Supplementary Material

Altered processing of communication signals in the subcortical auditory sensory pathway in autism

Stefanie Schelinski 1,2, Alejandro Tabas 1,2 & Katharina von Kriegstein 1,2

a Technische Universität Dresden, Faculty of Psychology, Chair of Cognitive and Clinical Neuroscience, Bamberger Str. 7, 01187 Dresden
b Max Planck Institute for Human Cognitive and Brain Sciences, Neural Mechanisms of Human Communication, Stephanstrasse 1a, 04103 Leipzig, Germany

Supplementary Methods

Participants

Medication. Two participants in the voice identity recognition and the vocal sound experiment (1 control, 1 ASD participant) and one control participant in the speech-in-noise recognition experiment took a histamine antagonist for allergies when needed. Three ASD participants (2 in the voice identity recognition experiment and 1 in the speech-in-noise recognition experiment) took antihypertensive medication. Two ASD participants in the speech-in-noise recognition experiment took thyroid medication for hypothyroidism. None of the participants reported to have a neurological disease. In the voice identity recognition and the vocal sound experiment, three additional ASD participants were not included in the analysis due to incidental findings in an anatomical MRI-scan (for details see Schelinski, Borowiak, & von Kriegstein, 2016).

Recruitment and reimbursement. We recruited people with ASD via autism outpatient clinics and announcements in communities for people with ASD, such as self-help groups and online fora. We only included participants into the ASD group who could provide a clinical diagnosis. That means that independent clinical experts made the diagnoses of all ASD participants before participating in the studies. We recruited the control group participants from the participant database of the Max-Planck-Institute for Human Cognitive and Brain Sciences Leipzig. The database contains participants who have contacted the institute because they are interested in taking part in scientific studies or have been recruited in the context of other
studies. The database contains volunteers with e.g., different age ranges and different socioeconomic status or educational backgrounds. All participants received reimbursement for their participation.

Sample size. We attempted to maximise the number of clinical participants we could recruit in a given time (approximately one-year for each study). Since analyses in experiments 1 and 2 were performed over data previously collected within the scope of another project, the sample size was not calculated beforehand given the sample size and results shown in this report.

Experiments

Experimental procedure
The TD and the ASD group participants underwent all experiments under the same testing conditions and in the same experimental order. The MRI machine was the same for all 3 experiments located in the Clinic for Cognitive Neurology at the University Clinic Leipzig. During all three fMRI experiments participants wore hearing protection plugs in addition to MRI-compatible headphones (Mark II; MR confon, Germany). Initial sound levels were defaults in the MR confon system, similar for all participants and adjusted to a comfortable hearing level if needed.

Voice identity recognition experiment
Stimuli. All recordings were made in a soundproofed chamber under the same conditions and using the same equipment (condenser microphone, TLM 50; Neumann, Germany; preamplifier, Mic-Amp F-35; Lake People, Germany; soundcard, Power Mac 5; Apple Inc., USA), and the software Sound Studio 3 (Felt Tip Inc., USA) with a 48 000 Hz sampling rate and a resolution of 16 bits. Stimuli were postprocessed and root-mean-square (RMS) adjusted using Matlab (version 7.7; The MathWorks, Inc, Natick, Massachusetts, USA).

Speaker familiarisation. The speaker familiarisation was performed outside the MRI room on a laptop with headphones (HD 201, Sennheiser, Germany). Participants listened to 20 sentences from each of the three speakers organised in blocks (two blocks per speaker, one block containing 10 sentences). Next, pairs of sentences spoken by one or two of the three speakers were presented consecutively and participants indicated whether the sentences were spoken by the same speaker or by two different speakers. Answers were made by pressing a corresponding
button on the keyboard. In total, we presented 54 pairs of sentences (18 sentences per speaker). Visual feedback indicating whether the answer was correct (green cross) or wrong (red cross) was provided immediately after each response. We told the participants that this test served to learn the three speaker’s voices for a subsequent voice identity and speech recognition test. Results of this speaker familiarisation are presented and discussed elsewhere (Schelinski, Roswandowitz, & von Kriegstein, 2017).

Task familiarisation. After the speaker familiarisation, participants received task instructions for the fMRI experiment and were familiarised with the voice identity and the speech tasks (two practice blocks per condition). The order of the conditions within the task familiarisation was randomised across participants. In order to ensure that all participants understood the task equally well, the task familiarisation was repeated if a participant performed less than 70% correct in the practice trials.

Speech-in-noise recognition experiment

Stimuli. We generated the auditory stimuli from audio-visual recordings which were made from each speaker using a digital video camera (HD-Camcorder LEGRIA HSF100; Canon Inc., Tokyo, Japan) with an external directional microphone (Sennheiser Kondensator M. System K6; Sennheiser, Wedemark, German; 44100 kHz sampling rate, 16 bit resolution). The raw audio material was pre-processed (i.e., choosing best quality speaker recordings and first assembly) using Audacity software (version 1.3.5 beta, http://audacity.sourceforge.net). For further processing of the audio material we used Matlab software (version 8.2, The MathWorks, Inc, Natick, Massachusetts, USA). All audio stimuli were cut to have 50 ms of silence at onset and 150 ms of silence at offset. The audio stimuli from all six speakers were adjusted to the same RMS (RMS = 0.0765).

Speaker familiarisation. The six speakers were familiarised during an audio-visual training before the speech-in-noise recognition experiment. In this audio-visual training, participants learned to recognise voices, names, and faces of all speakers. The design from the audio-visual training was adapted from previous studies in which we investigated the influence of audio-visual speaker familiarisation on auditory-only perception (Schelinski, Riedel, & von Kriegstein, 2014; von Kriegstein et al., 2008). Voices of the six speakers were learned together with a visual stimulus in two different learning conditions (see below stimuli and design description for details). The two conditions were part of a different study, which will be reported
elsewhere. Within the scope of the current study we averaged the results for speech recognition for the noise and the no noise task condition over both learning conditions. There was a break of approximately 45 minutes between the audio-visual training and the actual scanning session in which participants were allowed to rest and were transferred to the MR-centre (10 min walk).

**Task familiarisation.** To familiarise the participants with the task and noise of the speech-in-noise recognition experiment, participants received the task instructions and performed example blocks before the fMRI-acquisition. The task familiarisation was repeated if a participant had more than one missed response in the practise trials. Sentences presented during the task familiarisation were not presented during the fMRI-experiment. To familiarise participants with the noise and task inside the MRI-machine, we additionally presented two training blocks (one block for the noise and one block for no-noise condition) once participants had entered the MRI-machine.

**Stimuli.** We used audio-visual recordings from the six speakers of the speech-in-noise recognition experiment. Stimuli used in the audio-visual training were not used during the speech-in-noise recognition experiment. We recorded the face of each speaker in 16:9/ XP+ with a resolution of 1920 x 1080 pixels. Video recordings were edited using FinalCutPro (Video Studio). Videos were cut so that each speaker was shown 500ms before and after face-movement. Recordings were done against a uniform black background. We extracted static face images with closed mouth from each speaker from the video recordings.

**Experimental design.** The training consisted of at least two and maximal three learning and evaluation sessions. During the learning session, we instructed participants to memorise names and voices of six speakers. In each session, there were two learning conditions. In the ‘static face learning condition’, participants learned the voices and names of three of the speakers together with a static picture of their face. In the ‘moving face learning condition’, voices and names of the other three speakers were learned together with a video showing the speaker articulating. In one learning session, each speaker was presented for 15 times articulating 5-word sentences (15 trials x 3 speakers x 2 learning conditions = 90 trials in total). In each trial, the name of the speaker was presented for 1 sec, followed by a sentence spoken by that speaker and the simultaneously presented face (static or moving depending on the learning condition). Trials were presented in blocks of five trials in which the same speaker identity was presented (one block per speaker). In the following blocks, trials with different speaker identities were
randomly intermixed. There were two sets of speakers: speaker set 1 containing speakers A, B and C and speaker set 2 containing speakers D, E and F. Half of the participants learned speaker set 1 in the ‘static face learning condition’ and speaker set 2 in the ‘moving face learning condition’. The other half of participants learned speaker set 1 in the ‘moving face learning condition’ and speaker set 2 in the ‘static face learning condition’. Static face learning and moving face learning were presented separately. The order of the two learning conditions was balanced over the whole sample: half of the participants started with the static face learning, whereas the other half of participants started with the moving face learning. During the training eye movements and button press responses were recorded (Tobii Studio, Tobii Technology AB, Danderyd, Sweden).

Evaluation. Learning success was evaluated after each learning session. Learning and evaluation sessions were performed twice. If the participant did not reach a criterion of 75% correct performance after the second repetition, learning and evaluation were repeated a third time. In the evaluation, the voice of one of the six speakers was presented either followed by a written name or by a picture of a face from one of the three speakers learned in the corresponding learning condition. Participants had to decide whether the voice and the face belonged to same person or not. An audio-visual feedback of the correct combination was provided immediately after each response. In this feedback, the voice was presented together with the name and the static or moving (i.e., articulating) face respectively. In total, each evaluation session included 48 trials (8 trials x 6 speakers). In half of the trials, voices were followed by a name, in the other half voices were followed by the face.

Image acquisition

Structural MRI. MPRAGE parameters were: TR = 2.3 sec; TE = 2.98 ms; TI = 900 ms; flip angle = 9°; FOV = 256 mm x 240 mm; voxel size = 1 mm³ (isotropic resolution); 176 sagittal slices) with nonselective excitation and linear phase encoding. Magnetisation preparation consisted of a nonselective inversion pulse. For one participant we used a 12-channel head coil (with an identical MR-protocol), as the 32-channel head coil was too tight for the participant’s head size. MRI-acquisition time for the structural scan was 9 min 14 sec.
Data analyses

Regions of interest

*Voice identity and vocal sound experiments*. To test robustness of our results in the IC, we additionally created masks defined as spheres centered on coordinates reported in previous studies. We chose a radius of 4mm, because this yields a ROI size (268 mm$^3$) that comprises the sizes of functional clusters of ICs (146 mm$^3$ - 263 mm$^3$, reported in Sitek et al., 2019). We chose the coordinate reported in a pioneering fMRI study on auditory subcortical sensory pathway responses ($x = 6$, $y = -36$, $z = -10$, MNI space, SPM99; Griffiths, Uppenkamp, Johnsrudef, Josephs, & Patterson, 2001) and another one of a previous clinical study including patients with schizophrenia and major depression ($x = 6$, $y = -33$, $z = -9$, MNI space, reported in Talairach space in Gaebler et al., 2020; coordinate originally reported in Parsons et al., 2014). See Supplementary Figure 1 for an overlay of these masks with our initial mask from Sitek et al., (2019) used in the voice identity and the vocal sound experiment.

Control analyses

To control for potential effects due to differences in task difficulty, we include task performance as a regressor of no interest in all analyses where we find main effects of task (within-group comparisons) or group (between-group comparisons). To test whether results can be explained by task difficulty, we additionally perform analyses with task performance as a regressor of interest for conditions in which we do not find behavioural differences between the ASD and the control group.

*Design matrix for control analyses in the voice identity recognition experiment*. To account for group differences in performance in the voice identity recognition task, we included the performance difference (percent correct) between the speech and voice identity task as a covariate of no interest for the between group analyses for the contrast ‘voice identity task + speech task’.

*Generalised lower BOLD responses in ASD*. To test whether there is generalised lower BOLD response in ASD in contrast to controls, we additionally report results from pooled responses over both experimental conditions for each experiment (i.e., ‘voice identity task + speech task’, ‘vocal sounds + non-vocal sounds’, ‘speech task noise + speech task no noise’) for the IC and MGB ROIs.
Temporal signal-to-noise ratio for IC and MGB

The tSNR (e.g., Welvaert & Rosseel, 2013) is the ratio between the average signal and its variability along the duration of the data acquisition. High tSNRs result in better estimations of the BOLD response. tSNRs larger or in the range of 30 are sufficient to robustly detect effects of 0.5% in data of the duration of our experiments (Murphy, Bodurka, & Bandettini, 2007). We used the nipype confound toolbox (Gorgolewski et al., 2011) to estimate the tSNR in the realigned, registered and normalised data for each subject, experiment and ROI. Since we performed our measurements using the preprocessed data our estimations of the variance across the MRI sessions include task-related variance. This means that the reported tSNRs are lower bounds of the nominal values.

We additionally tested whether group differences in tSNR contribute to the results. For this, we computed the tSNR of the coregistered MRI data for each subject and experiment using nipype's native confound toolbox (version 1.1.2; Gorgolewski et al., 2011). We then used the left and right IC and left and right MGB to mask each of the tSNR images, and computed the average tSNR for each masked region.

Supplementary Results

Voice identity recognition and vocal sound experiment

Behavioural results

Vocal sound experiment. For the analyses of the behavioural data in the vocal sound experiment we used the absolute number of recalled sounds for each sound category. The numbers of sounds for each category presented in the experiment are unequal. Thus, we report direct group comparisons i.e., independent t-test for each sound category but did not perform repeated measures ANOVAs for the total numbers of recalled sounds. The ASD group recalled significantly less non-vocal and vocal sounds than the control group ($t(30) = 2.622, p = 0.014$). The ASD group recalled significantly less non-speech vocal sounds as compared to the control group after the fMRI scan ($t(30) = 2.396, p = 0.023$). In contrast, there were no significant group differences in the amount of recalled speech vocal sounds ($t(30) = 1.202, p = 0.239$; Supplementary Figure 5; Supplementary Table 4).

After neuroimaging data acquisition, we asked participants whether they fell asleep during the experiment. None of the participants reported to have fallen asleep during the fMRI
scan for a noticeable amount of time and all participants were able to recall sounds. Further, it is unlikely that there was a systematic effect of attention during the vocal sound experiment, since we found group differences for the number of recalled sounds in some but not all categories (see Supplementary Table 4).

fMRI results

Robustness of the fMRI effects for different ROI-definitions. In both, the voice identity recognition and the vocal sound experiment, we found significant group differences (i.e., control > ASD) in the right IC with a peak coordinate at x = 6, y = -31, z = -10. We tested whether the results are also present when using IC coordinates from previous studies (Supplementary Methods; Supplementary Figure 1). Using these masks, the results remained qualitatively the same; i.e., there was significantly greater BOLD response in the right IC in the control group as compared to the ASD group for the contrasts ‘voice identity task > speech task’ and the contrast ‘vocal sounds > non-vocal sounds’ (Supplementary Table 5).

Control analyses

No effect of head movements during MRI. The ASD and the control group did not differ significantly in the average amount of head movements (all p values > 0.1 in all three experiments; Supplementary Table 6; Schelinski et al., 2016).

Controlling for task difficulty differences between groups in Experiment 1 and 3. All analyses contained a regressor of no interest to control as far as possible for potential task difficulty effects in experiment 1 and 3. Furthermore we tested whether the IC-responses correlated with the amount of task performance in those tasks where there was no task difficulty difference between ASD and controls (i.e., speech task in the voice identity recognition experiment; speech task for the no noise and the noise condition in the speech-in-noise recognition experiment). The correlations were performed across both groups to maximise statistical power (Experiment 1 n = 32; Experiment 3 n = 34). We found no significant correlations in the right or left IC with none of the behaviourlal scores (all ps > 0.2 FWE corrected for the ROI).

Prompted by a reviewer, we additionally performed correlation analyses between IC-responses and task performance in all tasks (voice identity recognition and speech-in-noise recognition experiment). There were no significant correlations between task performance and BOLD-
responses in the right or left IC for none of the groups (all \( ps > .13 \); Supplementary Figure 7; Supplementary Table 8).

**Overall BOLD responses were comparable in the ASD and control group.** We also checked whether there is a generalised lower BOLD response in ASD. This was not the case: When we tested the unspecific contrasts of all conditions with sounds in contrast to the silence baseline, we did not find significant differences between groups and there was also no systematic difference at very lenient statistical thresholds (Supplementary Table 7).

**Good temporal signal-to-noise ratio for IC and MGB**

We observed generally large average tSNRs, ranging from 30 to 60, in the four ROIs (i.e., left and right IC and left and right MGB) in all three experiments (Supplementary Figure 3). These good tSNRs illustrate that our experimental setup was appropriate to assess subcortical nuclei responses.

We additionally tested for group differences in tSNR. There were no significant differences for the right and left IC and right and left MGB for the voice identity recognition and the vocal sound experiments (independent \( t \)-tests: all \( p \)-values > 0.1; Supplementary Figure 4). However, there were significant group difference for tSNRs in the speech-in-noise recognition experiment with the ASD group showing a lower tSNR as compared to the control group in the right IC (\( t(32) = 3.064, p = 0.004 \)), left IC (\( t(32) = 3.379, p = 0.002 \)) and the left MGB (\( t(32) = 2.740, p = 0.010 \)). We therefore checked if tSNR group differences would change the results in the speech-in-noise experiment by adding tSNR for each ROI as an additional covariate of no interest in the GLM. The results remained qualitatively the same (Supplementary Table 3).
Supplementary Figures

Supplementary Figure 1

Overview of masks used for the right IC in the voice identity recognition and the vocal sound experiment. (A) Masks provided in an independent atlas on the human subcortical auditory system (Sitek et al., 2019) (yellow, same as in Figure 2C). (B, C) Masks defined based on centre coordinates (green cross) reported in previous studies by Gaebler et al. (2020) (B, blue) and Griffiths et al. (2001) (C, red). P = posterior; S = superior; A = anterior; I = inferior; L = left; R = right; x, y, z = coordinates in MNI space.
Supplementary Figure 2

Coarse regions of the left and right IC (red and blue squares) and left and right MGB (yellow and pink circles) for creating ROIs for the speech-in-noise recognition experiment. The regions were created based on visual inspection of the MNI template. IC-L = left inferior colliculus; IC-R = right inferior colliculus; MGB-L = left medial geniculate body; MGB-R = right medial geniculate body; y = coordinate in MNI space.
Supplementary Figure 3

tSNR of the raw data across the four ROIs (i.e., left and right IC and left and right MGB) in the three experiments. Violin plots are kernel density estimations of the tSNR across participants. Crosses mark the average and standard error of the mean. Each grey circle is the average tSNR computed for one subject across all the voxels comprised in the ROI. tSNRs values were computed on the realigned, registered, normalised data before smoothing. tSNR = temporal signal-to-noise ratio; IC-L = left inferior colliculus; IC-R = right inferior colliculus; MGB-L = left medial geniculate body; MGB-R = right medial geniculate body.
Supplementary Figure 4
Kernel density estimation of the tSNR in the regions of interest of the ASD (red) and control (blue) groups. Each plot corresponds to one of the experiments included in the study. Each grey circle is one sample of each distribution, corresponding to the average tSNR in the ROI and experiment for one participant. Black error bars are the group mean and standard deviation.
Supplementary Figure 5

Total amount of recalled sounds after the vocal sound experiment (Schelinski et al., 2016; see also Supplementary Results and supplementary Table 4). The ASD group recalled significantly less non-vocal and vocal sounds than the control group. The vocal sound condition contained both speech and non-speech sounds. The ASD group recalled a comparable number of speech sounds but less non-speech sounds as compared to the control group. Error bars represent +/- 1 SE; * p < 0.05; ** p < 0.005; n.s. not significant.
Supplementary Figure 6

1st eigenvariate extracted at the statistical maximum (MNI-coordinate: x = 6, y = -31, z = -10) for the voice identity recognition experiment (A) and the vocal sound experiment (B) for the significant interactions between group (controls, ASD) and task condition (voice identity recognition experiment: voice identity task, speech task; vocal sound experiment: vocal sounds, non-vocal sounds).
Supplementary Figure 7

Control analyses: Scatter plots for the relation between task performance in the voice identity recognition and the speech-in-noise recognition experiment and BOLD-responses in the right and left IC for the respective task (averaged mean 1st eigenvariates extracted for the whole left and right ROI (Sitek et al., 2019); for the voice identity task, we extracted the 1st eigenariate for the coordinate for which we found a significant group differences, MNI coordinate: x=6, y=-10, z=-31). Each dot represents one participant in the control group (blue) or the ASD group (red), respectively.
## Supplementary Table 1

Overview of diagnostic scores in the ASD group in the voice identity recognition and the vocal sound experiment (A) (1) and the speech-in-noise recognition experiment (B).

### A Voice identity recognition and vocal sound experiment

| Diagnostic test | M (cut-offs for autism / autism spectrum) | SD |
|-----------------|------------------------------------------|----|
| **Participants as informant** | | |
| Interview | ADOS<sup>a</sup> (<i>n</i> = 15) | | |
| Social Interaction & Communication | 11.00 (12 / 7) | 2.78 |
| Social Interaction | 7.20 (7 / 4) | 1.97 |
| Communication | 3.80 (3 / 2) | 1.27 |
| **Parents as informant** | | |
| Questionnaire | SCQ<sup>b</sup> (<i>n</i> = 9) | 20.33 (15) | 5.70 |
| Interview | ADI-R<sup>c</sup> (<i>n</i> = 9) | | |
| Social Interaction & Communication | 36.22 | 8.04 |
| Social Interaction | 21.11 (17) | 5.09 |
| Communication | 13.89 (8) | 4.37 |

### B Speech-in-noise recognition experiment

| Diagnostic test | M (cut-offs for autism / autism spectrum) |
|-----------------|------------------------------------------|
| **Participants as informant** | | |
| Interview | ADOS<sup>a</sup> (<i>n</i> = 16) | | |
| Social Interaction & Communication | 10.31 (12 / 7) | 2.18 |
| Social Interaction | 6.58 (7 / 4) | 1.83 |
| Communication | 3.76 (3 / 2) | 1.13 |
| **Parents as informant** | | |
| Questionnaire | SCQ<sup>b</sup> (<i>n</i> = 9) | 20.89 (15) | 6.01 |
| Interview | ADI-R<sup>c</sup> (<i>n</i> = 11) | | |
| Social Interaction & Communication | 30.18 | 11.37 |
| Social Interaction | 17.09 (17) | 7.30 |
| Communication | 12.18 (8) | 4.81 |

<sup>M</sup> = mean; <sup>SD</sup> = standard deviation.

<sup>a</sup>ADOS = Autism Diagnostic Observation Schedule (Lord et al., 2000; German version: Rühl, 2004).

<sup>b</sup>SCQ = Social Communication Questionnaire (Rutter, 2003; German version: Bölte & Poustka, 2006).

<sup>c</sup>ADI-R = Autism Diagnostic Interview- Revised (Lord, Rutter, & Le Couteur, 1994; German version: Bölte, Rühl, Schmötzer & Poustka, 2003).
Supplementary Table 2

Results in the speech-in-noise recognition experiment for the ROIs provided in an independent atlas on the human subcortical auditory system (Sitek et al., 2019) and used in the voice identity recognition and vocal sound experiment. *P*-values are FWE corrected at peak level and Bonferroni corrected for the four regions of interest (*p* < 0.0125 FWE corrected). For information purposes only, grey font *z*-scores and *p*-values indicate results at FWE corrected thresholds (*p* > 0.05 FWE corrected).

| ROIs   | Control group | ASD group |
|--------|---------------|-----------|
|        | *x* | *y* | *z* | *Z*  | *p* | *x* | *y* | *z* | *Z*  | *p* |
| right IC | 6   | -31 | -10 | 3.10 | 0.011 | -   |   |   | 1.26 | 0.381 |
| left IC  | -   |   |   | 1.74 | 0.177 | -   |   |   | 2.28 | 0.072 |
| right MGB | -   |   |   | 0.33 | 0.627 | -   |   |   | -0.20 | 0.694 |
| left MGB | -   |   |   | 0.40 | 0.674 | -   |   |   | -0.19 | 0.728 |

| ROIs   | Controls > ASD | ASD > Controls |
|--------|----------------|----------------|
|        | *x* | *y* | *z* | *Z*  | *p* | *x* | *y* | *z* | *Z*  | *p* |
| right IC | -   |   |   | 1.75 | 0.199 | -   |   |   | 0.35 | 0.646 |
| left IC  | -   |   |   | 1.32 | 0.332 | -   |   |   | 0.69 | 0.552 |
| right MGB | -   |   |   | 1.34 | 0.317 | -   |   |   | 0.32 | 0.641 |
| left MGB | -   |   |   | 1.89 | 0.166 | -   |   |   | 0.69 | 0.611 |

*IC* = inferior colliculus; *MGB* = medial geniculate body; *x*, *y*, *z* = peak coordinates in MNI space (in mm).
Supplementary Table 3
MNI-coordinates for BOLD-responses for group comparisons in the speech-in-noise recognition experiment ($p < 0.0125$ FWE corrected at peak level and Bonferroni corrected for four regions of interest) with average tSNR included as an additional covariate of no interest. As in the main analysis there were no significant differences between groups. For information purposes only, grey font $z$-scores and $p$-values indicate results at $p > 0.0125$ FWE corrected.

|                     | Controls > ASD | ASD > Controls |
|---------------------|----------------|----------------|
|                     | $Z$     | $p$     | $Z$     | $p$     |
| right IC            | -       | 1.36    | 0.252   | -       | 2.00    | 0.087   |
| left IC             | -       | 1.38    | 0.225   | -       | 1.10    | 0.319   |
| right MGB           | -       | 1.92    | 0.087   | -       | 0.19    | 0.607   |
| left MGB            | -       | 1.22    | 0.315   | -       | 0.42    | 0.668   |

IC = inferior colliculus; MGB = medial geniculate body; $x$, $y$, $z$ = peak coordinates in MNI space (in mm).
Supplementary Table 4
Summary of average recognition accuracy (% correct) for the voice identity and the speech-in-noise recognition experiment. Scores are summarised as average over group with standard deviation (SD) and \( p \)-values from independent \( t \)-tests.

| Experiment                                | ASD                | Controls           | \( P \)  |
|-------------------------------------------|--------------------|--------------------|--------|
|                                           | \( M \) | \( SD \) | \( M \) | \( SD \) |        |
| **Voice identity recognition experiment** | (\( n_{ASD} = 16; n_{Controls} = 16 \)) |                     |        |
| Voice identity task                       | 76.36  | 11.61 | 87.36  | 7.15   | 0.003* |
| Speech task                               | 89.41  | 9.28  | 91.75  | 9.06   | 0.475  |
| **Vocal sound experiment** (\( n_{ASD} = 16; n_{Controls} = 16 \)) |                     |        |        |
| Total                                     | 11.88  | 6.25  | 18.13  | 7.20   | 0.014* |
| Vocal sounds                              | 5.38   | 2.16  | 8.25   | 4.12   | 0.019* |
| Speech                                    | 2.13   | 1.71  | 2.88   | 1.82   | 0.239  |
| Non-speech                                | 3.25   | 1.73  | 5.38   | 3.10   | 0.023* |
| Non-vocal sounds                          | 6.50   | 4.93  | 9.88   | 4.27   | 0.047* |
| Nature                                    | 1.19   | 1.64  | 1.44   | 1.75   | 0.680  |
| Animals                                   | 2.31   | 2.70  | 3.31   | 2.21   | 0.261  |
| Modern environment                        | 2.31   | 1.30  | 4.31   | 2.12   | 0.003* |
| Musical instruments                       | 0.69   | 0.95  | 0.81   | 0.98   | 0.716  |
| **Speech-in-noise recognition experiment** | (\( n_{ASD} = 17; n_{Controls} = 17 \)) |                     |        |
| No noise condition                        | 91.43  | 7.86  | 94.44  | 3.98   | 0.168  |
| Noise condition                           | 68.08  | 8.74  | 69.93  | 7.86   | 0.521  |

\( M \) = mean; \( SD \) = standard deviation. * Significant group differences (\( p < 0.05 \)).
Supplementary Table 5

Results for the right IC in the voice identity recognition experiment and the vocal sound experiment for the ROIs derived from previously published coordinates: x = 6, y = -33, z = -9 (MNI space) in Gaebler et al. (2020) and x = 6, y = -36, z = -10 (MNI space) in Griffiths et al. (2001). *P*-values are FWE corrected at peak level and Bonferroni corrected for four regions of interest (*p* < 0.0125 FWE corrected). For information purposes only, grey font *z*-scores and *p*-values indicate results at FWE corrected thresholds (*p* > 0.05 FWE corrected).

ROI Gaebler et al. (2020) | ROI Griffiths et al. (2001)
---|---
**Voice identity task > Speech task**

|        |        |       |   |        |        |       |   |
|--------|--------|-------|---|--------|--------|-------|---|
| Controls | 6      | -31   | -10| 3.61   | 0.001  | 6     | -34| -10| 3.24   | 0.004  |
| ASD     | -      |       |   | 0.86   | 0.556  | -     | 0.82| 0.565 |
| Controls > ASD | 6 | -31   | -10| 3.29   | 0.004  | 6     | -34| -10| 2.79   | 0.018  |

**Vocal sounds > Non-vocal sounds**

|        |        |       |   |        |        |       |   |
|--------|--------|-------|---|--------|--------|-------|---|
| Controls | 6      | -31   | -10| 2.91   | 0.013  | -     | 1.58| 0.358 |
| ASD     | -      |       |   | 0.40   | 0.652  | -     | 0.45| 0.647 |
| Controls > ASD | 6 | -31   | -10| 3.45   | 0.002  | -     | 2.30| 0.076 |

IC = inferior colliculus; x, y, z = peak coordinates in MNI space (in mm).
Supplementary Table 6
Group comparisons of the average movement (in mm) for all directions (x, y, and z) in the speech-in-noise recognition experiment.

| Average movement (mm) | ASD     | Controls | p     |
|-----------------------|---------|----------|-------|
|                       | M       | SD       | M     | SD    | p     |
| Translation           |         |          |       |       |       |
| x                     | 0.73    | .041     | 0.80  | 0.65  | 0.719 |
| y                     | 0.80    | 0.68     | 0.68  | 0.50  | 0.590 |
| z                     | 1.72    | 0.99     | 1.38  | 0.47  | 0.209 |
| Rotation              |         |          |       |       |       |
| x                     | 0.02    | 0.01     | 0.02  | 0.01  | 0.330 |
| y                     | 0.01    | 0.01     | 0.01  | 0.01  | 0.956 |
| z                     | 0.01    | 0.01     | 0.01  | 0.01  | 0.795 |

Directions: x = medial-lateral axis; y = anterior-posterior axis; z = dorsal-ventral axis. Values represent the maximum movement along one axis. M = mean; SD = standard deviation; p = p-value.
Supplementary Table 7
MNI-coordinates for BOLD-responses in the voice identity recognition, the vocal sound and the speech-in-noise recognition experiment at a statistical threshold of $p = 0.001$ uncorrected. Italic coordinates indicate results at a very lenient threshold of $p < 0.05$ uncorrected.

|                 | Voice identity task + Speech task |               | Vocal sounds + Non-vocal sounds |               | Speech task noise + Speech task no noise |               |
|----------------|----------------------------------|---------------|---------------------------------|---------------|----------------------------------------|---------------|
|                 | Control group                    | ASD group     | Control group                    | ASD group     | Control group                          | ASD group     |
|                 | $x$     | $y$     | $z$     | $Z$     | $x$     | $y$     | $z$     | $Z$     | $x$     | $y$     | $z$     | $Z$     | $x$     | $y$     | $z$     | $Z$     | $x$     | $y$     | $z$     | $Z$     |
| right IC        | 6      | -34     | -13     | 1.84    | 6      | -34     | -10     | 2.36    | 6      | -37     | -10     | 4.26    | 6      | -34     | -10     | 3.05    | 6      | -34     | -10     | 3.05    |
| left IC         | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       |
| right MGB       | 15     | -25     | -4      | 3.73    | -       | -       | -       | -       | 18     | -25     | -4      | 4.75    | -       | -       | -       | -       | 12     | -25     | -7      | 2.86    |
| left MGB        | -15    | -25     | -4      | 2.90    | -15    | -25     | -4      | 2.28    | -15    | -25     | -7      | 3.99    | -       | -       | -       | -       | -9     | -25     | -7      | 3.76    |
| controls > ASD  | -       | -       | -       | -       | ASD > Controls | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       |
| controls > ASD  | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       |
| right MGB       | 15     | -25     | -7      | 5.69    | 18     | -25     | -4      | 4.75    | 12     | -25     | -7      | 3.46    | 12     | -25     | -7      | 2.86    | 12     | -25     | -7      | 2.86    |
| left MGB        | -15    | -25     | -7      | 3.99    | -15    | -25     | -4      | 4.65    | -15    | -25     | -7      | 3.99    | -       | -       | -       | -       | -9     | -25     | -7      | 3.76    |
| controls > ASD  | -       | -       | -       | -       | ASD > Controls | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       |
| controls > ASD  | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       | -       |
| left IC         | -6     | -31     | -10     | 2.49    | -6     | -31     | -10     | 4.54    | -6     | -31     | -10     | 3.97    | -6     | -31     | -10     | 3.97    | -6     | -31     | -10     | 3.97    |
| left MGB        | -9     | -25     | -7      | 4.08    | -9     | -25     | -7      | 3.46    | -9     | -25     | -7      | 3.46    | -         | -       | -       | -       | -       | -9     | -25     | -7      | 3.46    |

IC = inferior colliculus; MGB = medial geniculate body; $x$, $y$, $z$ = peak coordinates in MNI space (in mm).
Supplementary Table 8
Results for correlation analyses (Pearson correlation, two-tailed) within the control and the ASD group for correlations between task performances in the voice identity and speech-in-noise recognition experiment and BOLD-responses in the left and right inferior colliculus (IC) (averaged mean 1st eigenvariates extracted for the whole left and right ROI (Sitek et al., 2019); for the voice identity task (voice identity recognition experiment), we extracted the 1st eigenvariate for the coordinate for which we found a significant group differences, i.e. MNI coordinate x=6, y=-10, z=-31). The results showed no significant correlations.

|                                | Controls | ASD  |
|--------------------------------|----------|------|
| **Voice identity recognition experiment** | n=16     | n=16 |
| **Voice identity task**         |          |      |
| left IC                        | p = .293; r = .280 | p = .275; r = -.290 |
| right IC                       | p = .889; r = .038  | p = .142; r = -.381 |
| **Speech task**                |          |      |
| left IC                        | p = .819; r = .062  | p = -.402; r = -.225 |
| right IC                       | p = .923; r = .023  | p = .243; r = -.310 |
| **Speech-in-noise recognition experiment** | n=17     | n=17 |
| **Speech task**                |          |      |
| left IC                        | p = .417; r = .210  | p = .204; r = -.324 |
| right IC                       | p = .339; r = .247  | p = .154; r = -.361 |
| **Speech task no noise**       |          |      |
| left IC                        | p = .135; r = .378  | p = .337; r = -.248 |
| right IC                       | p = .303; r = .265  | p = .391; r = -.222 |
**Supplementary References**

Bölte, S., Poustka, F. (2006). *Fragebogen zur Sozialen Kommunikation (FSK)*. Bern: Verlag Hans Huber.

Bölte, S., Rühl, D., Schmötzer, G., Poustka, F. (2003). *Diagnostisches Interview für Autismus – Revidiert (ADI-R)*. Bern: Verlag Hans Huber.

Gaebler, A. J., Zweerings, J., Koten, J. W., Konig, A. A., Turetsky, B. I., Zvyagintsev, M., & Mathiak, K. (2020). Impaired subcortical detection of auditory changes in schizophrenia but not in major depression. *Schizophrenia Bulletin, 46*(1), 193-201. doi:10.1093/schbul/sbz027

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics, 5*, 13. doi:10.3389/fninf.2011.00013

Griffiths, T. D., Uppenkamp, S., Johnsrude, I., Josephs, O., & Patterson, R. D. (2001). Encoding of the temporal regularity of sound in the human brainstem. *Nature Neuroscience, 4*(6), 633-637. doi:10.1038/88459

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Jr., Leventhal, B. L., DiLavore, P. C., . . . Rutter, M. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders, 30*(3), 205-223. doi:10.1023/A:1005592401947

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24*(5), 659-685. doi:10.1007/BF02172145
Murphy, K., Bodurka, J., & Bandettini, P. A. (2007). How long to scan? The relationship between fMRI temporal signal to noise ratio and necessary scan duration. *Neuroimage, 34*(2), 565-574. doi:10.1016/j.neuroimage.2006.09.032

Parsons, C. E., Young, K. S., Joensson, M., Brattico, E., Hyam, J. A., Stein, A., . . . Kringelbach, M. L. (2014). Ready for action: a role for the human midbrain in responding to infant vocalizations. *Social Cognitive and Affective Neuroscience, 9*(7), 977-984. doi:10.1093/scan/nst076

Rühl, D., Bölte, S., Feineis-Matthews, S., Poustka, F. (2004). *Diagnostische Beobachtungsskala für Autistische Störungen (ADOS)*. Bern: Verlag Hans Huber.

Rutter, M., Bailey, A., Lord, C. . (2003). *Social Communication Questionnaire (SCQ)*. Los Angeles, CA: Western Psychological Services.

Schelinski, S., Borowiak, K., & von Kriegstein, K. (2016). Temporal voice areas exist in autism spectrum disorder but are dysfunctional for voice identity recognition. *Social Cognitive and Affective Neuroscience, 11*(11), 1812-1822. doi:10.1093/scan/nsw089

Schelinski, S., Riedel, P., & von Kriegstein, K. (2014). Visual abilities are important for auditory-only speech recognition: evidence from autism spectrum disorder. *Neuropsychologia, 65*, 1-11. doi:10.1016/j.neuropsychologia.2014.09.031

Schelinski, S., Roswandowitz, C., & von Kriegstein, K. (2017). Voice identity processing in autism spectrum disorder. *Autism Research, 10*(1), 155-168. doi:10.1002/aur.1639

Sitek, K. R., Gulban, O. F., Calabrese, E., Johnson, G. A., Lage-Castellanos, A., Moerel, M., . . . De Martino, F. (2019). Mapping the human subcortical auditory system using histology, postmortem MRI and in vivo MRI at 7T. *Elife, 8*. doi:10.7554/eLife.48932

von Kriegstein, K., Dogan, O., Gruter, M., Giraud, A. L., Kell, C. A., Gruter, T., . . . Kiebel, S. J. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences of the United States of America, 105*(18), 6747-6752. doi:10.1073/pnas.0710826105
Welvaert, M., & Rosseel, Y. (2013). On the definition of signal-to-noise ratio and contrast-to-noise ratio for FMRI data. *Plos One, 8*(11), e77089. doi:10.1371/journal.pone.0077089