listening to sentences (Fig. 3C). Nonetheless, our results found significant differences in behavioral measures of task performance and self-reported task difficulty levels between four conditions, both of which were significantly correlated with pupillometry measures. Despite the differences in syntactic processing, these results suggest that the designed four conditions in the current study were of significantly varying task demands, and that listening effort exerted to understand these sentences, measured using pupillometry, was a function of task difficulty and speech intelligibility.

6.2. fNIRS measures of speech processing but not effort

Using the same stimulus manipulations, we expected that the changes in pupil dilation and fNIRS responses in *a priori* ROIs would reveal consistent trends across conditions, and that both measures would be associated with the behavioral measures of task difficulty and task performance. The fNIRS responses were first examined for the left LFC where cortical activity has been reported as a neural marker for listening effort in previous studies using fMRI (Alain et al., 2018; Wild et al., 2012b) and fNIRS (Wijayasiri et al., 2017). We expected that, as the conditions became more difficult, cerebral oxygenation in the left LFC would correspondingly increase. However, counter to our expectation, the cerebral oxygenation in the left LFC showed a significantly *negative* correlation with the self-reported task difficulty levels, and a significantly *positive* correlation with the task performance scores measured in the pupillometry session. Our results also demonstrated significantly smaller cerebral oxygenation amplitudes in the left IFG in the self-reported most difficult condition where the sentence was vocoded-interrupted compared to the other conditions (Fig. 5). We posit again that the extra processing needed to reorganize a sentence in the shuffled versus unshuffled conditions may have consumed greater mental resources, which was manifested as greater changes in the fNIRS measures in the shuffled conditions in the left LFC, which is essential for processing speech information including syntax and semantics (Friederici, 2012). In support of this theory, an fMRI study (Kristensen et al., 2013) that investigated the context/syntax on speech processing also reported greater activation in the left IFG when processing sentences with word order being manipulated. They presented written sentences in the canonical (subject-before-object) and shuffled order (object-initial clauses) and demonstrated greater activity in the left IFG in the shuffled condition compared to the canonical condition.

An alternate interpretation of the results may be that fNIRS measures in the LFC in the current study reflected changes in speech processing rather than changes in effort resulting from these manipulations. The two shuffled conditions, which were not self-reported as the hardest, involved more syntactic processing compared to the unshuffled conditions. Additionally, in the vocoded-interrupted condition, which was self-reported as the hardest, the amount of acoustic processing was reduced to half due to the interruptions. Therefore, the LFC responses, which were negatively correlated with task demands here, may in fact reveal a positive relation with the amount of speech processing. This interpretation is further supported by a significant and *positive* correlation between the left LFC and AC responses (repeated measure correlation, $r = 0.49$, $p < 0.001$, Fig. 7). As the LFC and AC are at lower and higher nodes of the speech processing pathways, respectively, this significant and positive correlation suggests that LFC may be involved in multiple aspects of speech perception. Contrary to our results, previous studies that have examined effortful speech perception of degraded speech found *greater* LFC responses in more effortful conditions in the LFC (Lawrence et al., 2018; Wijayasiri et al., 2017; Wild et al., 2012b), opposite to that in the AC. For instance, Lawrence et al. (2018) varied the degrees of degradation in the speech spectrum and found that, as the intelligibility increased from 25% correct to 100% correct, responses in

the ACs increased and responses in the left LFC (IFG) decreased. They interpreted the results as changes in the left IFG being related to effortful perception, and the changes in the AC being related to speech intelligibility. The results in the current study and in previous studies suggest that different configurations of stimulation among studies such as spectral degradation, interrupting or shuffling the order of words, or speech with different types or levels of masking noise could reveal some roles of the left LFC more for speech processing compared to the varying effort related to these manipulations.

The fNIRS responses in the AC were in line with our expectations, as task performance scores decreased across conditions, the cerebral oxygenation in the left AC decreased (Fig. 6). We were specifically interested in the left AC as previous studies have demonstrated markers of speech intelligibility in the left AC (Davis and Johnsrude, 2003; Narain et al., 2003; Scott et al., 2000). However, our results did not find a significant difference between the two hemispheres. Poeppel (2014) proposed that speech processing might be less left-lateralized than once believed, as speech perception and lexical level comprehension have been demonstrated in both hemispheres. In line with our results and the perspective of Poeppel (2014), ACs in both hemispheres have been reported to show greater activity to speech with better intelligibility or clarity (Lawrence et al., 2018; Obleser et al., 2007; Okada et al., 2010; Wild et al., 2012a). Studies that reported greater activity to more intelligible speech in both the left and right AC required participants to perform speech perception tasks related to the stimuli. Conversely, studies that only found activity in the left AC to be related to speech intelligibility across conditions had participants listening passively to the stimuli (Narain et al., 2003; Scott et al., 2000) or just indicating whether the speech was intelligible (Davis and Johnsrude, 2003). These results suggest that speech comprehension mainly involves the left AC, and the right AC contributes to certain degrees in varying configurations that require attentional listening.

6.3. *Differences across fNIRS and pupillometry paradigms, and limitations*

As noted above, peak proportional pupil dilation and fNIRS in the left AC and IFG were each correlated with self-reported task difficulty levels, and with task performance scores. However, the two objective measures were not correlated with each other. We will first consider the differences in methodologies. We must acknowledge that, while the study was intentionally designed to maximize similarity across the two measures, the tasks that participants performed were unavoidably different in the two situations. In the pupillometry session, participants listened to one sentence at a time, and verbally repeated back what they recognized after a 1.5 s waiting period. Each trial (sentence) was presented at participants' pace and pupil dilation in response to each sentence was measured. In the fNIRS session, designed for reliable fNIRS measurements, participants listened to five sentences in a block, with a 0.75 s interval between sentences. In this way, pupil dilation revealed effortful perception when listening to individual sentences in the degraded conditions. Whereas fNIRS measures reflected the continuous exposure to degraded speech and the cumulative change in cortical hemoglobin to blocks of stimuli; in fNIRS there was likely a greater amount of working memory involved to successfully perform the task. Further, sentences in the fNIRS session were presented more rapidly with shorter breaks, hence shorter time for recovery compared to in the pupillometry session. The short breaks between sentences and the great demand for working memory in the fNIRS session could have made the tasks more difficult, compared to that in the corresponding condition in the pupillometry session. Therefore, in the vocoded-interrupted condition when the task became too demanding (Fig. 6A), participants could have lost motivation, resulting in less listening effort being expended in the fNIRS session, but not in the pupillometry session for the same condition. There were also differences in the tasks between the two sessions. At the end of the block during the fNIRS session, a sentence was presented on the monitor, and participants were required to push a mouse button to indicate whether it was one of the sentences they had just heard. The use of a button-press response was done in order to avoid articulation which can cause motion artifacts from the movement of the temporalis muscle, resulting in contaminated fNIRS recordings in both frontal and temporal regions. However, the button-press could have made it a decision-making task, rather than a speech perception and recall task in the pupillometry session. These differences could have contributed to the lack of correlations between fNIRS measures in the left LFC and AC and pupillometry measures. Future work on this topic would be needed to determine whether more similar tasks can be developed or whether simultaneous data collection methods will produce parallel findings in pupillometry and fNIRS.

It is also possible that pupillometry measures revealed both effortful speech perception and non-effort-related changes in the physiological activity, including arousal, attention, and emotion (Sirois and Brisson, 2014; Winn et al., 2018). The effort and non-effort-related changes in physiology might be associated with cortical activation in different regions not limited to the left LFC and AC that were investigated in the current study, such as the working memory and meta-cognition network. To test this, future studies will need to implement a wider coverage of brain ROIs compared to the present study. Alternatively, fNIRS measures in the LFC in the current study might reveal speech processing rather than changes in effort resulting from these manipulations, as discussed earlier. This theory could also explain why our fNIRS measures in the LFC and AC were not correlated with pupillometry measures. Further, pupillometry measures showed greater pupil dilation for temporally degraded speech (by comparing vocoded-interrupted versus vocoded conditions) and for spectrally degraded speech (by comparing the shuffle-vocoded versus shuffled conditions). Whereas no such differences were observed in the fNIRS measures between the two pairs of conditions. These results further support that pupillometry measures reveal the relation between task demand and effort exerted, whereas fNIRS measures of the LFC may reflect the amount of speech processing involved. To further investigate the role of LFC in speech processing, future studies will need to better control the amount of speech processing and vary effort, or vice versa, to disassociate one from the other.

7. Conclusion

The current study investigated fNIRS measures of effortful speech perception in study-specific conditions that were designed to vary task demands. We validated these conditions in eliciting varying listening effort, by comparing task-evoked pupil dilation with behavioral measures of task difficulty and (speech perception) task performance from the same individuals. With the same stimuli and a similar protocol for the same individuals, fNIRS measures in the left LFC and AC were both significantly and *negatively* correlated with self-reported task difficulty levels, *positively* correlated with the task performance scores, but not correlated with pupillometry measures. The relation between cortical activity in the left AC and speech perception performance is consistent with what has been demonstrated in previous studies. Whereas, the unexpected relations between cortical activity in the left LFC and the two behavioral measures, and non-significant correlations between fNIRS and pupillometry measures suggest that our fNIRS measure in the LFC and AC revealed speech processing but not effort.

Data and code availability

fNIRS amplitude data, self-reported measures of task difficulty level, and (speech perception) task performance data that support the findings of this study are publicly available on Open Science Framework: https://osf.io/hvjsn/?view_only=036ee19a4f34452bb42f429abda32b06. The demographic information has been removed to protect the participants' privacy.

Author Contributions

XZ: collected and analyzed the data; all authors contributed to the interpretation of the data, and the writing and editing of this manuscript, All authors contributed to the conception of this study.

Author's note

Some of the data was presented at the Conference on Implantable Auditory Prostheses (CIAP, Lake Tahoe, CA, 2019).

Credit author statement

XZ contributed to the conception of this experiment, collected and analyzed the data, and wrote the first draft of this manuscript.

EB contributed to the conception of this experiment, the interpretation of the data, and editing of this manuscript.

AK contributed to the conception of this experiment, the interpretation of the data and editing of this manuscript.

RL contributed to the conception of this experiment, the interpretation of the data and editing of this manuscript.

Declaration of competing interest

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

Acknowledgments

This study was supported by NIH-NIDCD (R01DC003083 to RL and R03DC015321 to AK), UW-Madison's Office of the Vice Chancellor for Research, and a Core grant from NIH-NICHD (U54HD090256 to Waisman Center).

The authors appreciate the time and support from all the research participants.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crneur.2022.100052.

References

Aasted, C.M., Yucel, M.A., Cooper, R.J., Dubb, J., Tsuzuki, D., Becerra, L., Boas, D.A., 2015. Anatomical guidance for functional near-infrared spectroscopy: AtlasViewer tutorial. Neurophotonics 2 (2), 020801. https://doi.org/10.1117/1.NPh.2.2.020801.

Acharya, J.N., Hani, A., Cheek, J., Thirumala, P., Tsuchida, T.N., 2016. American clinical neurophysiology society guideline 2: guidelines for standard electrode position nomenclature. J. Clin. Neurophysiol. 33 (4), 308–311. https://doi.org/10.1097/WNP.0000000000000316.

Alain, C., Du, Y., Bernstein, L.J., Barten, T., Banai, K., 2018. Listening under difficult conditions: an activation likelihood estimation meta-analysis. Hum. Brain Mapp. 39 (7), 2695–2709. https://doi.org/10.1002/hbm.24031.

Alhanbali, S., Dawes, P., Lloyd, S., Munro, K.J., 2017. Self-reported listening-related effort and fatigue in hearing-impaired adults. Ear Hear. 38 (1), E39–E48. https://doi.org/10.1097/Aud.0000000000000361.

Anderson, C.A., Lazard, D.S., Hartley, D.E., 2016. Plasticity in bilateral superior temporal cortex: effects of deafness and cochlear implantation on auditory and visual speech processing. Hear. Res. 343, 138–149. https://doi.org/10.1016/j.heares.2016.07.013.

Ayaz, H., Shewokis, P.A., Bunce, S., Izzetoglu, K., Willems, B., Onaral, B., 2012. Optical brain monitoring for operator training and mental workload assessment. Neuroimage 59 (1), 36–47. https://doi.org/10.1016/j.neuroimage.2011.06.023.

Bakdash, J.Z., Marusich, L.R., 2017. Repeated measures correlation. Front. Psychol. 8, 456. https://doi.org/10.3389/fpsyg.2017.00456.

Beatty, J., 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychol. Bull. 91 (2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276.

Beatty, J., Lucero-Wagoner, B., 2000. The pupillary system. Handbook of psychophysiology 2, 142–162.

Bench, J., Kowal, A., Bamford, J., 1979. The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. Br. J. Audiol. 13 (3), 108–112. https://doi.org/10.3109/03005367909078884.

Bess, F.H., Hornsby, B.W., 2014. Commentary: listening can be exhausting—fatigue in children and adults with hearing loss. Ear Hear. 35 (6), 592.

Bhargava, P., Gaudrain, E., Baskent, D., 2016. The intelligibility of interrupted speech: cochlear implant users and normal hearing listeners. JARO J. Assoc. Res. Otolaryngol. 17 (5), 475–491. https://doi.org/10.1007/s10162-016-0565-9.

Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cerebr. Cortex 19 (12), 2767–2796.

Bortfeld, H., 2019. Functional near-infrared spectroscopy as a tool for assessing speech and spoken language processing in pediatric and adult cochlear implant users. Dev. Psychobiol. 61 (3), 430–443. https://doi.org/10.1002/dev.21818.

Burkey, J., 2013. In: LOWESS, Locally Weighted Scatterplot Smoothing for Linear and Nonlinear Data (Enhanced).

Butler, L.K., Kiran, S., Tager-Flusberg, H., 2020. Functional near-infrared spectroscopy in the study of speech and language impairment across the life span: a systematic review. Am. J. Speech Lang. Pathol 29 (3), 1674–1701. https://doi.org/10.1044/2020_AJSLP-19-00050.

Cabrera, L., Gervain, J., 2020. Speech perception at birth: the brain encodes fast and slow temporal information. Sci. Adv. 6 (30), eaba7830.

Chen, L.C., Puschmann, S., Debener, S., 2017. Increased cross-modal functional connectivity in cochlear implant users. Sci. Rep. 7 (1), 10043 https://doi.org/10.1038/s41598-017-10792-2.

Chen, L.C., Sandmann, P., Thorne, J.D., Bleichner, M.G., Debener, S., 2016. Cross-Modal functional reorganization of visual and auditory cortex in adult cochlear implant users identified with fNIRS, 2016 Neural Plast., 4382656. https://doi.org/10.1155/2016/4382656.

Cherry, E.C., Taylor, W.K., 1954. Some further experiments upon the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. 26 (4), 554–559. https://doi.org/10.1121/1.1907373.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. 74 (368), 829–836. https://doi.org/10.2307/2286407.

Cleveland, W.S., 1981. Lowess - a program for smoothing scatterplots by robust locally weighted regression. Am. Statistician 35 (1), 54. https://doi.org/10.2307/2683591.

Cristia, A., Minagawa-Kawai, Y., Egorova, N., Gervain, J., Filippin, L., Cabrol, D., Dupoux, E., 2014. Neural correlates of infant accent discrimination: an fNIRS study. Dev. Sci. 17 (4), 628–635. https://doi.org/10.1111/desc.12160.

Davis, M.H., Johnsrude, I.S., 2003. Hierarchical processing in spoken language comprehension. J. Neurosci. 23 (8), 3423–3431. Retrieved from <Go to ISI>://WOS:000182475200037.

Dawson, P.W., Hersbach, A.A., Swanson, B.A., 2013. An adaptive Australian sentence test in noise (AuSTIN). Ear Hear. 34 (5), 592–600. https://doi.org/10.1097/AUD.0b013e31828576fb.

Delpy, D.T., Cope, M., van der Zee, P., Arridge, S., Wray, S., Wyatt, J., 1988. Estimation of optical pathlength through tissue from direct time of flight measurement. Phys. Med. Biol. 33 (12), 1433.

Duan, L., Zhang, Y.J., Zhu, C.Z., 2012. Quantitative comparison of resting-state functional connectivity derived from fNIRS and fMRI: a simultaneous recording study. Neuroimage 60 (4), 2008–2018. https://doi.org/10.1016/j.neuroimage.2012.02.014.

Engelhardt, P.E., Ferreira, F., Patsenko, E.G., 2010. Pupillometry reveals processing load during spoken language comprehension. Q. J. Exp. Psychol. 63 (4), 639–645. https://doi.org/10.1080/17470210903469864.

Friederici, A.D., 2012. The cortical language circuit: from auditory perception to sentence comprehension. Trends Cognit. Sci. 16 (5), 262–268. https://doi.org/10.1016/j.tics.2012.04.001.

Hershman, R., Henik, A., Cohen, N., 2019. CHAP: Open-source software for processing and analyzing pupillometry data. Behav. Res. Methods. https://doi.org/10.3758/s13428-018-01190-1.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8 (5), 393–402. https://doi.org/10.1038/nrn2113.

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 Hear. 39 (5), 922–934. https://doi.org/10.1097/AUD.0000000000000553.

Huppert, T.J., Diamond, S.G., Franceschini, M.A., Boas, D.A., 2009. HomER: a review of time-series analysis methods for near-infrared spectroscopy of the brain. Appl. Opt. 48 (10), D280–D298.

Izzetoglu, K., Bunce, S., Onaral, B., Pourrezaei, K., Chance, B., 2004. Functional optical brain imaging using near-infrared during cognitive tasks. Int. J. Hum. Comput. Interact. 17 (2), 211–227. https://doi.org/10.1207/s15327590ijhc1702_6.

Jobsis, F.F., 1977. Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. Science 198 (4323), 1264–1267. Retrieved from. http://www.ncbi.nlm.nih.gov/pubmed/929199.

Kahneman, D., 1973. Attention and Effort, vol. 1063. Citeseer.

Koelewijn, T., Zekveld, A.A., Festen, J.M., Kramer, S.E., 2012. Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. Ear Hear. 33 (2), 291–300. https://doi.org/10.1097/AUD.0b013e3182310019.

Kristensen, L.B., Engberg-Pedersen, E., Nielsen, A.H., Wallentin, M., 2013. The influence of context on word order processing - an fMRI study. J. Neurolinguistics 26 (1), 73–88. https://doi.org/10.1016/j.jneuroling.2012.05.001.

Krueger, M., Schulte, M., Zokoll, M.A., Wagener, K.C., Meis, M., Brand, T., Holube, I., 2017. Relation between listening effort and speech intelligibility in noise. Am. J. Audiol. 26 (3), 378–392. https://doi.org/10.1044/2017_Aja-16-0136.

Kuchinsky, S.E., Ahlstrom, J.B., Vaden Jr., K.I., Cute, S.L., Humes, L.E., Dubno, J.R., Eckert, M.A., 2013. Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. Psychophysiology 50 (1), 23–34. https://doi.org/10.1111/j.1469-8986.2012.01477.x.

Laeng, B., Sirois, S., Gredeback, G., 2012. Pupillometry: a window to the preconscious? Perspect. Psychol. Sci. 7 (1), 18–27. https://doi.org/10.1177/1745691611427305.

Lawrence, R.J., Wiggins, I.M., Anderson, C.A., Davies-Thompson, J., Hartley, D.E., 2018. Cortical correlates of speech intelligibility measured using functional near-infrared spectroscopy (fNIRS). Hear. Res. 370, 53–64.

León-Carrión, J., León-Domínguez, U., 2012. Functional near-infrared spectroscopy (fNIRS): principles and neuroscientific applications. Neuroimaging-Methods.

Liakakis, G., Nickel, J., Seitz, R.J., 2011. Diversity of the inferior frontal gyrus-A meta-analysis of neuroimaging studies. Behav. Brain Res. 225 (1), 341–347. https://doi.org/10.1016/j.bbr.2011.06.022.

Liang, L.-Y., Getchell, N., Shewokis, P.A., 2016. Brain Activation in the Prefrontal Cortex during Motor and Cognitive Tasks in Adults.

Lloyd-Fox, S., Blasi, A., McCann, S., Rozhko, M., Katus, L., Mason, L., Team, B.P., 2019. Habituation and novelty detection fNIRS brain responses in 5-and 8-month-old infants: the Gambia and UK. Dev. Sci. 22 (5) https://doi.org/10.1111/desc.12817. ARTN e12817.

Lloyd-Fox, S., Papademetriou, M., Darboe, M.K., Everdell, N.L., Wegmuller, R., Prentice, A.M., Elwell, C.E., 2014. Functional near infrared spectroscopy (fNIRS) to assess cognitive function in infants in rural Africa. Sci. Rep. 4 https://doi.org/10.1038/srep04740. ARTN 4740.

Mao, D., Wunderlich, J., Savkovic, B., Jeffreys, E., Nicholls, N., Lee, O.W., McKay, C.M., 2021. Speech token detection and discrimination in individual infants using functional near-infrared spectroscopy. Sci. Rep. 11 (1) https://doi.org/10.1038/s41598-021-03595-z. ARTN 24006.

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. Int. J. Audiol.

McKay, C.M., Shah, A., Seghouane, A.K., Zhou, X., Cross, W., Litovsky, R., 2016. Connectivity in language areas of the brain in cochlear implant users as revealed by fNIRS. In: Van Dijk, P., Başkent, D., Gaudrain, E., De Kleine, E., Wagner, A., Lanting, C. (Eds.), Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing 894, 327–335.

Miller, G.A., Licklider, J.C.R., 1950. The intelligibility of interrupted speech. J. Acoust. Soc. Am. 22 (2), 167–173. https://doi.org/10.1121/1.1906584.

Molavi, B., Dumont, G.A., 2010. Wavelet based motion artifact removal for functional near infrared spectroscopy. Conf Proc IEEE Eng Med Biol Soc 5–8. https://doi.org/10.1109/IEMBS.2010.5626589, 2010.

Narain, C., Scott, S.K., Wise, R.J.S., Rosen, S., Leff, A., Iversen, S.D., Matthews, P.M., 2003. Defining a left-lateralized response specific to intelligible speech using fMRI. Cerebr. Cortex 13 (12), 1362–1368. https://doi.org/10.1093/cercor/bhg083.

Nazeer, H., Naseer, N., Khan, R.A., Noori, F.M., Qureshi, N.K., Khan, U.S., Khan, M.J., 2020. Enhancing classification accuracy of fNIRS-BCI using features acquired from vector-based phase analysis. J. Neural. Eng. 17 (5), 056025.

Noah, J.A., Ono, Y., Nomoto, Y., Shimada, S., Tachibana, A., Zhang, X., Hirsch, J., 2015. fMRI validation of fNIRS measurements during a naturalistic task. Jove-J. Vis. Exp. (100) https://doi.org/10.3791/52116. ARTN e52116.

Noah, J.A., Zhang, X., Dravida, S., DiCocco, C., Suzuki, T., Aslin, R.N., Hirsch, J., 2021. Comparison of short-channel separation and spatial domain filtering for removal of non-neural components in functional near-infrared spectroscopy signals. Neurophotonics 8 (1), 015004. https://doi.org/10.1117/1.NPh.8.1.015004.

Obleser, J., Wise, R.J., Dresner, M.A., Scott, S.K., 2007. Functional integration across brain regions improves speech perception under adverse listening conditions. J. Neurosci. 27 (9), 2283–2289. https://doi.org/10.1523/JNEUROSCI.4663-06.2007.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.H., Saberi, K., Hickok, G., 2010. Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cerebr. Cortex 20 (10), 2486–2495. https://doi.org/10.1093/cercor/bhp318.

Pichora-Fuller, M.K., Kramer, S.E., Eckert, M.A., Edwards, B., Hornsby, B.W., Humes, L. E., Wingfield, A., 2016. Hearing impairment and cognitive energy: the Framework for understanding effortful listening (FUEL). Ear Hear. 37 (Suppl. 1), 5S–27S. https://doi.org/10.1097/AUD.0000000000000312.

Piquado, T., Isaacowitz, D., Wingfield, A., 2010. Pupillometry as a measure of cognitive effort in younger and older adults. Psychophysiology 47 (3), 560–569. https://doi.org/10.1111/j.1469-8986.2009.00947.x.

Poeppel, D., 2014. The neuroanatomic and neurophysiological infrastructure for speech and language. Curr. Opin. Neurobiol. 28, 142–149. https://doi.org/10.1016/j.conb.2014.07.005.

Pollonini, L., Olds, C., Abaya, H., Bortfeld, H., Beauchamp, M.S., Oghalai, J.S., 2014. Auditory cortex activation to natural speech and simulated cochlear implant speech measured with functional near-infrared spectroscopy. Hear. Res. 309, 84–93. https://doi.org/10.1016/j.heares.2013.11.007.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 3.6.0. https://www.R-project.org/.

Rovetti, J., Goy, H., Pichora-Fuller, M.K., Russo, F.A., 2019. Functional near-infrared spectroscopy as a measure of listening effort in older adults who use hearing aids. Trends in Hearing 23. https://doi.org/10.1177/2331216519886722.

Saliba, J., Bortfeld, H., Levitin, D.J., Oghalai, J.S., 2016. Functional near-infrared spectroscopy for neuroimaging in cochlear implant recipients. Hear. Res. 338, 64–75. https://doi.org/10.1016/j.heares.2016.02.005.

Sato, T., Nambu, I., Takeda, K., Aihara, T., Yamashita, O., Isogaya, Y., Kawato, M., 2016. Reduction of global interference of scalp-hemodynamics in functional near-infrared spectroscopy using short distance probes. Neuroimage 141, 120–132.

Scholkmann, F., Wolf, M., 2013. General equation for the differential pathlength factor of the frontal human head depending on wavelength and age. J. Biomed. Opt. 18 (10), 105004 https://doi.org/10.1117/1.JBO.18.10.105004.

Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123 Pt 12, 2400–2406. Retrieved from. https://www.ncbi.nlm.nih.gov/pubmed/11099443.

Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270 (5234), 303–304.

Sirois, S., Brisson, J., 2014. Pupillometry. Wiley Interdiscip. Rev. Cogn. Sci. 5 (6), 679–692. https://doi.org/10.1002/wcs.1323.

Stanners, R.F., Coulter, M., Sweet, A.W., Murphy, P., 1979. The pupillary response as an indicator of arousal and cognition. Motiv. Emot. 3 (4), 319–340.

Steinbrink, J., Villringer, A., Kempf, F., Haux, D., Boden, S., Obrig, H., 2006. Illuminating the BOLD signal: combined fMRI-fNIRS studies. Magn. Reson. Imaging 24 (4), 495–505. https://doi.org/10.1016/j.mri.2005.12.034.

Studebaker, G.A., 1985. A rationalized arcsine transform. J. Speech Hear. Res. 28 (3), 455–462. https://doi.org/10.1044/jshr.2803.455.

van de Rijt, L.P., van Opstal, A.J., Mylanus, E.A., Straatman, L.V., Hu, H.Y., Snik, A.F., van Wanrooij, M.M., 2016. Temporal cortex activation to audiovisual speech in normal-hearing and cochlear implant users measured with functional near-infrared spectroscopy. Front. Hum. Neurosci. 10, 48. https://doi.org/10.3389/fnhum.2016.00048.

Vigneau, M., Beaucousin, V., Herve, P.Y., Duffau, H., Crivello, F., Houde, O., Tzourio-Mazoyer, N., 2006. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. Neuroimage 30 (4), 1414–1432. https://doi.org/10.1016/j.neuroimage.2005.11.002.

Vigneau, M., Beaucousin, V., Herve, P.Y., Jobard, G., Petit, L., Crivello, F., Tzourio-Mazoyer, N., 2011. What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing? Insights from a meta-analysis. Neuroimage 54 (1), 577–593. https://doi.org/10.1016/j.neuroimage.2010.07.036.

Villringer, A., Planck, J., Hock, C., Schleinkofer, L., Dirnagl, U., 1993. Near infrared spectroscopy (NIRS): a new tool to study hemodynamic changes during activation of brain function in human adults. Neurosci. Lett. 154 (1–2), 101–104. https://doi.org/10.1016/0304-3940(93)90181-j.

Wijayasiri, P., Hartley, D.E.H., Wiggins, I.M., 2017. Brain activity underlying the recovery of meaning from degraded speech: a functional near-infrared spectroscopy (fNIRS) study. Hear. Res. 351, 55–67. https://doi.org/10.1016/j.heares.2017.05.010.

Wild, C.J., Davis, M.H., Johnsrude, I.S., 2012a. Human auditory cortex is sensitive to the perceived clarity of speech. Neuroimage 60 (2), 1490–1502. https://doi.org/10.1016/j.neuroimage.2012.01.035.

Wild, C.J., Yusuf, A., Wilson, D.E., Peelle, J.E., Davis, M.H., Johnsrude, I.S., 2012b. Effortful listening: the processing of degraded speech depends critically on attention. J. Neurosci. 32 (40), 14010–14021. https://doi.org/10.1523/Jneurosci.1528-12.2012.

Winn, M.B., Edwards, J.R., Litovsky, R.Y., 2015. The impact of auditory spectral resolution on listening effort revealed by pupil dilation. Ear Hear. 36 (4), e153.

Winn, M.B., Moore, A.N., 2018. Pupillometry reveals that context benefit in speech perception can Be disrupted by later-occurring sounds, especially in listeners with cochlear implants. Trends in Hearing 22. https://doi.org/10.1177/2331216518808962.

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. Artn 2331216518800869 Trends in Hearing 22, 10.1177/2331216518800869.

Wobbrock, J.O., Findlater, L., Gergle, D., Higgins, J.J., 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In: Paper Presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Yucel, M.A., Luhmann, A.V., Scholkmann, F., Gervain, J., Dan, I., Ayaz, H., Wolf, M., 2021. Best practices for fNIRS publications. Neurophotonics 8 (1), 012101. https://doi.org/10.1117/1.NPh.8.1.012101.

Zekveld, A.A., Heslenfeld, D.J., Johnsrude, I.S., Versfeld, N.J., Kramer, S.E., 2014. The eye as a window to the listening brain: neural correlates of pupil size as a measure of cognitive listening load. Neuroimage 101, 76–86. https://doi.org/10.1016/j.neuroimage.2014.06.069.

Zekveld, A.A., Koelewijn, T., Kramer, S.E., 2018. The pupil dilation response to auditory stimuli: current state of knowledge. Trends Hear 22, 2331216518777174. 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 Hear. 31 (4), 480–490. https://doi.org/10.1097/AUD.0b013e3181d4f251.

Zekveld, A.A., Kramer, S.E., Festen, J.M., 2011. Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response. Ear Hear. 32 (4), 498–510. https://doi.org/10.1097/AUD.0b013e31820512bb.

Zhou, X., Seghouane, A.K., Shah, A., Innes-Brown, H., Cross, W., Litovsky, R.Y., McKay, C.M., 2018. Cortical speech processing in postlingually deaf adult cochlear implant users, as revealed by functional near-infrared spectroscopy. Trends Hear 22, 2331216518786850. https://doi.org/10.1177/2331216518786850.

Zhou, X., Sobczak, G., McKay, C.M., Litovsky, R.Y., 2020. Comparing fNIRS signal qualities between approaches with and without short channels. PLoS One 15 (12), e0244186. https://doi.org/10.1371/journal.pone.0244186.