Modulation of the primary auditory thalamus when recognising speech in noise

Paul Glad Mihai[1,2], Nadja Tschentscher[3], Katharina von Kriegstein[1]

[1]Chair of Cognitive and Clinical Neuroscience, Faculty of Psychology, Technische Universität Dresden, Germany

[2]Max Planck Institute for Cognitive and Brain Sciences, Leipzig, Germany

[3]Research Unit Biological Psychology, Department of Psychology, Ludwig-Maximilians-University Munich, Germany

Abstract

Recognising speech in background noise is a strenuous daily activity, yet most humans can master it. A mechanistic explanation of how the human brain deals with such sensory uncertainty is the Bayesian Brain Hypothesis. In this view, the brain uses a dynamic generative model to simulate the most likely trajectory of the speech signal. Such simulation account can explain why there is a task-dependent modulation of sensory pathway structures (i.e., the sensory thalami) for recognition tasks that require tracking of fast-varying stimulus properties (i.e., speech) in contrast to relatively constant stimulus properties (e.g., speaker identity) despite the same stimulus input. Here we test the specific hypothesis that this task-dependent modulation for speech recognition increases in parallel with the sensory uncertainty in the speech signal. In accordance with this hypothesis, we show—by using ultra-high-resolution functional magnetic resonance imaging in human participants—that the task-dependent modulation of the left primary sensory thalamus (ventral medial geniculate body, vMGB) for speech is particularly strong when recognizing speech in noisy listening conditions in contrast to situations where the speech signal is clear. Exploratory analyses showed that this finding was specific to the left vMGB; it was not present in the midbrain structure of the auditory pathway (left inferior colliculus, IC). The results imply that speech in noise recognition is supported by modifications at the level of the subcortical sensory pathway providing driving input to the auditory cortex.

Author contributions: PGM: collected data, analysed data, interpreted results, wrote the manuscript, edited the manuscript. NT: conceptualised experiment, programmed experiment, edited the manuscript. KvK: conceptualised experiment, interpreted results, wrote the manuscript, edited the manuscript.

1. Introduction

Honking horns and roaring engines, the hammering from a construction site, the mix of music and speech at a restaurant or pub, the chit-chat of many children in a classroom are just some examples of background noises which continuously accompany us. Nevertheless, humans have a remarkable ability to hear and understand the conversation partner, even under these severe listening conditions (Cherry, 1953).

Understanding speech in noise is a complex task that involves both sensory and cognitive processes (Adank, 2012; Alavash et al., 2019; Best et al., 2007; Bregman, 1994; Brokx and Noteboom, 1982; Bronkhorst, 2015; Darwin and Hukin, 2000; Moore et al., 1985; Parikh and Loizou, 2005; Peelle, 2018; Sayles and Winter, 2008; Shinn-Cunningham and Best, 2008; Song et al., 2010). Difficulties in understanding speech in noise can occur in age-related hearing impairment (Schoof and Rosen, 2016), as well as in developmental disorders like autism spectrum disorder (Alcántara et al., 2004), auditory processing disorder (Iliadou et al., 2017), or developmental dyslexia (Chandrasekaran et al., 2009; Ziegler et al., 2009). In contrast, early musical training is associated with better abilities in extracting speech from a noisy background (Parbery-Clark et al., 2009; Strait et al., 2012). To-date it is by-and-large unclear why the human brain is so robust to speech-in-noise perception. Understanding human speech-in-noise recognition on a mechanistic level would be important as it would advance the understanding of why some clinical populations have difficulties with speech-in-noise perception. Furthermore, a more mechanistic understanding of how the human brain recognises speech-in-noise might also trigger new insight on why artificial speech recognition systems still have difficulties when speech is presented in noise (Gupta et al., 2016; Qian et al., 2016; Scharenborg, 2007).

One mechanistic account of brain function that attempts to explain how the human brain deals with noise or uncertainty in the stimulus input is the Bayesian brain hypothesis. It assumes that the brain represents information probabilistically and uses an internal generative model and predictive coding for the most effective processing of sensory input (Friston, 2005; Friston and Kiebel, 2009; Kiebel et al., 2008; Knill and Pouget, 2004). Such type of processing has the potential to explain why the human brain is robust to sensory uncertainty, e.g., when recognising speech despite noise in the speech signal (Knill and Pouget, 2004; Srinivasan et al., 1982). Although predictive coding is often discussed in the context of cerebral cortex organization (Hesselmann et al., 2010; Shipp et al., 2013), it may also be a governing principle of the interactions between cerebral cortex and subcortical sensory pathway structures (Adams et al., 2013; Bastos et al., 2012; Huang and Rao, 2011; Mumford, 1992; Seth Anil K. and Friston Karl J., 2016; von Kriegstein et al., 2008). In accordance with this suggestion, studies in animals found that feedback from cerebral cortex areas changes the processing in the sensory pathway, i.e., the sensory thalamus and brainstem nuclei (Krupa et al., 1999; Sillito et al., 2006, 1994; Wang et al., 2018).

In humans, responses in the auditory sensory thalamus (medial geniculate body, MGB) are higher for speech tasks (that emphasise recognition of fast-varying speech properties) in contrast to control tasks (that require recognition of relatively constant properties of the speech signal, such as the speaker identity or the sound intensity level). This response difference holds even if the stimulus input is the same (Díaz et al., 2012; von Kriegstein et al., 2008). This task-dependent modulation seems to be behaviorally relevant for speech recognition: performance level in auditory speech recognition was positively correlated with the amount of task-dependent modulation in the MGB of the left hemisphere (Mihai et al., 2019; von Kriegstein et al., 2008). This behaviourally relevant task-dependent modulation was located in the ventral part of the MGB (vMGB), which is the primary subsection of the MGB, but not in other MGB subsections (Mihai et al., 2019). These findings could fit the Bayesian brain hypothesis on cortico-subcortical interactions: cerebral cortex areas provide dynamic predictions about the incoming sensory input to the sensory thalamus to optimally encode the trajectory of the fast-varying and predictable

speech input (Díaz et al., 2012; von Kriegstein et al., 2008). If this is the case, the specific hypothesis ensues that the task-dependent modulation of the vMGB is especially involved when the fast dynamics of speech have to be recognised in conditions with high sensory uncertainty (Díaz et al., 2012; Feldman and Friston, 2010; Van de Cruys et al., 2014; Yu and Dayan, 2005), for example when the incoming signal is disturbed (Feldman and Friston, 2010; Friston and Kiebel, 2009; Gordon et al., 2017; Yu and Dayan, 2005). The present study aimed to test this hypothesis.

An ecologically valid way to increase uncertainty about the speech input is the presentation of speech in background noise (Chandrasekaran and Kraus, 2010a). We, therefore, tested, whether the task-dependent modulation of the left vMGB for speech is higher when the speech stimuli are heard in a noisy as opposed to a clear background. We used ultra-high field fMRI at 7 T and a design that has been shown to elicit task-dependent modulation of the MGB in previous studies (Díaz et al., 2012; von Kriegstein et al., 2008). We complemented the design by a noise factor: the speech stimuli were presented with and without background noise. The experiment was a 2 × 2 factorial design with the factors task (speech task, speaker task) and noise (noise, clear). To test our hypothesis, we performed a task × noise interaction analysis. We predicted that the task-dependent modulation of the left vMGB increases with decreasing signal-to-noise ratios (i.e., increasing uncertainty about the speech sounds). We focused on the left vMGB for two reasons. First, its response showed behavioural relevance for speech recognition in previous studies (Mihai et al., 2019; von Kriegstein et al., 2008). Second, a study on developmental dyslexia – a condition that is often associated with speech-in-noise recognition difficulties (Chandrasekaran et al., 2009; Ziegler et al., 2009) – showed reduced task-dependent modulation of the left MGB in comparison to controls (Díaz et al., 2012).

In addition to testing our main hypothesis, the design also (i) served to test for replicability of previous findings on the involvement of the MGB in speech recognition as well as its relevance for speech recognition behaviour (Mihai et al., 2019; von Kriegstein et al., 2008), and (ii) allowed to explore the role of the inferior colliculus (IC) – the midbrain station of the auditory sensory pathway – in speech and speech-in-noise recognition.

2. Results

Participants listened to blocks of auditory syllables (e.g., /ada/, spoken by three different speakers) and performed either a speech or a speaker task (Figure 1). In the speech task, participants reported via button press whether the current syllable was different from the previous one (1-back task). In the speaker task, participants reported via button press whether the current speaker was different from the previous one. The speakers' voices were resynthesized from the recordings of one speaker's voice to only differ in constant speaker individuating features (i.e., the vocal tract length and the fundamental frequency of the voice). This ensured that the speaker task could not be done on dynamic speaker individuating features (e.g., idiosyncrasies in pronunciations of phonemes). Participants listened to either stimuli embedded in speech-shaped noise (noise condition) or without background noise (clear condition).

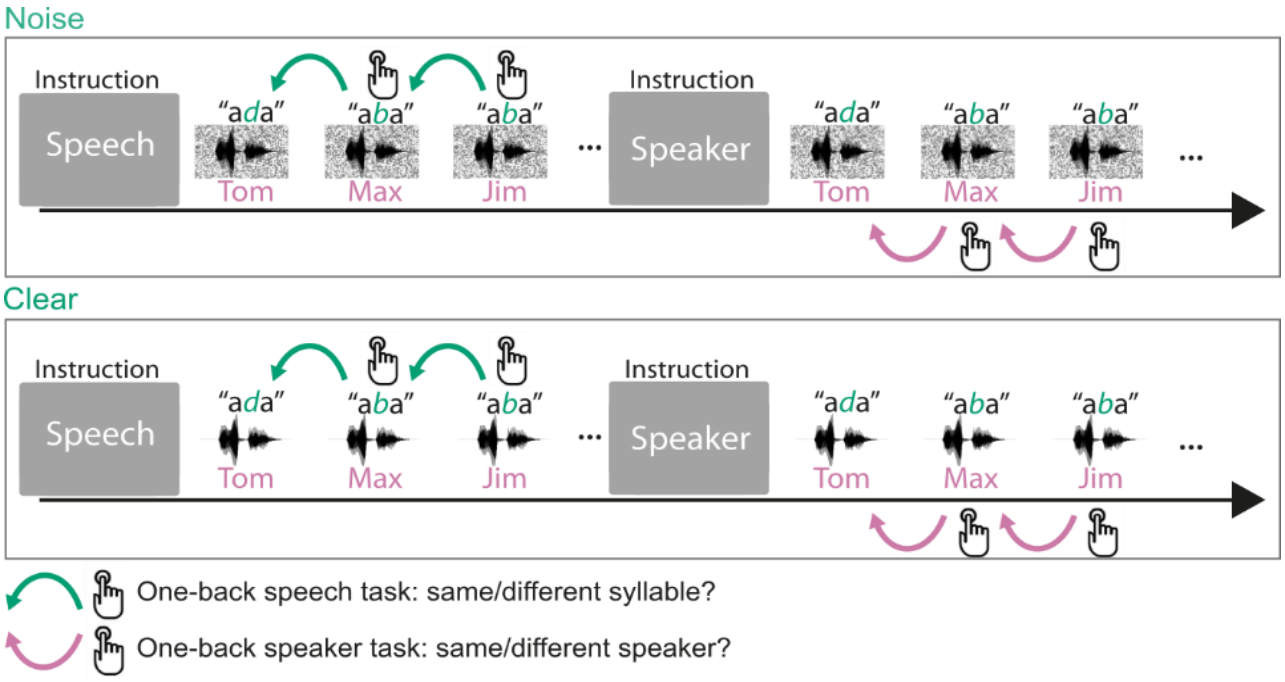

Figure 1. Design and trial structure of the experiment. In the speech task, listeners performed a one-back syllable task. They pressed a button whenever there was a change in syllable in contrast to the immediately preceding one, independent of speaker change. The speaker task used precisely the same stimulus material and trial structure. The task was to press a button when there was a change in speaker identity in contrast to the immediately preceding one,

independent of syllable change. An initial task instruction screen informed participants about which task to perform. Participants heard stimuli either with concomitant speech-shaped noise (noise condition) or without background noise (clear condition). Thus the experiment had four conditions: speech task/noise, speaker task/noise, speech task/clear, speaker task/clear. Stimuli in the speech and speaker tasks were precisely identical.

2.1 Behavioural results

Participants performed well above chance level in all four conditions (> 82% correct; Table 1; Figure 2A).

Table 1. The proportion of hits for each of the four conditions in the experiment. HDP: highest posterior density interval.

	Speech task/ Noise	**Speaker task/ Noise**	**Speech task/ Clear**	**Speaker task/ Clear**
% Mean [95% HPD]	0.82 [0.62, 0.95]	0.87 [0.74, 0.96]	0.92 [0.83, 0.98]	0.90 [0.81, 0.97]

Performing the tasks with background noise was more difficult than the conditions without background noise for both the speech and the speaker task (Figure 2B, for details on statistics, see figure legend). The rate of hits in the speech task was the same as in the speaker task (Figure 2C). There was a detectable interaction between task and noise (Figure 2D/E), but simple main effects (i.e., speech task/noise - speaker task/noise (Figure 2F) and speech task/clear - speaker task/clear (Figure 2G)) were not present.

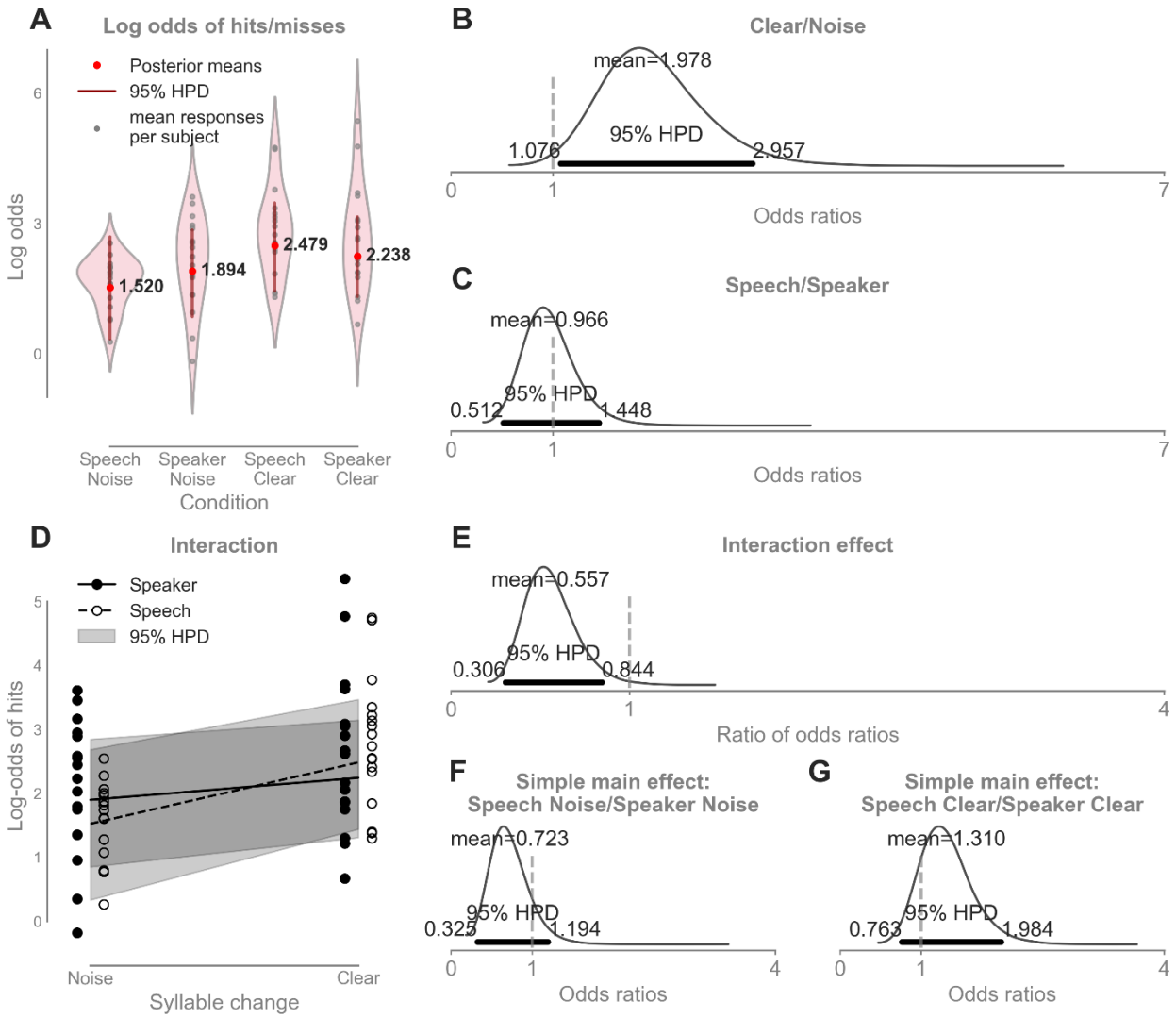

Figure 2. Behavioural results. We performed a binomial logistic regression to compute the rate of hits and misses in each condition because behavioural data were binomially distributed. For this reason, results are reported in log odds and odds ratios. The results showed a detectable main effect of noise and interaction between noise and task. There was no main effect of task, and no detectable simple main effects (speech task/noise - speaker task/noise; speech task/clear - speaker task/clear). ***A.*** *Log odds of hits and misses for each condition. The grey dots indicate mean responses for individual participants, the red dots and accompanying numbers denote the posterior mean per condition, and the dark red lines demarcate the 95% highest posterior density interval (HPD). The rate of hits compared to misses is plotted on a log scale to allow for a linear representation.* ***B.*** *Mean odds ratio for the clear and noise conditions. The odds of hits in the clear condition were on average twice as high as in the noise*

condition (the mean odds ratio was 1.978 [1.076, 2.957]). The HPD excluded 1 and indicated a detectable difference between conditions: No difference would be assumed if the odds ratio was 1 (50/50 chance or 1:1 ratio; Chen, 2003). ***C.*** *Mean odds ratio for the speech task - speaker task conditions. The mean odds ratio was ~1 indicating no difference between the speech and speaker task conditions.* ***D****. Visualization of the interaction (task × noise) as a comparison of slopes with 95% HPD.* ***E****. The ratio of odds ratios of the simple main effects speech task/noise - speaker task/noise and speech task/clear - speaker task/clear. The mean and 95% HPD was 0.557 [0.306, 0.844]. The HPD excluded 1 indicating an interaction effect.* ***F****. Mean odds ratio for the simple main effect speech task/noise - speaker task/noise. The rate of hits in the speech task/noise condition was on average ~1/3 lower than the rate of hits in the speaker task/noise condition; however, the HPD strongly overlapped 1 indicating that there was no difference between conditions.* ***G****. Mean odds ratio for the simple main effect speech task/clear - speaker task/clear. The rate of hits in the speech task/clear condition was on average ~1/3 higher than the rate of hits in the speaker task/clear condition; however, the HPD strongly overlapped 1 indicating that there was no detectable difference between conditions.*

2.2 fMRI Results

2.2.1 The task-dependent modulation of left vMGB increased for recognizing speech-in-noise in contrast to the clear speech condition

We localised the left vMGB based on an independent functional localizer (see section 4. Materials and Methods). Following our hypothesis, there was increased BOLD response for the task × noise interaction [(speech task/noise - speaker task/noise) - (speech task/clear - speaker task/clear)] in the left vMGB (Figure 3A/B). The interaction effect had a mean large effect size ranging from a small effect to a very large effect (g*=2.549 [0.211, 5.066]; Figure 3C and 3D). The 95% HPD of the interaction effect excluded 0, indicating that this was a robust effect (Bunce and McElreath, 2017; McElreath, 2018). Simple main effect analyses showed that the direction of the interaction was as expected. The speech task/noise condition yielded higher left vMGB responses in contrast to the speaker task/noise condition, ranging from a medium to a very large effect (g* = 1.104 [0.407,

1.798]; Figure 3E). Conversely, the left vMGB response difference between the speech task and speaker task in the clear condition had a small effect size (g* = 0.243 [-0.366, 0.854]; Figure 3F), ranging from a negative medium effect to a positive large effect, and the HPD overlapped 0.

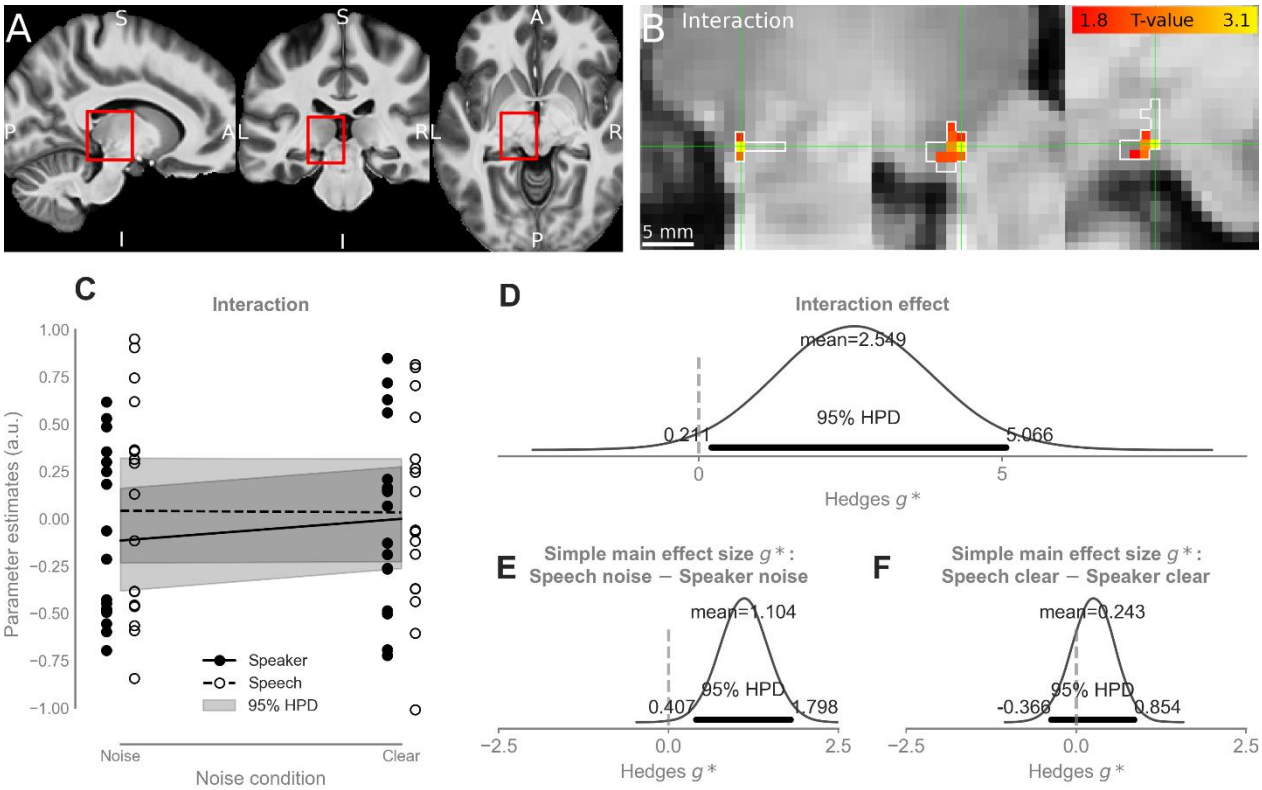

Figure 3. fMRI results. ***A.*** *The mean T1 structural image across participants in MNI space. Red rectangles denote the approximate location of the left MGB and encompass the zoomed-in views in B. Letters indicate anatomical terms of location: A, anterior; P, posterior; S, superior; I, inferior; L, left; R, right. Panels A and B share the same orientation across columns; i.e., from left to right: sagittal, coronal, and axial.* ***B.*** *Statistical parametric map of the interaction (yellow-red colour code): (speech task/noise - speaker task/noise) - (speech task/clear - speaker task/clear) overlaid on the mean structural T1 image. Crosshairs point to MNI coordinate (-11, -28, -6). The white outline shows the boundary of the vMGB mask.* ***C.*** *Parameter estimates (mean-centred) within the vMGB mask. Open circles denote parameter estimates of the speech task condition; filled circles denote parameter estimates of the speaker task condition. Dashed black line: the relationship between noise condition (noise, clear) and parameter estimates in the speech task. Solid black line: the relationship between noise condition (noise, clear) and parameter estimates in the speaker task. The shaded grey area shows the 95% HPD.* ***D-F*** *Bayesian Analysis of the parameter estimates.* ***D.*** *The effect size of the interaction: the effect size for the interaction effect was very large (2.549 [0.211, 5.066]) and the HPD excluded zero (indicated by the dashed vertical line).* ***E.*** *Simple main effect:*

speech task/noise – speaker task/noise. The mean effect size was large (1.104 [0.407, 1.798]). The HPD excluded zero. ***F****. Simple main effect: speech task/clear – speaker task/clear. The mean effect size was small (0.243 [-0.366, 0.854]). The HPD contained zero.*

The results showed that the task-dependent modulation of the left vMGB for the speech task was increased when participants recognised speech - speaker identity in background noise in contrast to speech - speaker identity without background noise. This finding cannot be explained by differences in stimulus input as the same stimulus material was used for the speech and the speaker task. The results are also unlikely due to differences in task difficulty between conditions, as the behavioural results showed no detectable differences in performance for the simple main effects.

2.2.2 Test for replication of previous findings

In addition to addressing the main hypothesis of the present paper, the data also allowed the testing for replication of previous findings (Díaz et al., 2012; Mihai et al., 2019; von Kriegstein et al., 2008), i.e., a test for a main effect of task (speech - speaker) in left and right MGB and a test for a correlation between speech recognition performance and main effect of task across participants in the left MGB.

Main effect of task: Consistent with previous reports (Díaz et al., 2012; von Kriegstein et al., 2008) there was a large positive main effect for the speech - speaker task in the left vMGB ranging from a small to a very large effect (g* = 0.679 [0.192, 1.200]; Figure 4 A & B). In the right vMGB, the main effect of task was small and the HPD overlapped 0 (g*=0.295 [-0.290, 0.882]).

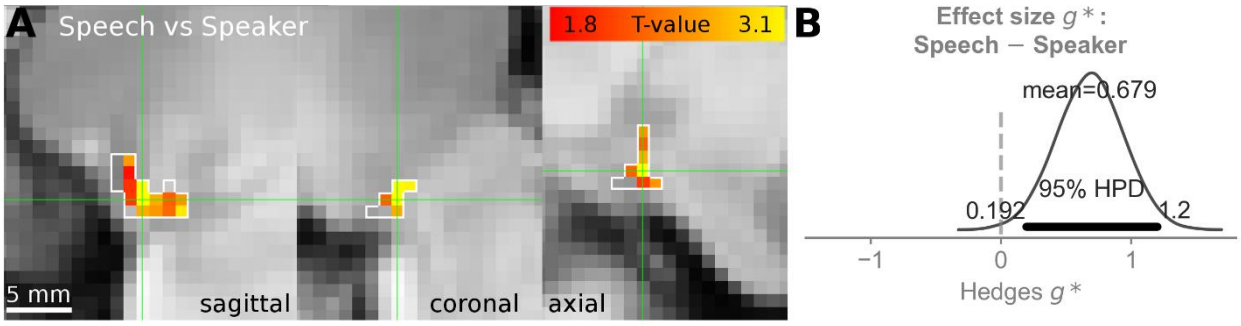

Figure 4. Main effect of task in the left vMGB. ***A.*** *Statistical parametric map of the main effect of task (yellow-red colour code) overlaid on a mean T1 image: speech task – speaker task. Crosshairs point to MNI coordinate (-12, -27, -7). The white outline shows the boundary of the vMGB mask. The orientation of the images is the same as in Figure 3A/B.* ***B.*** *Results of the Bayesian analysis of the parameter estimates for the main effect of task. There was a large effect size for the contrast speech – speaker task of 0.679 [0.192, 1.200]. The HPD excluded zero.*

Correlation between main effect of task and speech recognition performance: There was no significant correlation between the task-dependent modulation (i.e., parameter estimates for the contrast of speech - speaker) and the correct proportion of hits in the speech task; the effect size was very small and non-significant (mean Pearson's r = 0.15, p = 0.566; Figure S1A). A positive correlation between task-dependent modulation of the left MGB and speech task performance across participants has been reported in three previous experiments (experiments 1 and 2 of von Kriegstein et al. (2008) with n = 16 and n= 17 participants, (Mihai et al., 2019) with n = 33 participants), but was also not significant in one previous study (Díaz et al., 2012, with n = 14 participants). Since the previous studies did not include the factor noise, we also computed correlation coefficients between the simple main effect of task (speech/clear - speaker/clear task) and the proportion of hits in the speech/clear condition. Correlation coefficients were small and non-significant (r=0.03, p=0.917; Figure S1B).

To not wrongly treat variable results across studies as indicating a null-effect (Amrhein et al., 2019), we performed a random-effects meta-analysis (Figure 4) to test whether there is a meta-analytic significant correlation (speech - speaker task correlated with speech accuracy score across participants) across the present and previous studies. We included five studies in the meta-analysis: two experiments from von Kriegstein et al., (2008), results

from the control participants of Díaz et al. (2012), the experiment described in Mihai et al., (2019), and the current study. The meta-analysis yielded an overall effect size (Fisher z) of z=0.45 [0.23, 0.66], p<0.001 that corresponds to *r*=0.42. The direction of the correlation for all experiments was positive. The current study had a minimal correlation value that was not significant but was positive, thus in the same direction as the other studies.

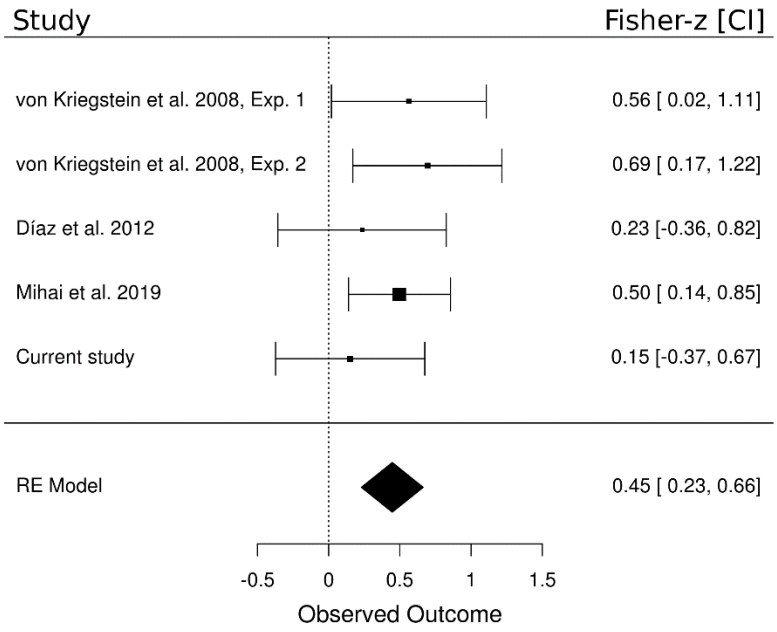

Figure 4. Meta-analysis of five experiments that investigated the correlation in the MGB between the contrast speech - speaker task and the proportion of hits in the speech task across participants. Experiment 1 of von Kriegstein et al. (2008) tested a speech - loudness task contrast correlated with performance in the speech task (n=16). All other experiments included a speech task - speaker task contrast correlated with performance in the speech task (i.e., experiment 2 of von Kriegstein et al. (2008) (n=17), the control participants of Díaz et al. (2012) (n=14), (Mihai et al., 2019) (n=33), and the current study (n=17)). The meta-analysis yielded an overall Fisher z = 0.45 [0.23, 0.66], p<0.001 which corresponds to an r=0.42. The area of the squares denoting the effect size is directly proportional to the weighting of the particular study when computing the meta-analytic overall score.

We attribute the non-significant correlation between the task-dependent modulation and the correct proportion of hits in the speech task in the present study to the fact that ~11% of the behavioural data in the speech task had ceiling or near to ceiling responses resulting in reduced correlation values (Bland and Altman, 2011). Many of the behavioural values were huddled towards the ceiling when plotted against BOLD responses (Figure S1). This

was not the case in previous studies, where there were no ceiling effects in the behavioural data *(Díaz et al., 2012; Mihai et al., 2019; von Kriegstein et al., 2008)*.

2.2.4 Exploratory analyses on the inferior colliculus

In exploratory analyses, we investigated the bilateral inferior colliculus' (IC) involvement during speech processing. The reason for these exploratory analyses were studies using auditory brainstem responses (ABR) during passive listening to speech sounds that have shown that the quality of speech sound representation (i.e., as measured by the frequency following response, FFR) explains inter-individual variability in speech-in-noise recognition abilities (Chandrasekaran et al., 2009; Schoof and Rosen, 2016; Selinger et al., 2016; Song et al., 2010). These findings indicated that there might be subcortical nuclei beyond the MGB that are involved in speech-in-noise perception, potentially also sources in the auditory brainstem, particularly the IC (Chandrasekaran and Kraus, 2010b). Four previous fMRI experiments, however, have shown that there is *no* significant task-dependent modulation (i.e., higher BOLD responses for a speech in contrast to a control task on the same stimuli) of the inferior colliculus (Díaz et al., 2012; Mihai et al., 2019; von Kriegstein et al., 2008). Two of them showed a significant positive correlation between the amount of BOLD response difference between a speech and a control task in the left IC and the speech recognition performance across participants (von Kriegstein et al., 2008), but the others did not. Thus the role of the IC in speech recognition and speech-in-noise recognition is to date unclear. In the present data, there was a small effect of task in the left IC (speech - speaker, left $g^*=0.309$ [-0.286, 0.902] and right $g^*= 0.126$ [-0.393, 0.646], however, the HPD overlapped zero. The task × noise interaction contained no explanatory power (left: $g^*=0.049$ [-0.103, 0.202], right: $g^*=-0.010$ [-0.136, 0.111]) and introduced overfitting. We, therefore, excluded it from the model, and the reported results were computed from the model without an interaction term.

The correlation between the task-dependent modulation (i.e., speech - speaker task contrast) and the speech recognition scores across participants in the left IC was not significant in the current study ($r=0.44$, $p=0.074$). We tested the left IC only since the correlations found in two previous experiments were restricted to the left IC (von

Kriegstein et al., 2008 experiment 1 and 2). We performed a random-effects meta-analysis to test whether there is, nevertheless, a consistent correlation effect in the left IC across studies. We included five studies in the meta-analysis: two experiments from von Kriegstein et al., (2008), the control participants of Díaz et al., (2012), the experiment described in Mihai et al., (2019), and the current study. The meta-analysis yielded an overall effect size (Fisher z) of z=0.36, p<0.001 that corresponds to *r*=0.35. The direction of the correlation for all experiments was positive.

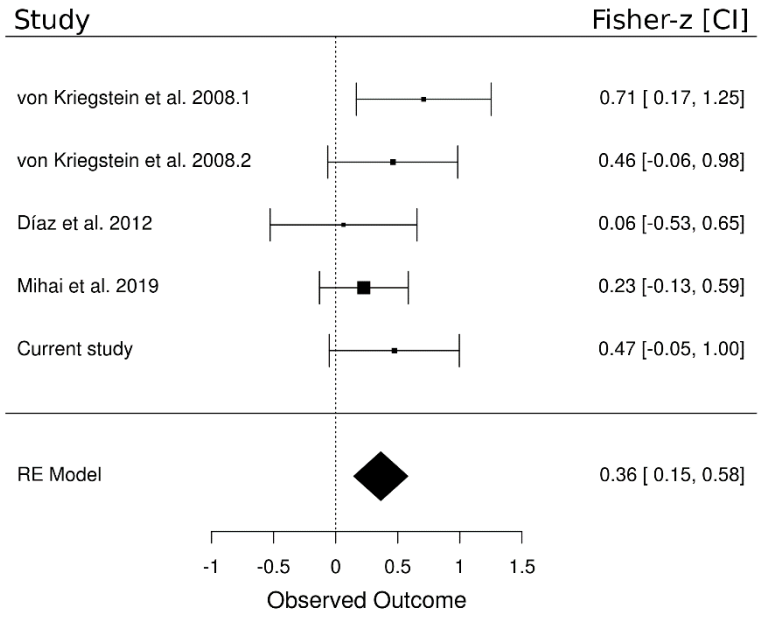

Figure 5. Meta-analysis of five experiments that investigated the correlation in the left IC between the contrast speech - speaker task and the proportion of hits in the speech task across participants. Experiment 1 of von Kriegstein et al. (2008) tested a speech - loudness task contrast correlated with performance in the speech task (n=16). All other experiments included a speech task - speaker task contrast correlated with performance in the speech task (i.e., experiment 2 of von Kriegstein et al. (2008) (n=17), (Díaz et al., 2012) (n=14), (Mihai et al., 2019) (n=33), and the current study (n=17)). The meta-analysis yielded an overall Fisher z = 0.36 [0.15, 0.58], p<0.001 which corresponds to an r=0.35. The area of the squares denoting the effect size is directly proportional to the weighting of the particular study when computing the meta-analytic overall score.

3. Discussion

We showed that the task-dependent modulation of the left hemispheric primary sensory thalamus (vMGB) for speech is particularly strong when recognising speech in noisy listening conditions in contrast to conditions where the speech signal is clear. This finding confirmed our a priori hypothesis which was based on explaining sensory thalamus function within a Bayesian brain framework. Exploratory analyses showed that there was no influence of noise on the responses for the contrast between speech and speaker task in the auditory midbrain, i.e., the inferior colliculi (IC). Besides answering our main hypothesis, we also provided three additional key findings. First, we replicated results from previous experiments (Díaz et al., 2012; von Kriegstein et al., 2008) that showed task-dependent modulation in the MGB for speech, and localised the task-dependent modulation in the vMGB (Mihai et al., 2019). Second, a meta-analysis of five studies showed that there was a positive correlation between the task-dependent modulation for speech in the left MGB and behavioural performance in the speech task across studies. Third, the same meta-analysis revealed a positive correlation between the task-dependent modulation for speech and the behavioural performance in the speech recognition task in the left IC.

Our main hypothesis in the present paper was based on the assumption that predictive coding might be a governing principle of how the human brain deals with background noise during speech recognition. Bayesian approaches to brain function propose that the brain uses internal dynamic models to predict the trajectory of the sensory input (Friston, 2005; Friston and Kiebel, 2009; Kiebel et al., 2008; Knill and Pouget, 2004). Thus, slower dynamics of the internal dynamic model (e.g., syllable and word representations) could be encoded by auditory cerebral cortex areas (Davis and Johnsrude, 2007; Giraud et al., 2000; Hickok and Poeppel, 2007; Mattys et al., 2012; Price, 2012; Wang et al., 2008), and provide predictions about the faster dynamics of the input arriving at lower levels of the anatomic hierarchy (Kiebel et al., 2008; von Kriegstein et al., 2008). In this view, dynamic predictions modulate the response properties of the first-order sensory thalamus to optimise the early stages of speech recognition (Mihai et al., 2019). In speech processing, such a mechanism might be especially useful as the signal includes rapid dynamics, as predictable (e.g., due to co-articulation or learned statistical regularities in words) (Saffran, 2003), and often has to

be computed online under conditions of (sensory) uncertainty. Uncertainty refers to the limiting reliability of sensory information about the world (Knill and Pouget, 2004). Examples include the density of hair cells in the cochlea that limit frequency resolution, the neural noise-induced at different processing stages, or – as was the case in the current study – background environmental noise that surrounds the stimulus of interest. An internal generative model about the fast sensory dynamics (Friston, 2005; Friston and Kiebel, 2009; Kiebel et al., 2008; Knill and Pouget, 2004) of speech could lead to enhanced stimulus representation in the subcortical sensory pathway and by that provides improved signal quality to the auditory cortex. Such a mechanism would result in more efficient processing when taxing conditions, such as background noise, confront the perceptual system. The interaction between task and noise in the left vMGB is in congruence with such a mechanism. It shows that the task-dependent modulation of the left vMGB is increased in a situation with high sensory uncertainty in contrast to the situation with lower sensory uncertainty.

Speech-in-noise recognition abilities are thought to rely (i) on additional cognitive resources that are recruited when recognising speech-in-noise (reviewed in Peelle, 2018) and (ii) on the fidelity of speech sound representation in brainstem nuclei, as measured by auditory brainstem response recordings (reviewed in Anderson and Kraus, 2010). For example, studies investigating speech-in-noise recognition at the level of the cerebral cortex found networks that include areas pertaining to linguistic, attentional, working memory, and motor planning (Bishop and Miller, 2008; Salvi et al., 2002; Scott et al., 2004; Wong et al., 2008). These results suggest that during speech recognition in challenging listening conditions additional cerebral cortex regions are recruited that likely complement the processing of sound in the core speech network (reviewed in Peelle, 2018). The present study showed that besides the additional cerebral cortex region recruitment, a specific part of the sensory pathway is also modulated during speech-in-noise recognition, the left vMGB.

Auditory brainstem response (ABR) recordings during passive listening to speech sounds have shown that the quality of speech sound representation (i.e., as measured by the frequency following response, FFR) explains inter-individual variability in speech-in-noise recognition abilities (Chandrasekaran et al., 2009; Schoof and Rosen, 2016; Selinger et al.,

2016; Song et al., 2010) and can be modulated by attention to speech in situations with two competing speech streams (Forte et al., 2017). It is difficult to directly relate the results of these FFR studies on participants with varying speech-in-noise recognition abilities (Chandrasekaran et al., 2009; Schoof and Rosen, 2016; Selinger et al., 2016; Song et al., 2010) to the studies on task-dependent modulation of structures in the subcortical sensory pathway (Díaz et al., 2012; Mihai et al., 2019; von Kriegstein et al., 2008): they involve very different measurement modalities and the FFR studies focus mostly on speech-in-noise perception in passive listening designs. One major candidate for the FFR source is the inferior colliculus. Particularly for speech, the FFR, as recorded by EEG, seems to be dominated by brainstem and auditory nerve sources (Bidelman, 2018; reviewed in Chandrasekaran et al., 2014). The results of the present study, however, do not provide evidence for a specific involvement of the inferior colliculus when recognising speech-in-noise. Whether the inferior colliculus plays a different role in speech-in-noise processing is an open question.

We speculate that the task-dependent vMGB modulation might be a result of feedback from cerebral cortex areas. The strength of the feedback could be enhanced when speech has to be recognised in background noise. The task-dependent feedback may emanate directly from primary auditory or association cortices, or indirectly via other structures such as the reticular nucleus with its inhibitory connections to the MGB (Rouiller and de Ribaupierre, 1985). Feedback cortico-thalamic projections from layer 6 in A1 to the vMGB, but also from association cortices such as the motion-sensitive planum temporale (Tschentscher et al., 2019), may modulate information ascending through the lemniscal pathway, rather than convey information to the vMGB (Lee, 2013; Llano and Sherman, 2008).

Difficulties in understanding speech-in-noise accompany developmental disorders like autism spectrum disorder, developmental dyslexia, and auditory processing disorders (Alcántara et al., 2004; Bellis and Bellis, 2015; Chandrasekaran et al., 2009; Schelinski and Kriegstein, 2019; Schoof and Rosen, 2016; Wong et al., 2009; Ziegler et al., 2009). In the case of developmental dyslexia, previous studies have found that developmental dyslexics do not have the same amount of task-dependent modulation of the left MGB for speech recognition as controls (Díaz et al., 2012) and also do not display the

same context-sensitivity of brainstem responses to speech sounds as typical readers (Chandrasekaran et al., 2009). In addition, diffusion-weighted imaging studies have found reduced structural connections between the MGB and cerebral cortex (i.e., the motion-sensitive planum temporale) of the left hemisphere in developmental dyslexics compared to controls (V5/MT; motion-sensitive planum temporale; Müller-Axt et al., 2017; Tschentscher et al., 2019). These deficient structures might account for the difficulties in understanding speech-in-noise in developmental dyslexia. Consider distinguishing speech sounds like "dad" and "had" in a busy marketplace. For typically developed individuals, vMGB responses might be modulated to optimally encode the subtle but predictable spectrotemporal cues that enable the explicit recognition of speech sounds. This modulation would enhance speech recognition. For developmental dyslexics, however, this vMGB modulation may be impaired and may explain their difficulty with speech perception in noise (Boets et al., 2007; Díaz et al., 2012; Ziegler et al., 2009).

In conclusion, the results presented here suggest that the left vMGB is particularly involved in decoding speech as opposed to identifying the speaker if there is background noise. This enhancement may be due to top-down processes that act upon subcortical sensory structures, such as the auditory thalamus, to better predict dynamic incoming signals in conditions with high sensory uncertainty.

4. Materials and Methods

4.1 Participants

The Ethics committee of the Medical Faculty, University of Leipzig, Germany, approved the study. We recruited 17 participants (mean age 27.7, SD 2.5 years, 10 female; 15 of these participated in a previous study: Mihai et al., 2019) from the database of the Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany. The participants were right-handed (as assessed by the Edinburgh Handedness Inventory (Oldfield 1971)), and native German speakers. Participants provided written informed consent. None of the participants reported a history of psychiatric or neurological disorders, hearing difficulties, or current use of psychoactive medications. Normal hearing abilities were confirmed with

pure tone audiometry (250 Hz to 8000 Hz; Madsen Micromate 304, GN Otometrics, Denmark) with a threshold equal to and below 25 dB). To exclude possible undiagnosed developmental dyslexics, we tested the participant's reading speed and reading comprehension using the German LGVT: 6-12 test (Schneider et al., 2007). The cut-off for both reading scores was set to those levels mentioned in the test instructions as the "lower average and above" performance range (i.e., 26% - 100% of the calculated population distribution). None of the participants performed below the cut off performance (mean 68.7%, SD 20.6%, lowest mean score: 36%). In addition, participants were tested on rapid automatized naming (RAN) of letters, numbers, and objects (Denckla and Rudel, 1976). The time required to name letters and numbers predicts reading ability and is longer in developmental dyslexics compared with typical readers, whereas the time to name objects is not a reliable predictor of reading ability in adults (Semrud-Clikeman et al., 2000). Participants scored well within the range of control participants for letters (mean 17.25, SD 2.52 s), numbers (mean 16.79, SD 2.63 s), and objects (mean 29.65, SD 4.47 s), based on results from a previous study (Díaz et al., 2012, letters: 16.09, SD 2.60; numbers: 16.49, SD 2.35; objects: 30.84, SD 5.85; age of participants was also comparable 23.5, SD 2.8 years). Furthermore, none of the participants exhibited a clinically relevant number of traits associated with autism spectrum disorder as assessed by the Autism Spectrum Quotient [AQ; mean: 15.9, SD 4.1; cut-off: 32-50; (Baron-Cohen et al., 2001)]. We tested AQ as autism can be associated with difficulties in speech-in-noise perception (Alcántara et al., 2004; Groen et al., 2009). Participants received monetary compensation for participating in the study.

4.2 Stimuli

We recorded 79 different vowel-consonant-vowel (VCV) syllables with an average duration of 784 ms, SD 67 ms. These were spoken by one male voice (age 29 years), recorded with a video camera (Canon Legria HFS10, Canon, Japan) and a Røde NTG-1 microphone (Røde Microphones, Silverwater, NSW, Australia) connected to a pre-amplifier (TubeMP Project Series, Applied Research and Technology, Rochester, NY, USA) in a sound-attenuated room.

The sampling rate was 48 kHz at 16 bit. Auditory stimuli were cut and flanked by Hamming windows of 15 ms at the beginning and end, converted to mono, and root-mean-square equalised using Python 3.6 (Python Software Foundation, www.python.org). The 79 auditory files were resynthesized with TANDEM-STRAIGHT (Banno et al., 2007) to create three different speakers: 79 auditory files with a vocal tract length (VTL) of 17 cm and glottal pulse rate (GPR) of 100 Hz, 79 with VTL of 16 cm and GPR of 150 Hz, and 79 with VTL of 14 cm and GPR of 300 Hz. This procedure resulted in 237 different auditory stimuli. The parameter choice (VTL and GPR) was motivated by the fact that a VTL difference of 25% and a GPR difference of 45% suffices for listeners to hear different speaker identities (Gaudrain et al., 2009; Kreitewolf et al., 2014). Additionally, we conducted pilot experiments (12 pilot participants which did not participate in the main experiment) in order to fine-tune the combination of VTL and GPR that resulted in a balanced behavioural accuracy score between the speech and speaker tasks. The pilot experiments were conducted outside the scanner, and each run included continuous recordings of scanner gradient noise to simulate a real scanning environment.

The 237 stimuli were embedded in background noise to create the stimuli for the condition with background noise. The background noise consisted of normally distributed random (white) noise filtered with a speech-shaped envelope. We calculated the envelope from the sum of all VCV stimuli presented in the experiment. We used speech-shaped noise as it has a stronger masking effect than stationary random non-speech noise (Carhart et al., 1975). Before each experimental run, the noise was computed and added to the stimuli included in the run with a signal-to-noise ratio (SNR) of 2 dB. The SNR choice was based on a pilot study that showed a performance decrease of at least 5% but no greater than 15% between the clear and noise condition. In the pilot study, we started at an SNR of -10 dB and increased this value until we converged on an SNR of 2 dB. Calculations were performed in Matlab 8.6 (The Mathworks Inc., Natick, MA, USA) on Ubuntu Linux 16.04 (Canonical Ltd., London, UK).

4.3 Procedure

We conceived the experiment as a 2 × 2 factorial design with the factors task (speech, speaker) and background noise (clear, noise). Participants listened to blocks of auditory VCV syllables and were asked to perform two types of tasks: a speech task and a speaker task. In the speech task, participants reported via button press whether the current syllable was different from the previous one (1-back task). In the speaker task, participants reported via button press whether the current speaker was different from the previous one. The blocks had either syllables with background noise (noise condition) or without background noise (clear condition).

Task instructions were presented for two seconds before each block and consisted of white written words on a black background (German words “Silbe” for syllable, and “Person” for person). After the instruction, the block of syllables started (Figure 1). Each block contained twelve stimuli. Each stimulus had a duration of approximately 784 ms, and the stimulus presentation was followed by 400 ms of silence. Within one block both syllables and speakers changed at least twice, with a theoretical maximum of nine changes. The theoretical maximum was derived from random sampling of seven instances from three possible change types: no change, speech change, speaker change, and change of speech and speaker. The average length of a block was 15.80 seconds, SD 0.52 seconds.

The experiment was divided into four runs. The first three runs had a duration of 12:56 min and included 40 blocks: 10 for each of the four conditions (speech task/noise, speaker task/noise, speech task/clear, speaker task/clear). A fourth run had a duration of 6:32 min and included 20 blocks (5 for each of the four conditions). For two participants, only the first three runs were recorded due to time constraints. Participants could rest for one minute between runs.

Participants were familiarised with the three speakers’ voices to ensure that they could perform the speaker-identity task of the main experiment. The speaker familiarisation took place 30 minutes before the fMRI experiment. It consisted of a presentation of the speakers and a test phase. In the presentation phase, the speakers were presented in six blocks, each containing nine pseudo-randomly chosen VCV stimuli from the 237 total. Each block

contained one speaker-identity only. Participants were alerted to the onset of a new speaker identity block by the presentation of white words on a black screen indicating speaker 1, speaker 2, or speaker 3. Participants listened to the voices with the instruction to memorise the speaker's voice. In the following test phase participants were presented with four blocks of nine trials that each contained randomly chosen syllable pairs spoken by the three speakers. The syllable pairs could be from the same or a different speaker. We asked participants to indicate whether the speakers of the two syllables were the same by pressing keypad buttons "1" for yes and "2" for no. Participants received visual feedback for correct (the green flashing German word for correct: "Richtig") and incorrect (the red flashing German word for incorrect: "Falsch") answers. The speaker familiarisation consisted of three 2:50 min runs (each run contained one presentation and one test phase). If participants scored below 80% on the last run, they performed an additional run until they scored above 80%. All participants exceeded the 80% cut-off value.

The experiments were programmed in the Matlab Psychophysics Toolbox [Psychtoolbox-3, www.psychtoolbox.com (Brainard, 1997)] running on Matlab 8.6 (The Mathworks Inc., Natick, MA, USA) on Ubuntu Linux 16.04 (Canonical Ltd., London, UK). The sound was delivered through a MrConfon amplifier and headphones (manufactured 2008; MrConfon GmbH, Magdeburg, Germany).

4.4 Data Acquisition and Processing

MRI data were acquired using a Siemens Magnetom 7 T scanner (Siemens AG, Erlangen, Germany) with an 8-channel head coil. We convened on the 8-channel coil, due to its spaciousness which allowed the use of higher quality headphones (manufactured 2008; MrConfon GmbH, Magdeburg, Germany). Functional MRI data were acquired using echo-planar imaging (EPI) sequences. We used partial brain coverage with 30 slices. The volume was oriented in parallel to the superior temporal gyrus such that the slices encompassed the MGB, the inferior colliculi (IC), and the Heschl's gyrus.

The EPI sequences had the following acquisition parameters: TR = 1600 ms, TE = 19 ms, flip angle 65°, GRAPPA (Griswold et al., 2002) with acceleration factor 2, 33% phase oversampling, matrix size 88, field of view (FoV) of 132 mm x 132 mm, phase partial Fourier 6/8, voxel size 1.5 mm isotropic resolution, interleaved acquisition, anterior to posterior phase-encode direction. The first three runs consisted of 485 volumes (12:56 min), and the fourth run consisted of 245 volumes (6:32 min). During functional MRI data acquisition, we also acquired physiological values (heart rate, and respiration rate) using a BIOPAC MP150 system (BIOPAC Systems Inc., Goleta, CA, USA).

To address geometric distortions in EPI images we recorded gradient echo based field maps which had the following acquisition parameters: TR = 1500 ms, TE1 = 6.00 ms, TE2 = 7.02 ms, flip angle 60°, 0% phase oversampling, matrix size 100, FoV 220 mm x 220 mm, phase partial Fourier off, voxel size 2.2 mm isotropic resolution, interleaved acquisition, anterior to posterior phase-encode direction. Resulting images from field map recordings were two magnitude images and one phase difference image.

Structural images were recorded using an MP2RAGE (Marques et al., 2010) T1 protocol: 700 µm isotropic resolution, TE = 2.45ms, TR = 5000 ms, TI1 = 900 ms, TI2 = 2750 ms, flip angle 1 = 5°, flip angle 2 = 3°, FoV 224 mm × 224 mm, GRAPPA acceleration factor 2, duration 10:57 min.

4.5 Behavioural Data Analysis

Button presses (hits, misses) were binomially distributed, and were thus modeled using a binomial logistic regression which predicts the probability of correct button presses based on four independent variables (speech task/noise, speaker task/noise, speech task/clear, speaker task/clear) in a Bayesian framework (McElreath, 2018).

To pool over participants and runs we modelled the correlation between intercepts and slopes. For the model implementation and data analysis, we used PyMC3 3.5 (Salvatier et al., 2016), a probabilistic programming package for Python 3.6. We sampled with a No-U-

Turn Sampler (Hoffman and Gelman, 2014) with four parallel chains. Per chain, we had 5,000 samples with 5,000 as warm-up. There were the following effects of interest: main effects (clear - noise, speech task - speaker task), the interaction (speech task/ noise - speaker task/ noise) - (speech task/ clear - speaker task/ clear), simple main effects (speech task/ noise - speaker task/ noise, speech task/ clear - speaker task/ clear). For the effects of interest, we calculated means from the posterior distributions and 95% highest posterior density intervals (HDP). The HPD is the probability that the mean lies within the interval (Gelman et al., 2013; McElreath, 2018), this means that we are 95% sure the mean lies within the specified interval bounds. If the posterior probability distribution of odds ratios does not strongly overlap one (i.e., the HPD excludes one), then it is assumed that there is a detectable difference between conditions (Bunce and McElreath, 2017; McElreath, 2018).

The predictors included in the behavioural data model were: task (x_S:1 = speech task, 0 = speaker task), and background noise (x_N: 1 = noise, 0 = clear). We also included the two-way interaction of task and noise condition. Because data were collected across participants and runs, we included random effects for both of these in the logistic model. Furthermore, since ~11% of the data exhibited ceiling effects (i.e., some participants scored at the highest possible level) which would result in underestimated means and standard deviations (Uttl, 2005), we treated these data as right-censored and modeled them using a Potential class (Jordan, 1998; Lauritzen et al., 1990) as implemented in PyMC3. This method integrates out the censored values using the log of the complementary normal cumulative distribution function (Gelman et al., 2013; McElreath, 2018). In essence, we sampled twice, once for the observed values without the censored data points, and once for the censored values only.

The model is described below.

$$L_{i,j} \sim Binomial(1, p_{i,j})$$

$$p_{i,j} = \begin{cases} p^*_{i,j}, & \text{for } p^*_{i,j} < c \\ c, & \text{for } p^*_{i,j} \geq c \end{cases}$$

$$logit(p^*_{i,j}) = A_{i,j} + B_{S,i,j}x_S + B_{N,i,j}x_N + B_{SN,i,j}x_S x_N, \text{for } i = 1,\ldots,I; j = 1,\ldots,J$$

$$A_{i,j} = \alpha + \alpha_{participant[i]} + \alpha_{run[j]}$$

$$B_{S,i,j} = \beta_S + \beta_{S,participant[i]} + \beta_{S,run[j]}$$

$$B_{N,i,j} = \beta_N + \beta_{N,participant[i]} + \beta_{N,run[j]}$$

$$B_{SN,i,j} = \beta_{SN} + \beta_{SN,participant[i]} + \beta_{SN,run[j]}$$

$$\begin{bmatrix} \alpha_{participant} \\ \beta_{S,participant} \\ \beta_{N,participant} \\ \beta_{SN,participant} \end{bmatrix} \sim MVNormal\left(\begin{bmatrix} \alpha \\ \beta_S \\ \beta_N \\ \beta_{SN} \end{bmatrix}, S_{participant}\right)$$

$$\begin{bmatrix} \alpha_{run} \\ \beta_{S,run} \\ \beta_{N,run} \\ \beta_{SN,run} \end{bmatrix} \sim MVNormal\left(\begin{bmatrix} \alpha \\ \beta_S \\ \beta_N \\ \beta_{SN} \end{bmatrix}, S_{run}\right)$$

$$S_{subject} = \begin{bmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta_S} & 0 & 0 \\ 0 & 0 & \sigma_{\beta_N} & 0 \\ 0 & 0 & 0 & \sigma_{\beta_{SN}} \end{bmatrix} R_{subject} \begin{bmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta_S} & 0 & 0 \\ 0 & 0 & \sigma_{\beta_N} & 0 \\ 0 & 0 & 0 & \sigma_{\beta_{SN}} \end{bmatrix}$$

$$S_{run} = \begin{bmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta_S} & 0 & 0 \\ 0 & 0 & \sigma_{\beta_N} & 0 \\ 0 & 0 & 0 & \sigma_{\beta_{SN}} \end{bmatrix} R_{run} \begin{bmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta_S} & 0 & 0 \\ 0 & 0 & \sigma_{\beta_N} & 0 \\ 0 & 0 & 0 & \sigma_{\beta_{SN}} \end{bmatrix}$$

$$\alpha \sim Normal(0,5)$$

$$\beta_S \sim Normal(0,5)$$

$$\beta_N \sim Normal(0,5)$$

$$\beta_{SN} \sim Normal(0,5)$$

$$(\sigma_{participant}, \sigma_{run}) \sim HalfCauchy(1)$$

$$\sigma_{corr,participant} \sim HalfCauchy(1)$$

$$\sigma_{corr,run} \sim HalfCauchy(1)$$

$$R_{participant} \sim LKJcorr(4, \sigma_{corr,participant})$$

$$R_{run} \sim LKJcorr(4, \sigma_{corr,run})$$

I represents the participants and J the runs. The model is compartmentalized into sub-models for the intercepts and slopes. $A_{i,j}$ is the sub-model for the intercept for observations i, j. Similarly, $B_{S,i,j}$, $B_{N,i,j}$, and $B_{SN,i,j}$ are the sub-models for the speech task – speaker task slope, clear-noise slope and the interaction slope, respectively; $S_{subject}/S_{run}$ are the covariance matrices for participant/run. $R_{subject}/R_{run}$ are the priors for the correlation matrices modelled as LKJ probability densities (Lewandowski et al., 2009). Weakly informative priors for the intercept (α) and additional coefficients (e.g., β_S), random effects for participant and run ($\beta_{S,subject}$, $\beta_{S,run}$), and multivariate priors for participants and runs identify the model by constraining the position of $p_{i,j}$ to reasonable values. Here we used normal distributions as priors. Furthermore, $p_{i,j}$ is defined as the ramp function equal to the proportion of hits when these are known and below the ceiling (c), and set to the ceiling if they are equal to or greater than the ceiling *c.*

4.6 Functional MRI Data Analysis

4.6.1 Preprocessing of fMRI data

The MP2RAGE images were first segmented using SPM's segment function (SPM 12, version 12.6906, Wellcome Trust Centre for Human Neuroimaging, UCL, UK, http://www.fil.ion.ucl.ac.uk/spm) running on Matlab 8.6 (The Mathworks Inc., Natick, MA, USA) in Ubuntu Linux 16.04 (Canonical Ltd., London, UK). The resulting grey and white matter segmentations were summed and binarised to remove voxels that contain air, scalp, skull and cerebrospinal fluid from structural images using the ImCalc function of SPM.

We used the template image created for a previous study (Mihai et al., 2019) using structural MP2RAGE images from the 28 participants of that study. We chose this template since 15 participants in the current study are included in this image, and the vMGB mask (described below) is in the same space as the template image. The choice of this common template reduces warping artefacts, which would be introduced with a different template, as both the vMGB mask and the functional data of the present study would need to be warped to a common space. The template was created and registered to MNI space with ANTs (Avants et al., 2008) and the MNI152 template provided by FSL 5.0.8 (Smith et al., 2004). All MP2RAGE images were preprocessed with Freesurfer (Fischl et al., 2004; Han and Fischl, 2007) using the recon-all command to obtain boundaries between grey and white matter, which were later used in the functional to structural registration step.

Preprocessing and statistical analyses pipelines were coded in nipype 1.1.2 (Gorgolewski et al., 2011). Head motion and susceptibility distortion by movement interaction of functional runs were corrected using the Realign and Unwarp method (Andersson et al., 2001) in SPM 12. This step also makes use of a voxel displacement map (VDM), which addresses the problem of geometric distortions in EPI caused by magnetic field inhomogeneity. The VDM was calculated using field map recordings, which provided the absolute value and the phase difference image files, using the FieldMap Toolbox (Jezzard and Balaban, 1995) of SPM 12. Outlier runs were detected using ArtifactDetect (composite threshold of translation and rotation: 1; intensity Z-threshold: 3; global threshold: 8; https://www.nitrc.org/projects/artifact_detect/). Coregistration matrices for realigned functional runs per participant were computed based on each participant's structural image using Freesurfer's BBregister function (register mean EPI image to T1). We used a whole-brain EPI volume as an intermediate file in the coregistration step to avoid registration problems due to the limited FoV of the functional runs. Warping using coregistration matrices (after conversion to the ITK coordinate system) and resampling to 1 mm isovoxel was performed using ANTs. Before model creation, we smoothed the data in SPM12 using a 1 mm kernel at full-width half-maximum.

4.6.2 Physiological data

Physiological data (heart rate and respiration rate) were processed by the PhysIO Toolbox (Kasper et al., 2017) to obtain Fourier expansions of each, in order to enter these into the design matrix (see section 4.6.3 Testing our hypothesis in the left vMGB). Since heartbeats and respiration result in undesired cortical and subcortical artefacts, regressing these out increases the specificity of fMRI responses to the task of interest (Kasper et al., 2017). These artefacts occur in abundance around the thalamus (Kasper et al., 2017).

4.6.3 Testing our hypothesis in the left vMGB

Models were set up in SPM 12 using the native space data for each participant. We modelled five conditions of interest: speech task/noise, speaker task/noise, speech task/clear, speaker task/clear, and task instruction. Onset times and durations were used to create boxcar functions, which were convolved with the hemodynamic response function (HRF) provided by SPM 12. The design matrix also included the following nuisance regressors: three cardiac, four respiratory, and a cardiac × respiratory interaction regressor. We additionally entered the outlier regressors from the ArtifactDetect step.

Parameter estimates were computed for each condition at the first level using restricted maximum likelihood (REML) as implemented in SPM 12. Parameter estimates for each of the four conditions of interest (speech task/noise, speaker task/noise, speech task/clear, speaker task/clear) were registered to the MNI structural template using a two-step registration in ANTs. First, a quick registration was performed on the whole head using rigid, affine and diffeomorphic transformations (using Symmetric Normalization, SyN), and the mutual information similarity metric. Second, the high-quality registration was confined to the volume that was covered by the 30 slices of the EPI images. These volumes include the IC, MGB, and primary and secondary auditory cortices. This step used affine and SyN transformations and mean squares and neighbourhood cross-correlation similarity

measures. We performed the registration to MNI space by linearly interpolating the contrast images using the composite transforms from the high-quality registration.

We extracted parameter estimates for each of the four conditions of interest per participant, averaged over all voxels from the region of interest, i.e., the left vMGB. To locate the left vMGB, we used the mask from (Mihai et al., 2019), which included 15 of the 17 participants of the present study.

We analysed the extracted parameter estimates in a Bayesian framework (McElreath, 2018). The model was implemented in PyMC3 with a No-U-Turn Sampler with four parallel chains. Per chain, we sampled posterior distributions which had 5000 samples with 5000 as warm-up. The predictors included in the model were: task (x_S: 1 = speech task, 0 = speaker task), and background noise (x_N: 1 = noise, 0 = clear). We also included the two-way interaction of task and noise condition. Because data were collected across participants, it was reasonable to include random effects. To pool over participants, we modelled the correlation between intercepts and slopes over participants. The interaction model is described below.

$$L_i \sim T(\mu_i, \nu, \lambda)$$

$$\mu_i = A_i + B_{S,i}x_S + B_{N,i}x_N + B_{SN,i}x_S x_N, \text{for } i = 1, \ldots, I$$

$$A_i = \alpha + \alpha_{participant[i]}$$

$$B_{S,i} = \beta_S + \beta_{S,participant[i]}$$

$$B_{N,i} = \beta_N + \beta_{N,participant[i]}$$

$$B_{SN,i} = \beta_{SN} + \beta_{SN,participant[i]}$$

$$\begin{bmatrix} \alpha_{participant} \\ \beta_{S,participant} \\ \beta_{N,participant} \\ \beta_{SN,participant} \end{bmatrix} \sim MVNormal\left(\begin{bmatrix} \alpha \\ \beta_S \\ \beta_N \\ \beta_{SN} \end{bmatrix}, S\right)$$

$$S = \begin{bmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta_S} & 0 & 0 \\ 0 & 0 & \sigma_{\beta_N} & 0 \\ 0 & 0 & 0 & \sigma_{\beta_{SN}} \end{bmatrix} R \begin{bmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta_S} & 0 & 0 \\ 0 & 0 & \sigma_{\beta_N} & 0 \\ 0 & 0 & 0 & \sigma_{\beta_{SN}} \end{bmatrix}$$

$$\alpha \sim T(0,1,3)$$

$$\beta_S \sim T(0,1,3)$$

$$\beta_N \sim T(0,1,3)$$

$$\beta_{SN} \sim T(0,1,3)$$

$$(\sigma_{participant}) \sim HalfCauchy(1)$$

$$\sigma_{corr} \sim HalfCauchy(1)$$

$$R \sim LKJcorr(4, \sigma_{corr})$$

$$\nu \sim Exponential(1/29) + 1$$

$$\sigma \sim HalfCauchy(2)$$

$$\lambda = \sigma^{-2}$$

I represents the participants. The model is compartmentalized into sub-models for the intercepts and slopes. A_i is the sub-model for the intercept for observations i. Similarly, $B_{S,i}$, $B_{N,i}$, and $B_{SN,i}$ are the sub-models for the speech task -speaker task slope, clear-noise slope and the interaction slope, respectively; S is the covariance matrix and R is the prior for the correlation matrix modelled as an LKJ probability density (Lewandowski et al., 2009). Weakly informative priors for the intercept (α) and additional coefficients (e.g., β_S), random effects for participant ($\beta_{S,subject}$), and multivariate priors for participants identify the model by constraining the position of μ_i to reasonable values. Here we used Student's-T distributions as priors.

From the model output, we calculated posterior distributions for each condition of interest (speech task/noise, speaker task/ noise, speech task/clear, speaker task/clear). Posterior distributions, in comparison to point estimates, have the advantage of quantifying

uncertainty about each parameter. We summarised each posterior distribution using the mean as a point estimate (posterior mean) together with a 95% highest posterior density interval (HPD). The HPD is the probability that the mean lies within the interval (Gelman et al., 2013; McElreath, 2018), e.g., we are 95% sure the mean lies within the specified interval bounds. We computed the following contrasts of interest: interaction (speech task/noise – speaker task/noise) – (speech task/clear – speaker task/clear); simple main effects (speech task/noise – speaker task/noise), (speech task/clear – speaker task/clear); main effect of task (speech task – speaker task). Differences between conditions were converted to effect sizes [Hedges g* (Hedges and Olkin, 1985)]. Hedges g*, like Cohen's d (Cohen, 1988), is a population parameter that computes the difference in means between two variables normalised by the pooled standard deviation with the benefit of correcting for small sample sizes. Based on Cohen (1988), we interpreted effect sizes on a spectrum ranging from small ($g^* \approx 0.2$), to medium ($g^* \approx 0.5$), to large ($g^* \approx 0.8$), and beyond. If the HPD did not overlap zero, we considered this to be a robust effect (Bunce and McElreath, 2017; McElreath, 2018). However, we caution readers that if the HPD includes zero, it does not mean that the effect is missing (Amrhein et al., 2019). Instead, we quantify and interpret the magnitude (by the point estimate) and its uncertainty (by the HPD) provided by the data and our assumptions (Anderson, 2019).

4.6.4 Test for replication of previous findings: Main effect of task in the left and right vMGB

We tested for replication of previous studies that have found a task-dependent modulation (speech - speaker task) in the left and the right MGB (Díaz et al., 2012; von Kriegstein et al., 2008). To do this we adopted the same procedure as described in section 4.6.3.1. For the right vMGB mask we used a mask described in (Mihai et al., 2019). Posterior means and 95% HPD were used to summarise results.

4.6.5 Test for replication of previous findings: Correlation between the main effect of task and speech recognition performance in the left vMGB

To test for the correlation between the main effect of task in the BOLD response and the speech recognition performance across participants, we performed a Pearson's correlation calculation between the estimated parameters from the Bayesian model across subjects in

the left vMGB for the speech - speaker contrast together with the proportion of hits in the speech task. Additionally, we performed the correlation between the simple main effect of task (speech task/clear – speaker task/clear) and the speech task/clear accuracy score.

4.6.6 Meta-analysis of the correlation (speech - speaker task correlated with speech accuracy score) in the left MGB

The lack of statistical significance for the correlation between speech - speaker task contrast and the proportion of hits in the speech task raised the question whether the correlation effect in the left MGB is different from the ones reported previously (Díaz et al., 2012; Mihai et al., 2019; von Kriegstein et al., 2008). We performed a random-effects meta-analysis to test whether the lack of task-dependent modulation in the present study was different from other studies that have reported a correlation in the MGB. We included five studies in the meta-analysis: two experiments from (von Kriegstein et al., 2008), the control participants of Díaz et al. (2012), the result of (Mihai et al., 2019), and the current study. Pearson correlation values were Fisher-z transformed (Fisher, 1915) to z-values and standard errors. These were then entered into a random-effects model that was estimated with restricted maximum likelihood using JASP 0.9 (jasp-stats.org). The resulting z-value was converted back to a correlation value for easier interpretation.

4.6.7 Analyses of the left inferior colliculus

To analyse the task × noise interaction and the main effect of task in the bilateral IC we used the same analysis procedures as described for the left vMGB (see section 4.6.3 Testing our hypothesis in the left vMGB). As region of interest, we used the IC masks described in (Mihai et al., 2019). Furthermore, to analyse the correlation (speech - speaker task correlated with speech accuracy score) in the left IC, we followed the same analysis procedures as for the left vMGB (see section 4.6.5 Test for replication of previous findings: Correlation between the main effect of task and speech recognition performance in the left vMGB).

Additionally, we computed a meta-analysis for the correlation between speech - speaker task contrast and the proportion of hits in the speech task in the left IC. We focused on the

left IC since previous studies only reported correlations in the left IC (Díaz et al., 2012; von Kriegstein et al., 2008). We included correlation coefficients from five studies: four previous studies [two experiments from von Kriegstein et al., (2008), the control participants of Díaz et al., (2012), one experiment from Mihai et al., (Mihai et al., 2019)] and the current study. Díaz et al. (2012) did not report the correlation coefficient in the IC. We took this value from the original study data that was part of our research group's archive. Pearson correlation coefficients were Fisher-z transformed (Fisher, 1915) to z-values and standard errors. These were then entered into a random-effects model that was estimated with restricted maximum likelihood using JASP 0.9 (jasp-stats.org). The resulting z-value was converted back to a correlation value for a more straightforward interpretation.

Acknowledgements

We thank the participants for taking part in the study.

Funding

The study was funded by the European Research Council ERC Consolidator Grant SENSOCOM (647051).

Supplementary Material

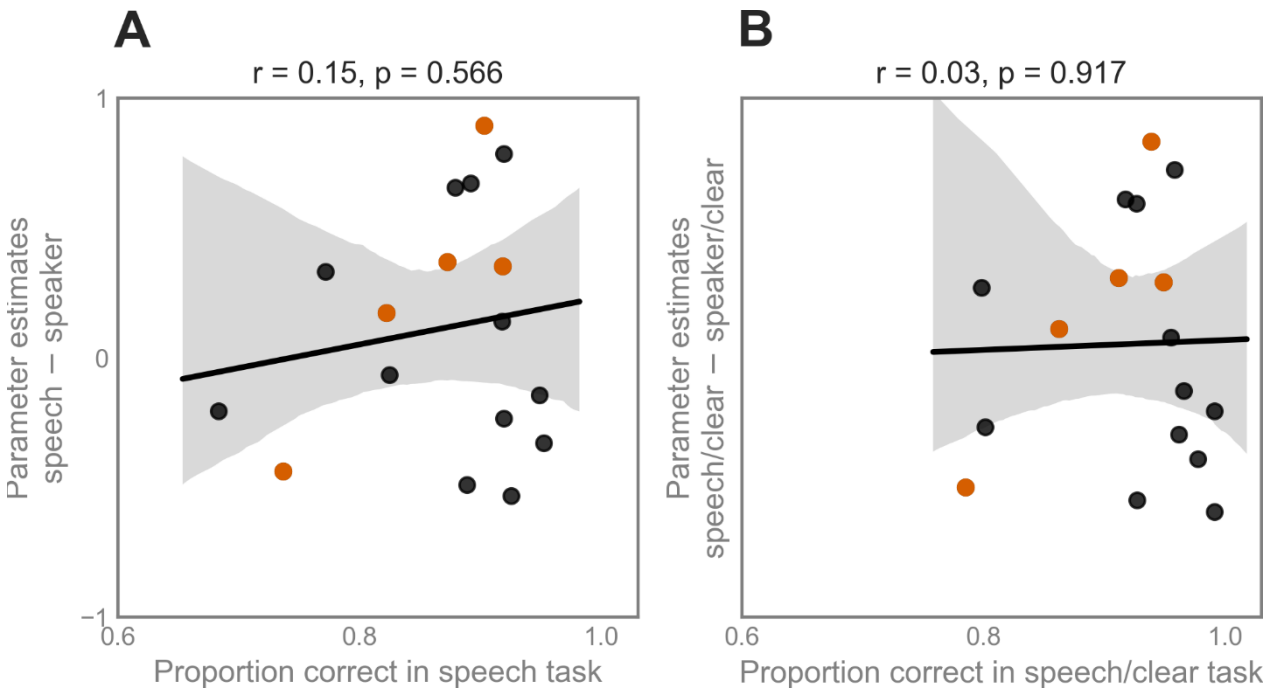

Figure S1. (A) Correlation between the contrast Speech – Speaker task and the proportion of hits in the speech task. (B) Correlation between the contrast speech/clear – speaker/clear task and the proportion of hits in the speech/clear task. Orange points denote those participants, that scored poorly on the reading speed and comprehension task. Most data points are close to the ceiling on the right of the behavioural score.

References

Anderson, Samira, and Nina Kraus. 2010. "Sensory-Cognitive Interaction in the Neural Encoding of Speech in Noise: A Review". *Journal of the American Academy of Audiology* 21 (9). American Academy of AAdams RA, Shipp S, Friston KJ. 2013. Predictions not commands: active inference in the motor system. *Brain Struct Funct* **218**:611–643. doi:10.1007/s00429-012-0475-5

Adank P. 2012. The neural bases of difficult speech comprehension and speech production: Two Activation Likelihood Estimation (ALE) meta-analyses. *Brain and Language* **122**:42–54. doi:10.1016/j.bandl.2012.04.014

Alavash M, Tune S, Obleser J. 2019. Modular reconfiguration of an auditory control brain network supports adaptive listening behavior. *PNAS* **116**:660–669. doi:10.1073/pnas.1815321116

Alcántara JI, Weisblatt EJL, Moore BCJ, Bolton PF. 2004. Speech-in-noise perception in high-functioning individuals with autism or Asperger's syndrome. *Journal of Child Psychology and Psychiatry* **45**:1107–1114. doi:10.1111/j.1469-7610.2004.t01-1-00303.x

Amrhein V, Greenland S, McShane B. 2019. Scientists rise up against statistical significance. *Nature* **567**:305. doi:10.1038/d41586-019-00857-9

Anderson AA. 2019. Assessing Statistical Results: Magnitude, Precision, and Model Uncertainty. *The American Statistician* **73**:118–121. doi:10.1080/00031305.2018.1537889

Anderson S, Kraus N. 2010. Sensory-Cognitive Interaction in the Neural Encoding of Speech in Noise: A Review. *Journal of the American Academy of Audiology* **21**:575–585. doi:10.3766/jaaa.21.9.3

Andersson JLR, Hutton C, Ashburner J, Turner R, Friston K. 2001. Modeling Geometric Deformations in EPI Time Series. *NeuroImage* **13**:903–919. doi:10.1006/nimg.2001.0746

Avants BB, Epstein CL, Grossman M, Gee JC. 2008. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, Special Issue on The Third International Workshop on Biomedical Image Registration – WBIR 2006 **12**:26–41. doi:10.1016/j.media.2007.06.004

Banno H, Hata H, Morise M, Takahashi T, Irino T, Kawahara H. 2007. Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation. *Acoustical Science and Technology* **28**:140–146. doi:10.1250/ast.28.140

Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E. 2001. The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Malesand Females, Scientists and Mathematicians. *J Autism Dev Disord* **31**:5–17. doi:10.1023/A:1005653411471

Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. 2012. Canonical Microcircuits for Predictive Coding. *Neuron* **76**:695–711. doi:10.1016/j.neuron.2012.10.038

Bellis TJ, Bellis JD. 2015. Central auditory processing disorders in children and adults. *Handb Clin Neurol* **129**:537–556. doi:10.1016/B978-0-444-62630-1.00030-5

Best V, Gallun FJ, Carlile S, Shinn-Cunningham BG. 2007. Binaural interference and auditory grouping. *The Journal of the Acoustical Society of America* **121**:1070–1076. doi:10.1121/1.2407738

Bidelman GM. 2018. Subcortical sources dominate the neuroelectric auditory frequency-following response to speech. *NeuroImage* **175**:56–69. doi:10.1016/j.neuroimage.2018.03.060

Bishop CW, Miller LM. 2008. A Multisensory Cortical Network for Understanding Speech in Noise. *Journal of Cognitive Neuroscience* **21**:1790–1804. doi:10.1162/jocn.2009.21118

Bland JM, Altman DG. 2011. Correlation in restricted ranges of data. *BMJ* **342**:d556. doi:10.1136/bmj.d556

Boets B, Wouters J, van Wieringen A, Ghesquière P. 2007. Auditory processing, speech perception and phonological ability in pre-school children at high-risk for dyslexia: A longitudinal study of the auditory temporal processing theory. *Neuropsychologia* **45**:1608–1620. doi:10.1016/j.neuropsychologia.2007.01.009

Brainard DH. 1997. The Psychophysics Toolbox. *Spatial Vision* **10**:433–436. doi:10.1163/156856897X00357

Bregman AS. 1994. Auditory scene analysis: The perceptual organization of sound. MIT press.

Brokx JP, Noteboom SG. 1982. Intonation and the perceptual segregation of competing voices. *J Phonetics* **10**:23–36.

Bronkhorst AW. 2015. The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Atten Percept Psychophys* **77**:1465–1487. doi:10.3758/s13414-015-0882-9

Bunce JA, McElreath R. 2017. Interethnic Interaction, Strategic Bargaining Power, and the Dynamics of Cultural Norms. *Hum Nat* **28**:434–456. doi:10.1007/s12110-017-9297-8

Carhart R, Johnson C, Goodman J. 1975. Perceptual masking of spondees by combinations of talkers. *The Journal of the Acoustical Society of America* **58**:S35–S35. doi:10.1121/1.2002082

Chandrasekaran B, Hornickel J, Skoe E, Nicol T, Kraus N. 2009. Context-dependent encoding in the human auditory brainstem relates to hearing speech in noise: Implications for developmental dyslexia. *Neuron* **64**:311–319. doi:10.1016/j.neuron.2009.10.006

Chandrasekaran B, Kraus N. 2010a. Music, Noise-Exclusion, and Learning. *MUSIC PERCEPT* **27**:297–306. doi:10.1525/mp.2010.27.4.297

Chandrasekaran B, Kraus N. 2010b. The scalp-recorded brainstem response to speech: Neural origins and plasticity. *Psychophysiology* **47**:236–246. doi:10.1111/j.1469-8986.2009.00928.x

Chandrasekaran B, Skoe E, Kraus N. 2014. An Integrative Model of Subcortical Auditory Plasticity. *Brain Topogr* **27**:539–552. doi:10.1007/s10548-013-0323-9

Chen JJ. 2003. COMMUNICATING COMPLEX INFORMATION: THE INTERPRETATION OF STATISTICAL INTERACTION IN MULTIPLE LOGISTIC REGRESSION ANALYSIS. *Am J Public Health* **93**:1376–1377.

Cherry EC. 1953. Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America* **25**:975–979. doi:10.1121/1.1907229

Cohen J. 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Lawrence Erlbaum Associates. doi:10.1016/C2013-0-10517-X

Darwin CJ, Hukin RW. 2000. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *The Journal of the Acoustical Society of America* **107**:970–977. doi:10.1121/1.428278

Davis MH, Johnsrude IS. 2007. Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research* **229**:132–147. doi:10.1016/j.heares.2007.01.014

Denckla MB, Rudel RG. 1976. Rapid 'automatized' naming (R.A.N.): Dyslexia differentiated from other learning disabilities. *Neuropsychologia* **14**:471–479. doi:10.1016/0028-3932(76)90075-0

Díaz B, Hintz F, Kiebel SJ, Kriegstein K von. 2012. Dysfunction of the auditory thalamus in developmental dyslexia. *PNAS* **109**:13841–13846. doi:10.1073/pnas.1119828109

Feldman H, Friston K. 2010. Attention, Uncertainty, and Free-Energy. *Front Hum Neurosci* **4**. doi:10.3389/fnhum.2010.00215

Fischl B, Salat DH, van der Kouwe AJW, Makris N, Ségonne F, Quinn BT, Dale AM. 2004. Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, Mathematics in Brain Imaging **23**:S69–S84. doi:10.1016/j.neuroimage.2004.07.016

Fisher RA. 1915. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika* **10**:507–521. doi:10.2307/2331838

Forte AE, Etard O, Reichenbach T. 2017. The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *eLife* **6**:e27203. doi:10.7554/eLife.27203

Friston K. 2005. A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **360**:815–836. doi:10.1098/rstb.2005.1622

Friston K, Kiebel S. 2009. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **364**:1211–1221. doi:10.1098/rstb.2008.0300

Gaudrain E, Li S, Ban VS, Patterson RD. 2009. The Role of Glottal Pulse Rate and Vocal Tract Length in the Perception of Speaker Identity. *Interspeech-2009* 148–151.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2013. Bayesian Data Analysis. Chapman and Hall/CRC. doi:10.1201/b16018

Giraud A-L, Lorenzi C, Ashburner J, Wable J, Johnsrude I, Frackowiak R, Kleinschmidt A. 2000. Representation of the Temporal Envelope of Sounds in the Human Brain. *Journal of Neurophysiology* **84**:1588–1598.

Gordon N, Koenig-Robert R, Tsuchiya N, van Boxtel JJ, Hohwy J. 2017. Neural markers of predictive coding under perceptual uncertainty revealed with Hierarchical Frequency Tagging. *eLife* **6**. doi:10.7554/eLife.22749

Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, Ghosh SS. 2011. Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Front Neuroinform* **5**. doi:10.3389/fninf.2011.00013

Griswold MA, Jakob PM, Heidemann RM, Nittka M, Jellus V, Wang J, Kiefer B, Haase A. 2002. Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic Resonance in Medicine* **47**:1202–1210. doi:10.1002/mrm.10171

Groen WB, van Orsouw L, Huurne N ter, Swinkels S, van der Gaag R-J, Buitelaar JK, Zwiers MP. 2009. Intact Spectral but Abnormal Temporal Processing of Auditory Stimuli in Autism. *J Autism Dev Disord* **39**:742–750. doi:10.1007/s10803-008-0682-3

Gupta S, Bhurchandi KM, Keskar AG. 2016. An efficient noise-robust automatic speech recognition system using artificial neural networks2016 International Conference on Communication and Signal Processing (ICCSP). Presented at the 2016 International Conference on Communication and Signal Processing (ICCSP). pp. 1873–1877. doi:10.1109/ICCSP.2016.7754495

Han X, Fischl B. 2007. Atlas Renormalization for Improved Brain MR Image Segmentation Across Scanner Platforms. *IEEE Transactions on Medical Imaging* **26**:479–486. doi:10.1109/TMI.2007.893282

Hedges LV, Olkin I. 1985. Statistical Methods for Meta-Analysis. Elsevier Science.

Hesselmann G, Sadaghiani S, Friston KJ, Kleinschmidt A. 2010. Predictive Coding or Evidence Accumulation? False Inference and Neuronal Fluctuations. *PLOS ONE* **5**:e9926. doi:10.1371/journal.pone.0009926

Hickok G, Poeppel D. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience* **8**:393–402. doi:10.1038/nrn2113

Hoffman MD, Gelman A. 2014. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**:1593–1623.

Huang Y, Rao RPN. 2011. Predictive coding. *WIREs Cogn Sci* **2**:580–593. doi:10.1002/wcs.142

Iliadou V (Vivian), Ptok M, Grech H, Pedersen ER, Brechmann A, Deggouj N, Kiese-Himmel C, Śliwińska-Kowalska M, Nickisch A, Demanez L, Veuillet E, Thai-Van H, Sirimanna T, Callimachou M, Santarelli R, Kuske S, Barajas J, Hedjever M, Konukseven O, Veraguth D, Stokkereit Mattsson T, Martins JH, Bamiou D-E. 2017. A European Perspective on

Auditory Processing Disorder-Current Knowledge and Future Research Focus. *Front Neurol* **8**. doi:10.3389/fneur.2017.00622

Jezzard P, Balaban RS. 1995. Correction for geometric distortion in echo planar images from B0 field variations. *Magnetic Resonance in Medicine* **34**:65–73. doi:10.1002/mrm.1910340111

Jordan MI, editor. 1998. Learning in Graphical Models, Nato Science Series D: Springer Netherlands.

Kasper L, Bollmann S, Diaconescu AO, Hutton C, Heinzle J, Iglesias S, Hauser TU, Sebold M, Manjaly Z-M, Pruessmann KP, Stephan KE. 2017. The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data. *Journal of Neuroscience Methods* **276**:56–72. doi:10.1016/j.jneumeth.2016.10.019

Kiebel SJ, Daunizeau J, Friston KJ. 2008. A Hierarchy of Time-Scales and the Brain. *PLOS Computational Biology* **4**:e1000209. doi:10.1371/journal.pcbi.1000209

Knill DC, Pouget A. 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences* **27**:712–719. doi:10.1016/j.tins.2004.10.007

Kreitewolf J, Gaudrain E, von Kriegstein K. 2014. A neural mechanism for recognizing speech spoken by different speakers. *NeuroImage* **91**:375–385. doi:10.1016/j.neuroimage.2014.01.005

Krupa DJ, Ghazanfar AA, Nicolelis MAL. 1999. Immediate thalamic sensory plasticity depends on corticothalamic feedback. *PNAS* **96**:8200–8205. doi:10.1073/pnas.96.14.8200

Lauritzen SL, Dawid AP, Larsen BN, Leimer H-G. 1990. Independence properties of directed markov fields. *Networks* **20**:491–505. doi:10.1002/net.3230200503

Lee CC. 2013. Thalamic and cortical pathways supporting auditory processing. *Brain and Language* **126**:22–28. doi:10.1016/j.bandl.2012.05.004

Lewandowski D, Kurowicka D, Joe H. 2009. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* **100**:1989–2001. doi:10.1016/j.jmva.2009.04.008

Llano DA, Sherman SM. 2008. Evidence for nonreciprocal organization of the mouse auditory thalamocortical-corticothalamic projection systems. *Journal of Comparative Neurology* **507**:1209–1227. doi:10.1002/cne.21602

Marques JP, Kober T, Krueger G, van der Zwaag W, Van de Moortele P-F, Gruetter R. 2010. MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. *NeuroImage* **49**:1271–1281. doi:10.1016/j.neuroimage.2009.10.002

Mattys SL, Davis MH, Bradlow AR, Scott SK. 2012. Speech recognition in adverse conditions: A review. *Language and Cognitive Processes* **27**:953–978. doi:10.1080/01690965.2012.705006

McElreath R. 2018. Statistical Rethinking : A Bayesian Course with Examples in R and Stan. Chapman and Hall/CRC. doi:10.1201/9781315372495

Mihai PG, Moerel M, de Martino F, Trampel R, Kiebel S, von Kriegstein K. 2019. Modulation of tonotopic ventral medial geniculate body is behaviorally relevant for speech recognition. *eLife* **8**:e44837. doi:10.7554/eLife.44837

Moore BCJ, Peters RW, Glasberg BR. 1985. Thresholds for the detection of inharmonicity in complex tones. *The Journal of the Acoustical Society of America* **77**:1861–1867. doi:10.1121/1.391937

Müller-Axt C, Anwander A, von Kriegstein K. 2017. Altered Structural Connectivity of the Left Visual Thalamus in Developmental Dyslexia. *Current Biology* **27**:3692-3698.e4. doi:10.1016/j.cub.2017.10.034

Mumford D. 1992. On the computational architecture of the neocortex. *Biol Cybern* **66**:241–251. doi:10.1007/BF00198477

Parbery-Clark A, Skoe E, Lam C, Kraus N. 2009. Musician Enhancement for Speech-In-Noise. *Ear and Hearing* **30**:653. doi:10.1097/AUD.0b013e3181b412e9

Parikh G, Loizou PC. 2005. The influence of noise on vowel and consonant cues. *The Journal of the Acoustical Society of America* **118**:3874–3888. doi:10.1121/1.2118407

Peelle JE. 2018. Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior. *Ear and Hearing* **39**:204. doi:10.1097/AUD.0000000000000494

Price CJ. 2012. A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, 20 YEARS OF fMRI **62**:816–847. doi:10.1016/j.neuroimage.2012.04.062

Qian Y, Bi M, Tan T, Yu K. 2016. Very Deep Convolutional Neural Networks for Noise Robust Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **24**:2263–2276. doi:10.1109/TASLP.2016.2602884

Rouiller EM, de Ribaupierre F. 1985. Origin of afferents to physiologically defined regions of the medial geniculate body of the cat: ventral and dorsal divisions. *Hearing Research* **19**:97–114. doi:10.1016/0378-5955(85)90114-5

Saffran JR. 2003. Statistical Language Learning: Mechanisms and Constraints. *Curr Dir Psychol Sci* **12**:110–114. doi:10.1111/1467-8721.01243

Salvatier J, Wiecki TV, Fonnesbeck C. 2016. Probabilistic programming in Python using PyMC3. *PeerJ Comput Sci* **2**:e55. doi:10.7717/peerj-cs.55

Salvi RJ, Lockwood AH, Frisina RD, Coad ML, Wack DS, Frisina DR. 2002. PET imaging of the normal human auditory system: responses to speech in quiet and in background noise. *Hearing Research*, Special Issue on the 38th Workshop on Inner Ear Biology , and regular research papers **170**:96–106. doi:10.1016/S0378-5955(02)00386-6

Sayles M, Winter IM. 2008. Ambiguous Pitch and the Temporal Representation of Inharmonic Iterated Rippled Noise in the Ventral Cochlear Nucleus. *J Neurosci* **28**:11925–11938. doi:10.1523/JNEUROSCI.3137-08.2008

Scharenborg O. 2007. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, Bridging the Gap between Human and Automatic Speech Recognition **49**:336–347. doi:10.1016/j.specom.2007.01.009

Schelinski S, Kriegstein K von. 2019. Speech-in-noise recognition and the relation to vocal pitch perception in adults with autism spectrum disorder and typical development. *PsyarXiv*. doi:10.31234/osf.io/u84vd

Schneider W, Schlagmüller M, Ennemoser M. 2007. LGVT 6-12: Lesegeschwindigkeits-und-verständnistest für die Klassen 6-12. Hogrefe Göttingen.

Schoof T, Rosen S. 2016. The Role of Age-Related Declines in Subcortical Auditory Processing in Speech Perception in Noise. *JARO* **17**:441–460. doi:10.1007/s10162-016-0564-x

Scott SK, Rosen S, Wickham L, Wise RJS. 2004. A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception. *The Journal of the Acoustical Society of America* **115**:813–821. doi:10.1121/1.1639336

Selinger L, Zarnowiec K, Via M, Clemente IC, Escera C. 2016. Involvement of the Serotonin Transporter Gene in Accurate Subcortical Speech Encoding. *J Neurosci* **36**:10782–10790. doi:10.1523/JNEUROSCI.1595-16.2016

Semrud-Clikeman M, Guy K, Griffin JD, Hynd GW. 2000. Rapid Naming Deficits in Children and Adolescents with Reading Disabilities and Attention Deficit Hyperactivity Disorder. *Brain and Language* **74**:70–83. doi:10.1006/brln.2000.2337

Seth Anil K., Friston Karl J. 2016. Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**:20160007. doi:10.1098/rstb.2016.0007

Shinn-Cunningham BG, Best V. 2008. Selective Attention in Normal and Impaired Hearing. *Trends in Amplification* **12**:283–299. doi:10.1177/1084713808325306

Shipp S, Adams RA, Friston KJ. 2013. Reflections on agranular architecture: predictive coding in the motor cortex. *Trends in Neurosciences* **36**:706–716. doi:10.1016/j.tins.2013.09.004

Sillito AM, Cudeiro J, Jones HE. 2006. Always returning: feedback and sensory processing in visual cortex and thalamus. *Trends in Neurosciences*, Neural substrates of cognition **29**:307–316. doi:10.1016/j.tins.2006.05.001

Sillito AM, Jones HE, Gerstein GL, West DC. 1994. Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex. *Nature* **369**:479–482. doi:10.1038/369479a0

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM. 2004. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, Mathematics in Brain Imaging **23**:S208–S219. doi:10.1016/j.neuroimage.2004.07.051

Song JH, Skoe E, Banai K, Kraus N. 2010. Perception of Speech in Noise: Neural Correlates. *Journal of Cognitive Neuroscience* **23**:2268–2279. doi:10.1162/jocn.2010.21556

Srinivasan MV, Laughlin SB, Dubs A. 1982. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London Series B Biological Sciences* **216**:427–459.

Strait DL, Parbery-Clark A, Hittner E, Kraus N. 2012. Musical training during early childhood enhances the neural encoding of speech in noise. *Brain and Language* **123**:191–201. doi:10.1016/j.bandl.2012.09.001

Tschentscher N, Ruisinger A, Blank H, Díaz B, Kriegstein K von. 2019. Reduced Structural Connectivity Between Left Auditory Thalamus and the Motion-Sensitive Planum Temporale in Developmental Dyslexia. *J Neurosci* **39**:1720–1732. doi:10.1523/JNEUROSCI.1435-18.2018

Uttl B. 2005. Measurement of Individual Differences: Lessons From Memory Assessment in Research and Clinical Practice. *Psychol Sci* **16**:460–467. doi:10.1111/j.0956-7976.2005.01557.x

Van de Cruys S, Evers K, Van der Hallen R, Van Eylen L, Boets B, de-Wit L, Wagemans J. 2014. Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review* **121**:649–675. doi:10.1037/a0037665

von Kriegstein K, Patterson RD, Griffiths TD. 2008. Task-Dependent Modulation of Medial Geniculate Body Is Behaviorally Relevant for Speech Recognition. *Current Biology* **18**:1855–1859. doi:10.1016/j.cub.2008.10.052

Wang W, Andolina IM, Lu Y, Jones HE, Sillito AM. 2018. Focal Gain Control of Thalamic Visual Receptive Fields by Layer 6 Corticothalamic Feedback. *Cereb Cortex* **28**:267–280. doi:10.1093/cercor/bhw376

Wang X, Lu T, Bendor D, Bartlett E. 2008. Neural coding of temporal information in auditory thalamus and cortex. *Neuroscience*, From Cochlea to Cortex: Recent Advances in Auditory Neuroscience **154**:294–303. doi:10.1016/j.neuroscience.2008.03.065

Wong PCM, Jin JX, Gunasekera GM, Abel R, Lee ER, Dhar S. 2009. Aging and cortical mechanisms of speech perception in noise. *Neuropsychologia* **47**:693–703. doi:10.1016/j.neuropsychologia.2008.11.032

Wong PCM, Uppunda, Ajith K., Parrish, Todd B., Dhar, Sumitrajit. 2008. Cortical Mechanisms of Speech Perception in Noise. *Journal of Speech, Language, and Hearing Research* **51**:1026–1041. doi:10.1044/1092-4388(2008/075)

Yu AJ, Dayan P. 2005. Uncertainty, Neuromodulation, and Attention. *Neuron* **46**:681–692. doi:10.1016/j.neuron.2005.04.026

Ziegler JC, Pech-Georgel C, George F, Lorenzi C. 2009. Speech-perception-in-noise deficits in dyslexia. *Developmental Science* **12**:732–745. doi:10.1111/j.1467-7687.2009.00817.x