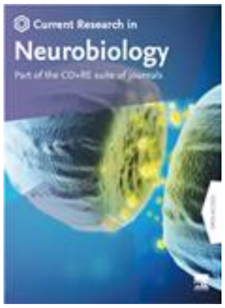
Peer Review Overview



Manuscript Title: Investigating effortful speech perception using fNIRS and pupillometry measures

Received	Dec 27, 2021
1st Decision	Apr 01, 2022
1st Revision Submitted	May 12, 2022
Accepted	Aug 12, 2022

1st Decision letter

Reference: CRNEUR-D-21-00079
Title: Investigating effortful speech perception using fNIRS and pupillometry measures
Journal: Current Research in Neurobiology

Dear Dr. Zhou,

Thank you for submitting your manuscript to Current Research in Neurobiology.

The reviewers recommend reconsideration of your manuscript following major revision. There are several substantial concerns that would need to be addressed and re-reviewed, as noted below in the reviewer comments.

I invite you to resubmit your manuscript after addressing the comments below. Please resubmit your revised manuscript by May 31, 2022.

When revising your manuscript, please consider all issues mentioned in the reviewers' comments carefully; outline every change made in response to their comments and provide suitable rebuttals for any comments not addressed. Please note that your revised submission will need to be re-reviewed.

Current Research in Neurobiology values your contribution and I look forward to receiving your revised manuscript.

CRNEUR aims to be a unique, community-led journal, as highlighted in the Editorial Introduction. As part of this vision, we will be regularly seeking input from the scientific community and encourage you and your co-authors to take the survey.

Kind regards,

Christopher I. Petkov
Editor in Chief
Current Research in Neurobiology

Comments from Editors and Reviewers:

Reviewer #1:

Review of CRNEUR-D-21-00079: Investigating effortful speech perception using fNIRS and pupillometry measures

Summary:
This study used fNIRS and pupillometry to evaluate effortful listening in normal-hearing participants. Correlations used to evaluate the relationship between the two objective measures and the behavioral measures of task performance and task difficultly. Both the fNIRS and pupillometry measures were correlated with the self-reported task difficulty levels and the task performance levels, but the direction of the relationship varied between the two objective measures of listening effort. Results suggested that both fNIRS and pupillometry are valid indicators of task demands and listening effort, but that protocol differences precluded a direct comparison between these two measures.
This is an interesting and timely study given the growing body of literature around objective measures of listening effort. I am unable to comment on the validity of the puillometry study design and analysis, so I will focus my comments on the fNIRS portion of the study. Mainly, it is unclear why certain pre-processing parameters were chosen for the fNIRS data, as well as the choice for statistical analyses. The processing and subsequent analysis of the fNIRS data as outlined in this paper is in conflict with the recent Best Practices for fNIRS Publications (Yucel et al., 2021). I have outlined some more specific concerns about the processing and analysis below.

Major Comments:
1. Many of the pre-processing steps used for these fNIRS data disregard the best practices for fNIRS data as outlined by Yucel et al (2021). I recommend that the authors re-analyze the fNIRS data using the agreed-upon standard practices in the field. There are a few examples of discrepancies between standard processing and analysis approaches for fNIRS detailed listed below:
a. The authors describe using a scalp-coupling index (correlations between two wavelengths in the heartrate range) to remove channels with poor signal quality. Why was a correlation of 0.35 chosen as the threshold? That seems like an oddly specific number, which is also a large departure from the Pollonini et al. (2014) recommendation of ~0.7.
b. There is no rationale for choosing HbC as the main variable of interest rather than HbO. It would be helpful if the authors could describe their rationale in more detail.
c. It is unclear why only the first two PCs were included in the short channel analysis. Did the authors try including different numbers of PCs before deciding on 2 PCs?
d. The band-pass filter upper cut-off frequency that was used is extremely low at 0.09 Hz. Yucel et al cautions setting this cut-off frequency substantially below 0.5 Hz as it can result in the removal of fluctuations in the brain response of interest. Indeed, the fNIRS time-series data as shown in Figure 5 has an overly filtered and smoothed appearance, which may have obscured variations in the peak response. This would not be a major issue if the time-series data was used simply for visualization of the temporal characteristics of the HB change and signal quality, but the authors use a block-averaging approach to the analysis, so the filter parameters could have a large impact on the peak and peak amplitude within the 5 sec averaging window.
e. There is no rationale for why ART ANOVAs were chosen for fNIRS analysis. Was there a violation of the parametric statistic assumptions? Also, why did the authors choose to a block-averaging approach

rather than a GLM approach to data analysis? Using a GLM approach would also allow the authors to regress out the "button press" from the fNIRS recording (assuming the button press timing was recorded for each person).

2. Correlation analyses were used to test the hypotheses; correlation plots are shown in Figures 3 and 6. These plots do not appear to represent a strong relationship between these two metrics. Although these correlations are reported as statistically significant, the visual relationships between these metrics are not compelling. I realize that repeated measures correlations were used so that multiple data points per single participant could be included (without violating the assumption of independence), but I wonder if included 4 data points per person has artificially inflated the stats here.

Reviewer #2:

The authors investigated the effects of task difficulty during speech processing on physiological measures of effort, including pupillometry and fNIRS. They found that pupillometry measures correlated with self-reported task difficulty and task performance, while activity in the left IFG and AC correlated with task difficulty and performance, albeit in opposite directions. Pupillometry measures did not correlate with fNIRS activity.

I find the pair of tasks interesting and the methods of each study relatively sound. However, I am confused as to why the authors are attempting to combine the studies, as the tasks themselves tap very different cognitive constructs with difference metrics of behavior, so it is unsurprising that the results do not correlate with each other. The paper is even laid out as two separate papers. Thus, I think this paper would make more sense as two separate manuscripts. I also have concerns about the operational constructs for motivation in the study. My more specific comments are found below.

Introduction:

The introduction is organized a little differently, e.g. a defense of many of the methods is found in the introduction well before the methods are presented. Nonetheless, I appreciate how thorough the authors were with their definitions and background in the introduction. I do have a concern regarding your references to motivation (here and in the Discussion). The motivation literature suggests that task performance can decline with increasing task difficulty, whether the participant has high motivation or not. This is even true in the context of listening effort. Thus, task performance is not a good proxy for motivation and should be avoided, especially since motivation was controlled in the current study. The only interpretation related to motivation might be diminished performance on the easy conditions, which was not found here.

In addition, the authors should use care when describing the locations that are being probed using fNIRS, as they did not collect MRI data on their participants so they cannot be certain what the anatomical correlates of their signal are. This is especially important when discussing specific sulci and gyri, as the authors are doing. The best that can be said about the current data is "frontal regions" and "parietal regions".

Methods and Results:

Please report your data in units that are more easily interpreted, rather than percentage. For example, the number of trials accepted should be in trials, not percentage. The same with number of channels and number of blocks excluded in the fNIRS dataset.

Again, the task designs between your pupillometry study and fNIRS study are completely different and the nature of how participants responded reflects different cognitive constructs. For the pupillometry study: Participants listened to a single sentence, and then were asked to either repeat the sentence back (if it was grammatically correct) or rearrange the words to make the sentence grammatically correct (in the case of the shuffled sentences). This is a listening effort, expressive language, and grammaticality task. In contrast, in the fNIRS study, the participants listened to five sentences and then were prompted with a sentence written on the screen and asked whether the sentence was one of the previous five sentences. In the shuffled condition, the 5 sentences were presented shuffled, but the prompt was not. This is clearly a working memory task with different levels of encoding difficulty. The two cannot be rightfully compared.

Discussion:
Again, the results cannot be interpreted in the context of motivation, because motivation was controlled in this study and the results do not suggest, by themselves, that motivation played a role.
The deactivation of the left neural regions could also be due to a difference in the neural dynamics at the whole brain level (i.e., more recruitment of other regions not recorded here). The authors should mention that.

Reviewer #3:

Review of manuscript nr. CRNEUR-D-21-00079 entitled "Investigating effortful speech perception using fNIRS and pupillometry measures"

This paper introduces a new paradigm to measure processing effort while normally hearing adult participants listed to various degraded speech conditions. There are a few methodological and other issue to be clarified, but the study is otherwise sound and the question is interesting.

My major concern is the interpretation / framing. The authors present the study as one that measures processing effort. While this may indeed be true for pupillometry, all the data clearly shows that for NIRS they simply measure auditory processing rather than processing effort (see point 6 below). This in itself is not a problem, but the paper should clearly be reframed to reflect this. Such a change would clearly take away from the novelty of the contributing, as there are many studies of speech perception with NIRS. However, the authors simply do not have grounds to claim that their NIRS study manages to dissociate processing effort from processing more generally.

Major concerns:

1. The Introduction very briefly and passingly discusses lateralization of speech processing. This is an issue with a huge literature that the paper doesn't seem to take into account. The processing of prosody, for instance, is lateralized to the right hemisphere and recent work suggests that speech processing may be more bilateral than previously believed (see Poeppel 2014 Current Opinion for a review).

2. I don't quite follow the logic of this predictin in the Introduction: "We also considered the possibility that fNIRS measures in the left AC and IFG and pupillometry measures might yield different outcomes related to our stimulus paradigm. Such a finding may suggest that protocol differences in conducting fNIRS and pupillometry data collection make it difficult to directly compare these listening effort measures."

The study explicitly says and sets out to test the *same* paradigm with NIRS and pupillometry. How could there be protocol differences then? Also, this prediction undermines the study objectives. If the paper is meant to be an empirical test of whether NIRS can be used to measure processing effort and pupillometry is an established measure of processing effort, then if the two are correlated, then the study can conclude that NIRS can also be used as a measure of effort. If the two end up not being correlated, then the conclusion needs to be that NIRS is not an appropriate measure of effort. This conclusion cannot be preempted by saying that NIRS may still be a measure of effort, albeit one that is uncorrelated with pupillometry. If tjis can be true, then there is no point running this study.

3. What motivates the experimental manipulations? Vocoding and interruption create degraded, challenging stimuli at the signal / auditory level, whereas shuffling does not affect the signal, but rather creates a challenge at the linguistic level. These conditions are thus not comparable and it is not clear what motivates these choices.

Relatedly, while the vocoder condition is ecologically valid and highly relevant given that it simulates the signal cochlear implant users hear, the motivation for the interrupted and the shuffled conditions are less clear.

Additionally, more details about the stimuli will be necessary, e.g. a list of the sentences in an appendix / supplementary material, figures illustrating the spectrograms of the stimuli etc.

4. The logic of the re-ordering task in the shuffled condition is not clear. Would asking participants to repeat back exactly what they heard not test better their speech perception accuracy? Relying on a reconstruction brings in linguistic knowledge (and other, related effects like word frequency, bi-word co-occurrence etc.) very heavily.
While participants certainly need to rely on such mechanisms in the other two conditions as well to some extent, the task is make a lot more complex and these mechanisms are made to be a lot more explicit in the re-order task. Not to mention that there may be ambiguity in this task (since the words Robin, kissed and Alex can be grammatically reconstructed as Robin kissed Alex and Alex kissed Robin, thus there are two, and not a single correct answer in some cases, at least potentially, further complicating the task).

5. In Study 1, despite the significant correlations between the pupillometry measure and the other two measures, self-reported effort and performance, qualitatively, there is a slight difference between them, in that the self-reported effort measure and performance are lowest/best for the shuffled condition, followed by the vocoded condition, whereas this pattern is reversed numerically for the pupillometry data and there seems to me no statistical difference.

6. As the most important concern, it is not clear whether NIRS measures processing or processing effort in this study. Indeed, the fact that activation was smallest in the vocoded interrupted condition suggests that NIRS is actually picking up on processing itself, which is the weakest / poorest in this condition. The directions of the significant correlations between NIRS and the behavioral measures also point in the same direction. The lack of correlation with pupillometry further suggest this.

To put it differently, if one was to use NIRS to measure auditory processing in these different conditions, one would proceed exactly as the authors did here (e.g. see Cabrera and Gervain 2020 for a NIRS study on newborns' processing of vocoded speech stimuli). There is nothing in the task that makes the NIRS responses measure effort rather than processing itself. The explanation in terms of dropping motivation should have been also visible in Study 1, soit cannot be used to justify the NIRS results.

Minor issues:
-The reference Wilcox and Biondi 2015 is not the best when referring to the use of NIRS in speech perception, language acquisition and cochlear implantation. There are many more specific references, e.g. Saliba et al. 2016, Gervain & Cabrera 2020 etc.
-Similarly, for the sentence "such as children and infants (see reviews by: Ferrari & Quaresima, 2012; Quaresima, Bisconti, & Ferrari, 2012; Vanderwert & Nelson, 2014; Wilcox & Biondi, 2015)", the reviews are not very appropriate, except maybe for the Wilcox & Biondi paper. More appropriate developmental reviews of NIRS are Lloyd-Fox et al 2010, Gervain et al. 2011, Minagawa-Kawai et al. 2008)
-section 2.2: What AuSTIN sentences are needs to be explained.
-section 2.2: Similarly, what vocoders are and how they work is not necessarily known by the broad readership of the journal. This needs to be explained and illustrated.
-section 3.1.1: what is meant by "trial type"? define
-section 3.1.2: "We used the term 'task performance' instead of 'speech intelligibility'" - nevertheless Figure 3B says speech intelligibility as does section 3.2

1st Author Response Letter

Response to comments from Editors and Reviewers:

Dear Editor and three reviewers, We really appreciate the comments from you all regarding our study and manuscript. In this revision, we have addressed all the comments and provided point-to-point responses. For clarity, we have labelled the three reviewers as R1, R2, and R3. As there were a few overlapping comments, we responded in detail to one and cross-referenced them for the rest to avoid repetitions. To help reviewers navigate the responses, we used hyperlinks for the cross-references within this document. Please see our responses below.

Comments from Reviewer 1

Review of CRNEUR-D-21-00079: Investigating effortful speech perception using fNIRS and pupillometry measures

R1. Summary:
This study used fNIRS and pupillometry to evaluate effortful listening in normal-hearing participants. Correlations used to evaluate the relationship between the two objective measures and the behavioral measures of task performance and task difficultly. Both the fNIRS and pupillometry measures were correlated with the self-reported task difficulty levels and the task performance levels,

but the direction of the relationship varied between the two objective measures of listening effort. Results suggested that both fNIRS and pupillometry are valid indicators of task demands and listening effort, but that protocol differences precluded a direct comparison between these two measures.

This is an interesting and timely study given the growing body of literature around objective measures of listening effort. I am unable to comment on the validity of the puillometry study design and analysis, so I will focus my comments on the fNIRS portion of the study. Mainly, it is unclear why certain pre-processing parameters were chosen for the fNIRS data, as well as the choice for statistical analyses. The processing and subsequent analysis of the fNIRS data as outlined in this paper is in conflict with the recent Best Practices for fNIRS Publications (Yucel et al., 2021). I have outlined some more specific concerns about the processing and analysis below.

R1. Major Comments:

1. Many of the pre-processing steps used for these fNIRS data disregard the best practices for fNIRS data as outlined by Yucel et al (2021). I recommend that the authors re-analyze the fNIRS data using the agreed-upon standard practices in the field. There are a few examples of discrepancies between standard processing and analysis approaches for fNIRS detailed listed below:

R1-C 1. The authors describe using a scalp-coupling index (correlations between two wavelengths in the heartrate range) to remove channels with poor signal quality. Why was a correlation of 0.35 chosen as the threshold? That seems like an oddly specific number, which is also a large departure from the Pollonini et al. (2014) recommendation of ~0.7.

Response: We appreciate this comment. In the revision (section 4.1.3), we have now included: 'A lower cut-off threshold was chosen here compared to the recommendation of 0.75 in Pollonini et al. (2014), for two 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, Sobczak, McKay, & Litovsky, 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 (p50) of SCI values were generally above 0.8, and the 25 percentile (p25) 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.' We have now included this information in the supplementary materials (please see S2. Fig).

- 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. doi:10.1016/j.heares.2013.11.007
- Sato T, Nambu I, Takeda K, Aihara T, Yamashita O, Isogaya Y, et al. Reduction of global interference of scalphemodynamics in functional near-infrared spectroscopy using short distance probes. NeuroImage. 2016;141:120–32. pmid:27374729
- 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.

R1-C 2. There is no rationale for choosing HbC as the main variable of interest rather than HbO. It would be helpful if the authors could describe their rationale in more detail.

Response: In the revision, we have now explained the rationale for running statistics on HbC amplitudes in section 4.1.3 (fNIRS data analysis): Further statistics were conducted on DHbC amplitudes for two reasons. First, DHbC amplitudes combined information from both DHbO and DHbR measures. Running statistics on one (DHbC) not only revealed information from both measures but also reduced the complexity of reporting results from both measures, separately. Second, DHbC responses have revealed changes in neuronal activity in the prefrontal cortex related to mental effort

(Ayaz et al., 2012; Liang, Getchell, & Shewokis, 2016; Nazeer et al., 2020; Rovetti, Goy, Pichora-Fuller, & Russo, 2019).

• 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.

• Liang, L.-Y., Getchell, N., & Shewokis, P. A. (2016). Brain activation in the prefrontal cortex during motor and cognitive tasks in adults.

• 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. Journal of Neural Engineering, 17(5), 056025.

• 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. Doi:10.1177/2331216519886722

R1-C 3. It is unclear why only the first two PCs were included in the short channel analysis. Did the authors try including different numbers of PCs before deciding on 2 PCs?

Response: As recommended by Noah et al. (2021) and Zhou et al. (2020), using the first and second PCs based on short-channel recordings can robustly remove non-neural components from the regular fNIRS channel. Therefore, we also included two PCs for GLM in the current study. In the revision, we added references to both papers.

• 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.

• 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.

R1-C 4. The band-pass filter upper cut-off frequency that was used is extremely low at 0.09 Hz. Yucel et al cautions setting this cut-off frequency substantially below 0.5 Hz as it can result in the removal of fluctuations in the brain response of interest. Indeed, the fNIRS time-series data as shown in Figure 5 has an overly filtered and smoothed appearance, which may have obscured variations in the peak response. This would not be a major issue if the time-series data was used simply for visualization of the temporal characteristics of the HB change and signal quality, but the authors use a block-averaging approach to the analysis, so the filter parameters could have a large impact on the peak and peak amplitude within the 5 sec averaging window.

Response: We appreciate this suggestion. Originally, we set the band-pass filter cut-off frequencies at [0.01 0.09]Hz according to the recommendation in Pinti et al. (2019) to reduce the effect of respiration and heartbeats on neuronal activity. However, we understand the concern raised about our data, and we re-analysed the data by setting the filter cut-off frequencies at [0.01 0.5]Hz. We also re-calculated DHbC amplitudes, and re-computed the statistics and repeated-measure correlation. Please see the updated Figure 5 and Figure 6. We agree with the reviewer that the updated fNIRS time-series data has a less smoothed appearance. However, the statistical results comparing the two band-pass filter cut-offs, i.e., [0.01 0.09]Hz and [0.01 0.5]Hz, are relatively comparable. Hence, the filter parameters, in this case, did not impact the previous conclusions.

• Pinti, P., Scholkmann, F., Hamilton, A., Burgess, P., & Tachtsidis, I. (2019). Current status and issues regarding pre-processing of fNIRS neuroimaging data: an investigation of diverse signal filtering methods within a general linear model framework. Frontiers in human neuroscience, 12, 505.

R1-C 5. There is no rationale for why ART ANOVAs were chosen for fNIRS analysis. Was there a violation of the parametric statistic assumptions? Also, why did the authors choose to a block-averaging approach rather than a GLM approach to data analysis? Using a GLM approach would also allow the authors to regress out the "button press" from the fNIRS recording (assuming the button press timing was recorded for each person).

Response: We appreciate this comment. Indeed, our fNIRS amplitude data were not normally distributed. In the revision, we included this sentence 'ART tests were conducted because fNIRS measures were not normally distributed, nor were their variances spherical.'

We appreciate the suggestion of using a GLM method. However, we did not use this method for two reasons. First, due to the differences in experimental conditions, we came across the challenges of applying a different HRF per condition. Second, according to Luke et al. (2021), GLM and averaging analyses of fNIRS data recorded in auditory tasks generated the same group-level experimental conclusions. Therefore, we decided to use the block-average results.

- Luke, R., Larson, E., Shader, M. J., Innes-Brown, H., Van Yper, L., Lee, A. K., ... & McAlpine, D. (2021). Analysis methods for measuring passive auditory fNIRS responses generated by a block-design paradigm. Neurophotonics, 8(2), 025008.

R1-C 6. Correlation analyses were used to test the hypotheses; correlation plots are shown in Figures 3 and 6. These plots do not appear to represent a strong relationship between these two metrics. Although these correlations are reported as statistically significant, the visual relationships between these metrics are not compelling. I realize that repeated measures correlations were used so that multiple data points per single participant could be included (without violating the assumption of independence), but I wonder if included 4 data points per person has artificially inflated the stats here.

Response: We appreciate this comment. We conducted the repeated-measure correlations because (1) they do not violate the assumption of independence, and (2) they estimate the common regression slope, the association shared among individuals, without ignoring intraindividual variances across conditions. We applied the method proposed in Bakdash and Marusich (2017), which has been widely implemented to determine the common within individual association for paired measures assessed on two or more conditions for multiple individuals.

In the revision, we included the sentence 'Rmcorr reveals the common regression slope, the association shared among individuals, without the violation of the independence of observations.' To further clarify this, in Figure 3 caption, we included the sentence that 'The orange lines in panels (E) and (F) indicate the common association between pupillometry measures and behavioral measures among individuals.' In Figure 6 caption, we included the sentence that 'The orange lines in panels (C-F) indicate the common association between fNIRS measures in two ROIs and behavioral measures among individuals.'

- Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. Frontiers in psychology, 8, 456.

Comments from Reviewer 2

R2. Summary :

The authors investigated the effects of task difficulty during speech processing on physiological measures of effort, including pupillometry and fNIRS. They found that pupillometry measures correlated with self-reported task difficulty and task performance, while activity in the left IFG and AC correlated with task difficulty and performance, albeit in opposite directions. Pupillometry measures did not correlate with fNIRS activity.

I find the pair of tasks interesting and the methods of each study relatively sound. However, I am confused as to why the authors are attempting to combine the studies, as the tasks themselves tap very different cognitive constructs with difference metrics of behavior, so it is unsurprising that the results do not correlate with each other. The paper is even laid out as two separate papers. Thus, I think this paper would make more sense as two separate manuscripts. I also have concerns about the operational constructs for motivation in the study. My more specific comments are found below.

Response: We appreciate this comment. We have a compelling reason for combining the two studies into a single paper. Over the years, pupillometry has become fairly accepted as a normative approach to measuring listening effort, or task engagement, by numerous labs including our lab. Pupillometry is not, however, a method that can be easily translated to special populations including young children. We were interested in designing an experimental paradigm that would be comparable across the two approaches (pupillometry and fNIRS) in order to ascertain whether fNIRS can provide a measure of effortful speech perception in a manner that is similar to that seen with pupillometry.

Study 1, using pupillometry measures, established that this newly designed stimulus paradigm can result in a systematic change in listening effort with varying task demands. Next, we implemented this paradigm using fNIRS. We acknowledge that there were a few task differences between study 1 and study 2, and have discussed the limitations in detail in text (please see section 5.3). These differences were mainly to avoid articulation-related artifacts in both frontal and temporal regions when measuring fNIRS response related to speech processing. Despite the differences, we believe that the two paradigms were similar to certain degrees, e.g., using the same stimuli, similar task difficulties in the corresponding conditions, and both tapping speech processing. By comparing fNIRS and pupillometry measures from the same individuals, we concluded that fNIRS measures in the LFC in the current study might reveal speech processing, rather than listening effort. Our findings are important, as previous studies have mainly identified LFC activity as being related to effort. Whereas, our results suggest that the relation between LFC activity and effort depends on experimental manipulations. Thus, if fNIRS is to be harnessed as an objective approach for assessing listening effort, comparing methods, brain regions and data outcomes with an established method such as pupillometry can be highly informative.

Please also see our detailed response to the next comment (R2. Introduction) regarding our interpretation of fNIRS measures, and to a comment from R3 (R3. Summary) regarding the comparisons between fNIRS and pupillometry measures.

R2. Introduction The introduction is organized a little differently, e.g. a defense of many of the methods is found in the introduction well before the methods are presented. Nonetheless, I appreciate how thorough the authors were with their definitions and background in the introduction. I do have a concern regarding your references to motivation (here and in the Discussion). The motivation literature suggests that task performance can decline with increasing task difficulty, whether the participant has high motivation or not. This is even true in the context of listening effort. Thus, task performance is not a good proxy for motivation and should be avoided, especially since motivation was controlled in the current study. The only interpretation related to motivation might be diminished performance on the easy conditions, which was not found here.

Response: We agree with this comment that the relation between task demand, motivation, effort, and performance is complicated. In the introduction, we only meant to use performance to indicate motivation when task demand is high, as good performance requires high motivation in difficult conditions. We agree that when task demand is high but performance is low, motivation could be high or low. In the revision, we rephrased the sentences as ‘if task demands become too high, listeners may lose motivation and listening effort subsides’. In the revision, we followed your suggestion and discussed that the result of low responses in the LFC (IFG) may be related to speech processing (section 5.2), rather than a drop in motivation:

‘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 might 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, Wiggins, Anderson, Davies-Thompson, & Hartley, 2018; Wijayasiri, Hartley, & Wiggins, 2017; Wild, Yusuf, et al., 2012), 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.’

- 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.

- 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. doi:10.1016/j.heares.2017.05.010

- Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful Listening: The Processing of Degraded Speech Depends Critically on Attention. Journal of Neuroscience, 32(40), 14010- 14021. doi:10.1523/Jneurosci.1528-12.2012

In addition, the authors should use care when describing the locations that are being probed using fNIRS, as they did not collect MRI data on their participants so they cannot be certain what the anatomical correlates of their signal are. This is especially important when discussing specific sulci and gyri, as the authors are doing. The best that can be said about the current data is "frontal regions" and "parietal regions".

Response: We appreciate the comment about fNIRS’ limited spatial resolution. In the revision, we replaced IFG as the lateral frontal cortex (LFC) and acknowledged this point in the introduction “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.”

R2. Methods and Results: Please report your data in units that are more easily interpreted, rather than percentage. For example, the number of trials accepted should be in trials, not percentage. The same with number of channels and number of blocks excluded in the fNIRS dataset.

Response: We appreciate this comment. In the revision (section 3.1.3), we rephrased the number of trials excluded as '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 conditions, respectively.'

In the revision (section 4.1.3), we rephrased the numbers of channels included as '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.'

Again, the task designs between your pupillometry study and fNIRS study are completely different and the nature of how participants responded reflects different cognitive constructs. For the pupillometry study: Participants listened to a single sentence, and then were asked to either repeat the sentence back (if it was grammatically correct) or rearrange the words to make the sentence grammatically correct (in the case of the shuffled sentences). This is a listening effort, expressive language, and grammaticality task. In contrast, in the fNIRS study, the participants listened to five sentences and then were prompted with a sentence written on the screen and asked whether the sentence was one of the previous five sentences. In the shuffled condition, the 5 sentences were presented shuffled, but the prompt was not. This is clearly a working memory task with different levels of encoding difficulty. The two cannot be rightfully compared.

Response: Please see our response to the comment above (R2. Summary).

R2. Discussion:

Again, the results cannot be interpreted in the context of motivation, because motivation was controlled in this study and the results do not suggest, by themselves, that motivation played a role.

Response: Please see our response to the comment above (R2. Introduction).

The deactivation of the left neural regions could also be due to a difference in the neural dynamics at the whole brain level (i.e., more recruitment of other regions not recorded here). The authors should mention that.

Response: We appreciate this comment. In the revision (section 5.3), we have now included the discussion: '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 & Brisson, 2014; Winn, Wendt, Koelewijn, & Kuchinsky, 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 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 may reveal 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.'

Comments from Reviewer 3

Review of manuscript nr. CRNEUR-D-21-00079 entitled "Investigating effortful speech perception using fNIRS and pupillometry measures".

R3. Summary

This paper introduces a new paradigm to measure processing effort while normally hearing adult participants listed to various degraded speech conditions. There are a few methodological and other issue to be clarified, but the study is otherwise sound and the question is interesting.

My major concern is the interpretation / framing. The authors present the study as one that measures processing effort. While this may indeed be true for pupillometry, all the data clearly shows that for NIRS they simply measure auditory processing rather than processing effort (see point 6 below). This in itself is not a problem, but the paper should clearly be reframed to reflect this. Such a change would clearly take away from the novelty of the contributing, as there are many studies of speech perception with NIRS. However, the authors simply do not have grounds to claim that their NIRS study manages to dissociate processing effort from processing more generally.

Response: We appreciate this comment. In the revision, we have taken the suggestion, discussed, and concluded our results that fNIRS measures in the LFC in the current study may reveal speech processing rather than effort. Please see our detailed response to the comment below (R3-C 6). We acknowledge that there were multiple differences between the paradigms in the two experiments (please see section 5.3, limitations). Still, we believe that our study design and findings contribute to our knowledge of fNIRS measures in the LFC. Please see our detailed response above to a comment from R2 (R2. Summary).

R3. Major concerns:

R3-C 1. The Introduction very briefly and passingly discusses lateralization of speech processing. This is an issue with a huge literature that the paper doesn't seem to take into account. The processing of prosody, for instance, is lateralized to the right hemisphere and recent work suggests that speech processing may be more bilateral than previously believed (see Poeppel 2014 Current Opinion for a review).

Response: We appreciate this comment. In the revision, we rephrased the sentences in the introduction as "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."

In the discussion section (section 5.2), we also rephrased our sentences as "We were specifically interested in the left AC as previous studies have demonstrated markers of speech intelligibility in the left AC. However, our results did not find a significant difference between the two hemispheres. Poeppel (2014) proposed that speech processing might be less leftlateralized 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, Wise, Dresner, & Scott, 2007; Okada et al., 2010; Wild, Davis, & Johnsrude, 2012)."

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

- 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.

- Obleser, J., Wise, R. J., Dresner, M. A., & Scott, S. K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. Journal of Neuroscience, 27(9), 2283-2289. doi: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. Cerebral Cortex, 20(10), 2486-2495. doi:10.1093/cercor/bhp318

- Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful Listening: The Processing of Degraded Speech Depends Critically on Attention. Journal of Neuroscience, 32(40), 14010- 14021. doi:10.1523/Jneurosci.1528-12.2012

R3-C 2. I don't quite follow the logic of this prediction in the Introduction: "We also considered the possibility that fNIRS measures in the left AC and IFG and pupillometry measures might yield different outcomes related to our stimulus paradigm. Such a finding may suggest that protocol differences in conducting fNIRS and pupillometry data collection make it difficult to directly compare these listening effort measures."

The study explicitly says and sets out to test the *same* paradigm with NIRS and pupillometry. How could there be protocol differences then? Also, this prediction undermines the study objectives. If the paper is meant to be an empirical test of whether NIRS can be used to measure processing effort and pupillometry is an established measure of processing effort, then if the two are correlated, then the study can conclude that NIRS can also be used as a measure of effort. If the two end up not being correlated, then the conclusion needs to be that NIRS is not an appropriate measure of effort. This conclusion cannot be preempted by saying that NIRS may still be a measure of effort, albeit one that is uncorrelated with pupillometry. If this can be true, then there is no point running this study.

Response: We apologize for the confusion. We were being straightforward about the fact that, while we did everything possible to design the study such that there would be an opportunity for discovering parallels between pupillometry and fNIRS, there were some limitations that precluded identical methods from being used. Perhaps we were over-zealous in acknowledging limitations before examining existing effects. In the revision, we rephrased the prediction as 'We also considered the possibility that fNIRS measures in the left AC and LFC and pupillometry measures might yield different outcomes. 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.

Accordingly, in the revision, we have now discussed our results by suggesting that fNIRS measures in the current study may reveal speech processing rather than effort. Please see our detailed responses to the comment below (R3-C 6).

R3-C 3. What motivates the experimental manipulations? Vocoding and interruption create degraded, challenging stimuli at the signal / auditory level, whereas shuffling does not affect the signal, but rather creates a challenge at the linguistic level. These conditions are thus not comparable and it is not clear what motivates these choices. Relatedly, while the vocoder condition is ecologically valid and highly relevant given that it simulates the signal cochlear implant users hear, the motivation for the interrupted and the shuffled conditions are less clear.

Response: We appreciate this comment. In the revision (section 2.2), we have further explained the rationale for shuffling and interrupting the sentences: "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 (but not spectral) information of 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 sentences in quiet is effortless for NH hearing adults, hence resulting in ceiling performance and minimal pupil dilation (Zekveld & Kramer, 2014). Second, sentences were shuffle-vocoded at the word level to mimic the scenario in which hearingimpaired listeners are around multiple persons, and they may confuse words from different people from time to time but have to fill the gap and follow the conversations."

Despite the differences in syntactic processing, the significant correlations between pupillometry measures and the two behavioral measures, i.e., self-reported task difficulty and task performance,suggest that pupillometry measures revealed variances in the effort. In the two comparable pairs, i.e., vocoded-interrupted vs. vocoded conditions, and shuffledvocoded vs. shuffled conditions, we found significant differences in both behavioral and pupillometry measures. However, no such differences were found in fNIRS measures. These results further supported that, unlike pupillometry, fNIRS measures in the current study revealed speech processing rather than effort. We have included this in the discussion (please see section 5.3).

- 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. doi:10.1111/psyp.12151

Additionally, more details about the stimuli will be necessary, e.g. a list of the sentences in an appendix / supplementary material, figures illustrating the spectrograms of the stimuli etc.

Response: We appreciate this suggestion. We have now attached the list of AuSTIN sentences used for the two studies. Please see the supplementary information which includes both the original list of

AuSTIN sentences in Australian English and the adapted sentences to North American English. We also included a figure (S1. Fig) to illustrate sentences in the four different conditions.

R3-C 4. The logic of the re-ordering task in the shuffled condition is not clear. Would asking participants to repeat back exactly what they heard not test better their speech perception accuracy? Relying on a reconstruction brings in linguistic knowledge (and other, related effects like word frequency, bi-word co-occurrence etc.) very heavily.

While participants certainly need to rely on such mechanisms in the other two conditions as well to some extent, the task is make a lot more complex and these mechanisms are made to be a lot more explicit in the re-order task. Not to mention that there may be ambiguity in this task (since the words Robin, kissed and Alex can be grammatically reconstructed as Robin kissed Alex and Alex kissed Robin, thus there are two, and not a single correct answer in some cases, at least potentially, further complicating the task).

Response: We appreciate this comment and agree that re-ordering the sentence increased the complexity of the task. By shuffling the sentences, we try to simulate the scenario in which hearing-impaired listeners may hear words from different people and have to fill the gap to follow the conversation. However, we do acknowledge that it is not a perfect simulation. The AuSTIN sentences we used consisted of 5-6 words, in the structure of 'subject-verbcomplement/object' or 'subject-verb-adverb'. Please see the supplementary list of sentences used. In the shuffled conditions, only the last 3 words that quite often were the complements were shuffled. For instance, participants might hear 'He LOCKED CAR the DOOR' instead of the original sentence 'He LOCKED the CAR DOOR'. We did not expect the ambiguity described here.

R3-C 5. In Study 1, despite the significant correlations between the pupillometry measure and the other two measures, self-reported effort and performance, qualitatively, there is a slight difference between them, in that the self-reported effort measure and performance are lowest/best for the shuffled condition, followed by the vocoded condition, whereas this pattern is reversed numerically for the pupillometry data and there seems to me no statistical difference.

Response: Thank you for pointing this out. We have now acknowledged this in our manuscript (section 3.2) as follows: 'To summarize …. the self-reported task difficulty increased and task performance decreased in the order of, shuffled, vocoded, shuffled-vocoded and vocodedinterrupted 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.' The significantly greater pupil response in the shuffle-vocoded and vocoded-interrupted conditions than in the vocoded and shuffled conditions contributed to the significant correlations with behavioral measures.

R3-C 6. As the most important concern, it is not clear whether NIRS measures processing or processing effort in this study. Indeed, the fact that activation was smallest in the vocoded interrupted condition suggests that NIRS is actually picking up on processing itself, which is the weakest / poorest in this condition. The directions of the significant correlations between NIRS and the behavioral measures also point in the same direction. The lack of correlation with pupillometry further suggest this.

To put it differently, if one was to use NIRS to measure auditory processing in these different conditions, one would proceed exactly as the authors did here (e.g. see Cabrera and Gervain 2020 for a NIRS study on newborns' processing of vocoded speech stimuli). There is nothing in the task that makes the NIRS responses measure effort rather than processing itself. The explanation in terms of dropping motivation should have been also visible in Study 1, so it cannot be used to justify the NIRS results.

Response: We appreciate this comment and agree that fNIRS might be measuring speech processing rather than effort. We also took the advice and deleted the discussion about low responses in the vocoded-interrupted condition being related to motivation. Instead, in the revision (section 5.2), we acknowledge the possibility of LFC revealing speech processing. For the detailed revision, please also see responses to a comment from R2 above (R2. Introduction)

In the revision (section 5.3), we have now included the discussion that "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."

- 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.

- 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. doi:10.1016/j.heares.2017.05.010

- Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful Listening: The Processing of Degraded Speech Depends Critically on Attention. Journal of Neuroscience, 32(40), 14010- 14021. doi:10.1523/Jneurosci.1528-12.2012

R3. Minor issues:

R3-C 7-The reference Wilcox and Biondi 2015 is not the best when referring to the use of NIRS in speech perception, language acquisition and cochlear implantation. There are many more specific references, e.g. Saliba et al. 2016, Gervain & Cabrera 2020 etc. -Similarly, for the sentence "such as children and infants (see reviews by: Ferrari & Quaresima, 2012; Quaresima, Bisconti, & Ferrari, 2012; Vanderwert & Nelson, 2014; Wilcox & Biondi, 2015)", the reviews are not very appropriate, except maybe for the Wilcox & Biondi paper. More appropriate developmental reviews of NIRS are Lloyd-Fox et al 2010, Gervain et al. 2011, Minagawa-Kawai et al. 2008)

Response: We appreciate the recommendations. In the revision, in the first paragraph, we included references to two recent review and perspective articles. The sentence reads as '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, Kiran, & Tager-Flusberg, 2020).'

In section 1.2, we referred to original research articles. The sentence reads as 'fNIRS has been implemented to examine auditory perception and cognitive functions in populations that are challenging for fMRI such as children and infants (Cabrera & Gervain, 2020; Cristia et al., 2014; Lloyd-Fox et al., 2019; Lloyd-Fox et al., 2014; Mao et al., 2021)'.

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

- 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. doi:10.1044/2020_AJSLP-19-00050

- Cabrera, L., & Gervain, J. (2020). Speech perception at birth: The brain encodes fast and slow temporal information. Science advances, 6(30), eaba7830.

- 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. Developmental Science, 17(4), 628-635. doi:10.1111/desc.12160

- 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. Developmental Science, 22(5). doi:ARTN e12817 10.1111/desc.12817

- 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. Scientific Reports, 4. doi:ARTN 4740 10.1038/srep04740

- 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. Scientific Reports, 11(1). doi:ARTN 24006 10.1038/s41598-021-03595-z

R3-C 8 -section 2.2: What AuSTIN sentences are needs to be explained.

Response: Thank you for this comment. In the revision, we included this information 'Stimuli consisted of a subset of AuSTIN sentences (Dawson, Hersbach, & Swanson, 2013) with five or six words, with 3-4 keywords each, recorded by an American female speaker. AuSTIN sentences are modelled based on the simple and short Bamford-Kowal-Bench (BKB) sentences (Bench, Kowal, & Bamford, 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.'

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

- Dawson, P. W., Hersbach, A. A., & Swanson, B. A. (2013). An adaptive Australian Sentence Test in Noise (AuSTIN). Ear and hearing, 34(5), 592-600. doi:10.1097/AUD.0b013e31828576fb

R3-C 9 -section 2.2: Similarly, what vocoders are and how they work is not necessarily known by the broad readership of the journal. This needs to be explained and illustrated.

Response: We appreciate this comment. In the revision, we included this information 'In the vocoded condition, the sentences were processed in AngelSimTM (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, Zeng, Kamath, Wygonski, & Ekelid, 1995). The vocoded sentences were to mimic the spectrally degraded input from cochlear implants, with the envelope information being transmitted but temporal fine information being compromised.'

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

R3-C 10 -section 3.1.1: what is meant by "trial type"? define

Response: Thank you for pointing this out. We apologize for the typo. The sentence was supposed to be "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." We have fixed this in the revision.

R3-C 11 -section 3.1.2: "We used the term 'task performance' instead of 'speech intelligibility'" - nevertheless Figure 3B says speech intelligibility as does section 3.2

Response: Thank you for pointing this out. We have fixed this problem. Please see updated Fig 3.

Accept Letter

Dear Dr. Zhou,

Thank you for submitting your manuscript to Current Research in Neurobiology.

I am pleased to inform you that your manuscript has been accepted for publication.

My comments, and any reviewer comments, are below.

Your accepted manuscript will now be transferred to our production department. We will create a proof which you will be asked to check, and you will also be asked to complete a number of online forms required for publication. If we need additional information from you during the production process, we will contact you directly.

We appreciate and value your contribution to Current Research in Neurobiology. We regularly invite authors of recently published manuscript to participate in the peer review process. If you were not already part of the journal's reviewer pool, you have now been added to it. We look forward to your continued participation in our journal, and we hope you will consider us again for future submissions.

CRNEUR aims to be a unique, community-led journal, as highlighted in the Editorial Introduction. As part of this vision, we will be regularly seeking input from the scientific community and encourage you and your co-authors to take the survey.

We would also like to invite you to take part in our CRNEUR Author Question & Answer (Q&A), which could get published alongside your article and help to promote it. We suspect you might have an interesting story of perseverance or team work that was required for the research study to complete, or a diversity of perspectives that you might share, as a way of inspiring others about neuroscience.

Kind regards,
Christopher I. Petkov
Editor in Chief
Current Research in Neurobiology

Editor and Reviewer comments:

Reviewer 2: The authors have responded to all of my comments. I appreciate that the manuscript has been reframed to illuminate the differences in pupillometry and fNIRS measures, and that the reviewer responded not only to my comments but the excellent comments from the other reviewers.

-------- End of Review Comments --------