

Dear recommenders,

The authors thank you for your careful review of our registered report. This round of review allowed us to clarify several points. We have addressed the issues raised and provide detailed responses. Changes made to the report are highlighted in yellow.

**You consider the inconsistent results in prior literature to be due to the discrepancies in higher functioning levels of the two ASD samples (e.g. IQ). In the introduction, the discrepancy between the ASD groups of each study is one of the main motivations for further examination into the vocal-sound processing in ASD. It is still not totally clear how you will control for higher functioning level in your ASD sample. It is plausible that with a larger sample of ASD participants, you will cover a broader variance in higher-level functioning, culminating in a result which is more applicable to the ASD population as a whole. If this is the case, you would need to consider your sample size to account for the wide range of IQs in the population. However, it is also plausible that without controlling for higher level function (in this case, IQ), you may recruit closer to one end or the other which could significantly impact interpretation of the result, falling into the same trap as the prior literature. If you indeed plan to control for variables such as IQ to overcome discrepancies in the prior literature, a more substantial reconsideration of planned analyses would be necessary.**

The authors thank the recommenders for this comment, which allows us to clarify this point. Discrepancies between the two studies regarding the characteristics of the ASD samples are indeed one of the motivations for this study. However, as stated at the end of the introduction, we plan to investigate associations between voice processing and individual participants' characteristics, such as IQ, with exploratory correlational analyses, that do not need to be detailed at this stage of the registered report. We believe that recruiting a larger sample will allow us to cover a broader variance in IQ, but also in other autism-related measures such as the Autism-spectrum Quotient (AQ), so that we will be able to explain variability in voice processing abilities to a greater extent than in previous studies.

*"Exploratory analyses will be conducted in order to investigate links between voice brain processing and individual participants' characteristics (e.g. IQ, Autism-spectrum Quotient; AQ, ...)."*

**You have included a memory task to determine if there is a difference in attention to the stimuli between the two groups. This addition seems important, but you should add a hypothesis and run a power analysis to ensure you have the power for your planned one-way ANOVA.**

The authors thank the recommenders for this comment. However, this task is only intended to control for potential between-group differences in attention during the voice localizer task. It is a control measure, just like IQ, Autism-spectrum Quotient, etc. Thus, the "attention/memory" task is not part of the global aims of the study and therefore the authors consider that no hypothesis are associated to this task and related statistical tests do not require power analyses.

**On that note, it seems like a t-test would be more appropriate for this analysis as there are only two groups. Could you either clarify why an ANOVA is the correct route or change the manuscript accordingly.**

A mistake was made in writing the manuscript. A t-test is indeed the appropriate analysis. The authors thank the recommenders for noticing this mistake and the manuscript has been modified accordingly:

*“Global scores from the “attention/memory” task will be compared between groups using a t-test.”*

**Furthermore, more information is necessary with regard to the task: how many sounds will be tested? Will the participants be told of the memory task prior to the fMRI data collection? If so, how might the knowledge of a memory task influence the processing of the stimuli during the scan? We believe it would be better to inform your participants of the memory task prior to the scan, as a post-scan behavior difference could explain any functional differences. However, informing the participants prior to the scan changes the interpretation of a potential functional difference between groups. The interpretation cannot be of an automatic processing of early voice processing, but a task-relevant processing of early voice processing.**

We thank the recommenders for allowing us to clarify these points. The authors believe that it is better not to inform the participants of the memory task prior to the scanning session, in particular as this was not done in the two previous studies. In fact, and as stated by the recommenders, informing the participants of the memory task prior to scanning will influence the way they process the sounds. Since we want to investigate automatic voice processing, and as the memory task is only intended to control a posteriori for differences in attention levels, we will not inform the participants about this task beforehand.

Manuscript has been modified accordingly:

*“An incidental “attention/memory” task will be administered to the participants after the scanning session. During this task, participants will be presented with some of the sounds displayed during the experiment. More precisely, stimuli consist of 10 vocal and 10 non-vocal sounds, arbitrarily drawn from the set of sounds displayed during the task. For each sound, the participants will be asked to indicate whether they remember hearing it during the scanning session. This task aims to monitor and detect differences in attention level during the task. In addition, participants will be asked to rate their level of engagement with the sounds during the voice-localizer task on a 5-points Likert scale.”*

**In the Design Table you should add/clarify some information. For example, you should state the power level you used (it's in the text but should be there as well). Also, under Rationale for deciding sensitivity you only write  $p < 0.02$  - you should support the p-value with a rationale by moving the information about previous effect size from Sampling plan into here. You already explain in the text why this is the effect size you use, so you would not need to add any extra this information.**

**It is important that you also remove their references to exploratory analyses in the study table. The mention of the individual differences exploratory analysis in the introduction is an exception. For readability, you need to explain the collection of data that will only**

support exploratory analyses (i.e. the individual evaluations) in the introduction. Other exploratory analyses are acceptable at Stage 2, but then they are clearly labelled and will be scrutinised accordingly. With your current design (which does not include an equivalence test), if your chi-square test finds no significant difference, the main message of your Stage 2 would be that the null hypothesis cannot be rejected.

The authors thank the recommenders for these comments concerning the study design table which has been modified accordingly:

Question	Hypothesis	Sampling plan	Analysis Plan	Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis	Interpretation given different outcomes	Theory that could be shown wrong by the outcomes
Is the ASD population characterized by a different proportion of individuals showing a specific response to vocal sound than the non-ASD population?	We predict a lower proportion of individual TVA activation in the ASD group than in the non-ASD group	Power analysis was conducted using an effect size estimated from the literature in order to estimate the minimum required sample size. Results indicated a minimum required sample size of n = 24 per group with a power level of .90.	A $\chi^2$ test will be conducted in order to compare the proportion of individuals showing a TVA activation in the ASD and non-ASD groups.	Considering the effect size inferred from the literature ( $w=11.81$ ), power analyses indicated that the study is powered enough to detect a group difference at the $p < .02$ alpha level.	A different proportion of individuals showing a TVA activation in the ASD group would suggest that at least a subset of ASD individuals does not process the vocal sounds in a typical way.  The failure to reject the null hypothesis will be interpreted as an absence of evidence towards either the null or the alternative hypothesis (i.e., the null hypothesis cannot be rejected).	The rejection of the null hypothesis would suggest that ASD individuals suffer from low level deficits in voice processing, which may eventually lead to higher order social dysfunction.  If the statistical test fails to reject the null hypothesis, no strong conclusions can be drawn considering an (absence of) impairment of voice processing in ASD.

We thank the recommenders for their comments, and we hope that we successfully clarified the issues raised during this round of review.

Sincerely,

Raphaël Gautier