Modulation of tonotopic ventral MGB is behaviorally relevant for speech recognition.

Paul Glad Mihai [1,2], Michelle Moerel [3, 4, 5], Federico de Martino [3, 4, 6], Robert Trampel [1], Stefan Kiebel [2], Katharina von Kriegstein [1, 2]*

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

[2]Chair of Cognitive and Clinical Neuroscience, Faculty of Psychology, Technische Universität Dresden, 01187 Dresden, Germany

[3] Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, the Netherlands

[4]Maastricht Brain Imaging Center (MBIC), Maastricht, the Netherlands

[5] Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, the Netherlands

[6] Center for Magnetic Resonance Research, University of Minnesota, Minneapolis, USA

Corresponding author: Paul Glad Mihai, pgmihai+scipubs@gmail.com*

Keywords: ventral MGB, thalamus, speech, 7T fMRI, auditory, top-down modulation

Impact statement: Ultra-high field neuroimaging dissects the ventral medial geniculate body (vMGB) of the primary auditory pathway from other MGB subregions and reveals that vMGB top-down modulation is relevant for speech recognition.

Abstract

Sensory thalami are central sensory pathway stations for information processing. Their role for human cognition and perception, however, remains unclear. Recent evidence suggests an involvement of the sensory thalami in speech recognition. In particular, the auditory thalamus (medial geniculate body, MGB) response is modulated by speech recognition tasks and the amount of this task-dependent modulation is associated with speech recognition abilities. Here we tested the specific hypothesis that this behaviorally relevant modulation is present in the MGB subsection that corresponds to the primary auditory pathway (i.e., the ventral MGB [vMGB]). We used ultra-high field 7T fMRI to identify the vMGB, and found a significant positive correlation between the amount of task-dependent modulation and the speech recognition performance across participants within left vMGB, but not within the other MGB subsections. These results imply that modulation of thalamic driving input to the auditory cortex facilitates speech recognition.

Introduction

Human communication relies on fast and accurate decoding of speech—the most important tool available to us for exchanging information. Understanding the neural decoding mechanisms for speech recognition is important for understanding human brain function (Rauschecker and Scott, 2009), but also for understanding communication disorders such as developmental dyslexia (Galaburda et al., 1994, Müller-Axt et al., 2017). Since the early findings of Wernicke (Wernicke, 1874) neuroscientific models of speech recognition have mainly focused on cerebral cortex mechanisms (Hickok and Poeppel, 2007, Friederici and Gierhan, 2013). Yet, more recently it has been suggested that a full understanding of speech recognition mechanisms might need to take the subcortical sensory pathways – particularly the sensory thalami – into account (Kriegstein et al., 2008, Díaz et al., 2012, Díaz et al., 2018, Chandrasekaran et al., 2009, Chandrasekaran et al., 2011).

The classic view, that the sensory thalamus is a passive relay station has been by-and-large abandoned over the last two decades. First of all, it is well known that there are strong corticofugal projections to the sensory thalamus (Sherman and Guillery, 2006, Winer and Prieto, Lee and Sherman, 2012, Lee and Winer, 2011). Furthermore, experimental evidence in humans and other mammals in the visual as well as the auditory modality has shown that sensory thalamus responses are modulated by attention (Saalmann and Kastner, 2011), percept (Haynes et al., 2005), context (Antunes and Malmierca, 2011, McAlonan et al., 2008, O'Connor et al., 2002), and task (Díaz et al., 2012, von Kriegstein et al., 2008, Díaz et al., 2018).

Based on these findings, the sensory thalamus has become accepted as a structure that is modulated by cognitive demands and is more involved in active information regulation (Saalmann and Kastner, 2011, Haynes et al., 2005, Antunes and Malmierca, 2011, McAlonan et al., 2008, O'Connor et al., 2002, Díaz et al., 2012, von Kriegstein et al., 2008, for a different take see Camarillo et al., 2012).

In the case of speech, previous studies showed a task-dependent modulation in the auditory sensory thalamus for auditory speech recognition, (MGB;von Kriegstein et al., 2008, Díaz et al., 2012) as well as a task-dependent modulation in the visual sensory thalamus for visual speech recognition (LGN;Díaz et al., 2018). In the MGB there were two findings: First, bilateral MGB showed significantly higher responses to an auditory speech recognition task than to control tasks, independent of attentional load or task-difficulty (von Kriegstein et al., 2008, Díaz et al., 2012). Second, the performance level in the auditory speech recognition task was significantly correlated with the task-dependent modulation in the MGB of the left hemisphere across participants (von Kriegstein et al., 2008).

Following the Bayesian brain hypothesis, (Knill and Pouget, 2004, Friston and Kiebel, 2009, Friston, 2005, Kiebel et al., 2008) and based on findings in non-human animals (Krupa et al., 1999, Sillito et al., 1994, Wang et al., 2018b), one possible explanation for the MGB task-dependent modulation for speech is that cerebral cortex areas tune the sensory thalamus depending on behavioral demand, and that this tuning is particularly relevant for fast-varying and predictable stimuli such as speech (von Kriegstein et al., 2008, Díaz et al., 2012). This view entails that the task-dependent modulation occurs already in those parts of the MGB that drive the cerebral cortex representations (von Kriegstein et al., 2008) – the so-called first-order sensory thalamus (Sherman and Guillery, 1998).

The MGB consists of three divisions. Only the ventral MGB (vMGB) can be considered first-order sensory thalamus (Malmierca et al., 2015 [review] , Winer et al., 2005 [review]), as vMGB receives driving inputs from sources that relay information from the sensory periphery and projects this information to the cerebral cortex (Sherman and Guillery, 1998). Ventral MGB also receives modulatory input from cerebral cortex (Sherman and Guillery, 1998). In contrast, the other two MGB divisions, the dorsal (dMGB) and medial MGB (mMGB), do not show major projections to primary auditory cortices (Vasquez-Lopez et al., 2017, Anderson et al., 2007a, Mothe et al., 2006), and are not considered to be part of the first order (i.e., lemniscal) auditory pathway (Anderson et al., 2007b, Anderson et al., 2009, Calford, 1983, Cruikshank et al., 2001, Gonzalez-Lima and Cada, 1994, Hackett et al., 1998, Morest, 1964, Winer et al., 1999); although see (Anderson and Linden, 2011).

83 The goal of the present study was to test whether the behaviorally relevant task-dependent modulation
84 for speech is located in the first-order auditory thalamus; i.e., the vMGB (von Kriegstein et al., 2008). A
85 localization of the behaviorally relevant task-dependent modulation for speech to the vMGB would
86 provide a crucial step forward in understanding sensory thalamus function for human cognition in vivo,
87 as it would imply that the stimulus representation in the auditory sensory pathway is modulated when
88 humans recognize speech.

89 Due to the relatively small size of human MGB (ca. 5×4×5 mm, Winer, 1984) and the spatial limitations of
90 the non-invasive imaging techniques used in previous studies (von Kriegstein et al., 2008; Diaz et al., 2012),
91 it was so far not possible to differentiate in which of the three major MGB divisions there is behaviorally
92 relevant task-dependent modulation for speech. Here we therefore used ultra-high field functional
93 magnetic resonance imaging (fMRI) at 7 Tesla, enabling high spatial resolution measurements (Duyn,
94 2012). The vMGB has a strong tonotopic organization (Calford, 1983, Rodrigues-Dagaeff et al., 1989,
95 Anderson et al., 2007b) while the other two MGB subsections have only a weak tonotopic organization
96 (i.e., broadly tuned neurons, Anderson and Linden, 2011, Calford, 1983, Bartlett and Wang, 2011,
97 Rodrigues-Dagaeff et al., 1989, Ohga et al., 2018). We planned to distinguish the vMGB based on its
98 tonotopic organization as well as its topographic (i.e., ventral) location.

99 We employed three fMRI paradigms – an MGB localizer, a tonotopy localizer, and the speech experiment.
100 In the MGB localizer and the tonotopy localizer (Figure 1A), participants listened to natural sounds (human
101 voices, animal cries, tool sounds) (Moerel et al., 2015). The MGB localizer served to identify the left and
102 right MGB. The tonotopic maps resulting from the tonotopy localizer were used to localize the left and
103 right vMGB. These served as regions of interest for hypothesis testing in the speech experiment. In the
104 speech experiment (Figure 1 B & C), participants listened to blocks of auditory syllables (e.g., /aba/), and
105 performed either a speech or a speaker task. In the speech task, participants reported via button press
106 whether the current syllable was different from the previous one (1-back task). In the speaker task,
107 participants reported via button press whether the current speaker was different from the previous one.

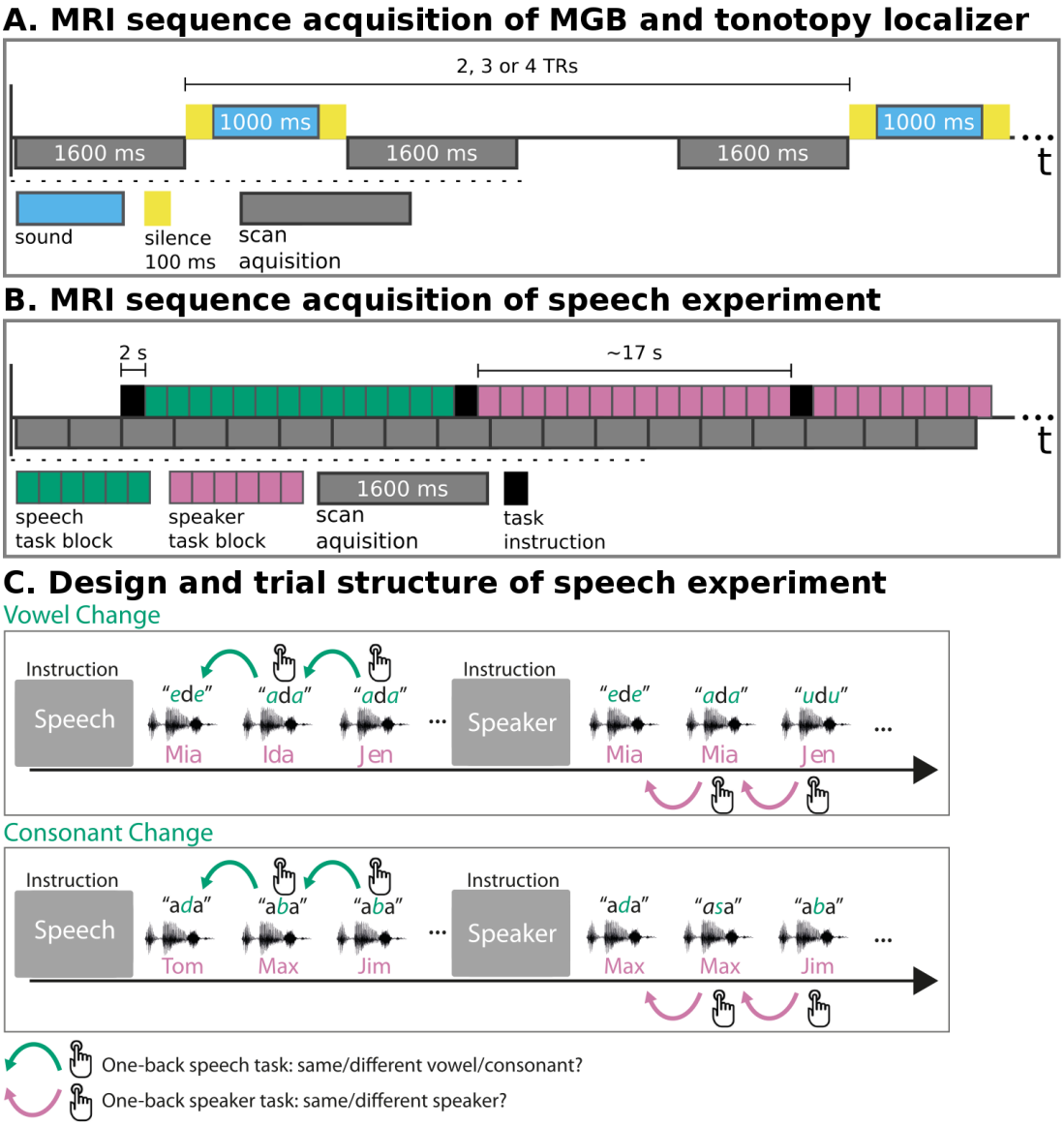

Figure 1. A. MRI sequence acquisition of MGB and tonotopy localizer. Stimuli ('sound') were presented in silence periods between scan acquisitions and jittered with 2, 3, or 4 TRs. TR: repetition time of volume acquisition. **B.** MRI sequence acquisition of the speech experiment. Each green or magenta rectangle of a block symbolizes a syllable presentation. Blocks had an average length of 17 s. Task instructions ('speech', 'speaker') were presented for 2 s before each block. MRI data were acquired continuously ('scan acquisition') with a TR of 1600 ms. **C.** Design and trial structure of speech 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 exactly the same stimulus material and trial structure. The task was, however, to press a button when there was a change in speaker identity in contrast to the immediately preceding one, independent of syllable change. Syllables differed either in vowels or in consonants within one block of trials. An initial task instruction screen informed participants about which task to perform.

121 In previous studies we found the task-dependent modulation for speech (i.e., higher response in the
122 speech in contrast to a control task on the same stimulus material) in both the left and right MGB and a
123 correlation of the task-dependent modulation with speech recognition performance only in the left MGB
124 (Kriegstein et al., 2008, Díaz et al., 2012). Since our aim of the present paper was to test whether the
125 behaviourally-task dependent modulation of MGB is present in the vMGB, we tested two hypotheses. We
126 hypothesized (i) a higher response to the speech than to the control (speaker) task in the tonotopically
127 organized left and right vMGB, and (ii) a positive correlation between speech recognition performance
128 and the task-dependent modulation for speech in the tonotopically organized left vMGB. Within our
129 design these hypotheses could be addressed by (i) the main effect of task (speech task vs speaker task) in
130 bilateral vMGB and by (ii) a correlation between the contrast speech task vs speaker task with speech
131 recognition performance across participants in left vMGB.

132 Results

133 Tonotopy localizer – replication of tonotopy in MGB

134 First, we replicated the MGB tonotopy reported previously by Moerel et al. (2015) with a larger participant
135 sample. Participants listened to natural sounds (human voices, animal cries, tool sounds) in a fast event-
136 related scheme during silent gaps of the clustered imaging technique (Moerel et al., 2015) (Figure 1A).
137 Using a model that mimics peripheral sound processing (Chi et al., 2005), each sound was represented as
138 a spectrogram. The resulting spectrograms were averaged over time and divided into ten equal
139 bandwidths in octaves. Onsets for each bin were convolved with the hemodynamic response function and
140 entered into the general linear model. Each voxel within each participant's left and right MGB localizer
141 mask was labeled according to the frequency bin to which it responded strongest, i.e., which had the
142 highest parameter estimate (Moerel et al., 2015). Thus, voxels would have values from 1-10 corresponding
143 to the frequency bin that they best represented. This resulted in a map of frequency distributions from
144 low to high frequencies in the left and right MGB for each participant.

145 Similar as in Moerel et al. (2015), we found two tonotopic gradients within the MGB in the group analysis.
146 On visual inspection, one high frequency region in the middle of the MGB was flanked by gradually lower
147 frequency components dorsally and ventrally (Figure 2A and B). The two gradients run dorso-lateral to
148 ventro-medial.

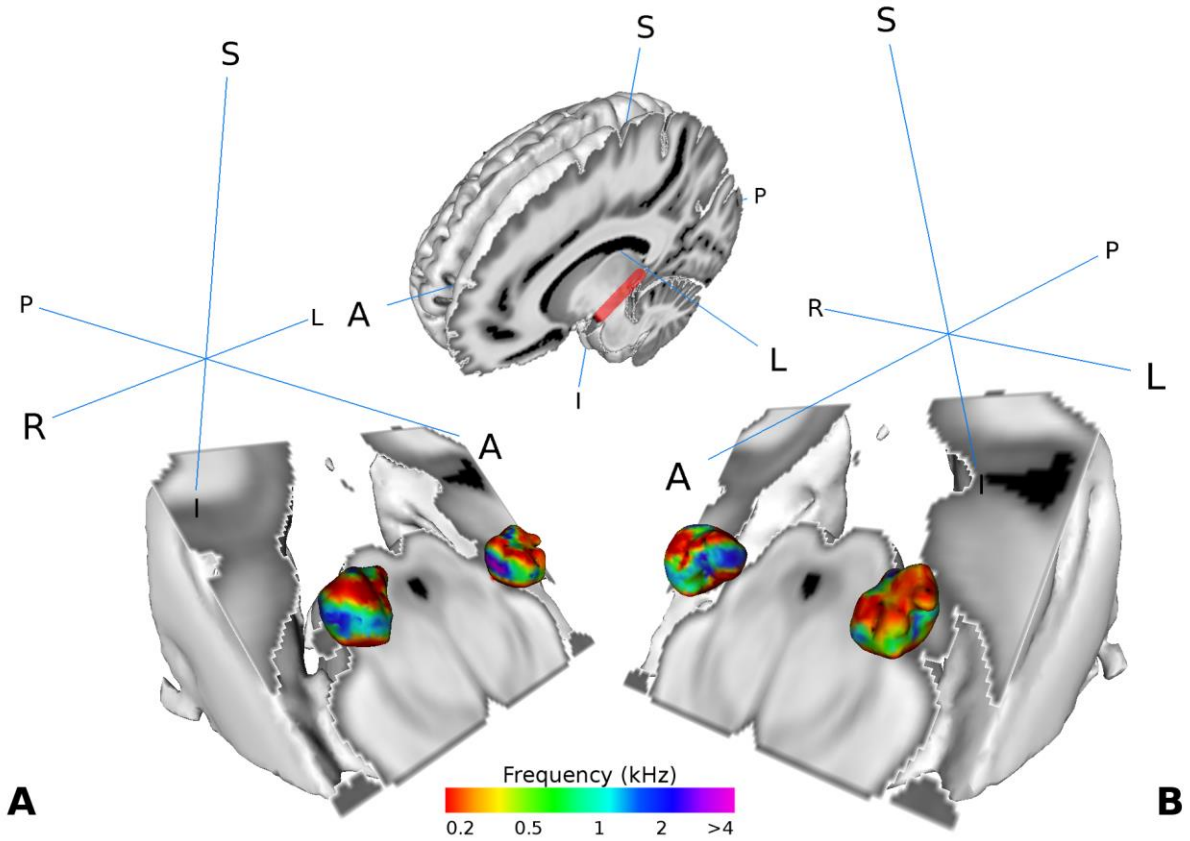

Figure 2. Visualization of the average tonotopy across participants (n = 28) found in the MGB using the tonotopic localizer. The half-brain image at the top shows the cut through the brain with a red line denoting the -45° oblique plane used in the visualizations in panels A-B. **A.** Three dimensional representation of the tonotopy in the left and right MGB with two low-high frequency gradients. **B.** Same as in A with a different orientation. Crosshairs denote orientation (A: anterior, P: posterior, L: left, R: right, S: superior, I: inferior).

The regions of low and high frequency preference could be observed in the sagittal view. To quantify the tonotopic gradient direction, we calculated gradient angles in ten slices of the left and right tonotopic map in sagittal orientation. Histograms of gradient angles in 5° steps were calculated for each slice. The histograms of the gradients were then averaged first over slices per participant, followed by an average over participants. The analysis of the mean gradient distributions across individuals (Figure 3, black line with standard error of the mean in grey) for the left MGB had maxima at 130° and 300° (dashed red lines, Figure 3). In the right MGB the mean across individual distributions had maxima at 130° and 310°.

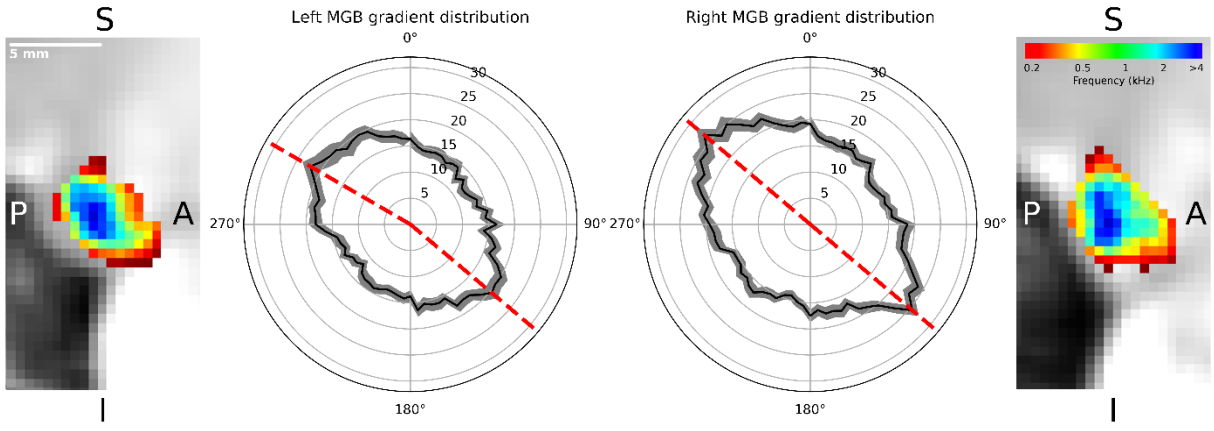

Figure 3. Distribution of gradients in a sagittal plane for ten slices averaged over participants (n = 28). The mean number of angle counts in 5° steps (black line with standard error of the mean in grey, numbers indicate counts) for the left MGB have maxima at 130° and 300° (red dashed lines). For the right MGB the maximum gradients are at 130° and 310° (red dashed lines). We interpreted these as two gradients in each MGB: one from anterior-ventral to the center (130°) and the other from the center to anterior-dorsal-lateral (300°, 310°). The two outer images display a slice of the mean tonotopic map in the left and right MGB in sagittal view (S: superior, I: inferior, P: posterior, A: anterior).

For completeness we ran the same tonotopy analysis also on the inferior colliculi (IC). This analysis (n=28) revealed a single gradient in the IC similarly to previous reports in the macaque (n=3) (Baumann et al., 2011) and human (n=6; n=5) (Moerel et al., 2015, De Martino et al., 2013) (Figure 3-figure supplement 1).

Tonotopy localizer - Localisation of vMGB

We used the high frequency components in the middle of the MGB as a reference to subdivide the MGB volume into two regions per hemisphere (Figure 4 A and B). For the left MGB, gradient 1 was located ventrally and slightly medial compared to gradient 2, which was situated more anterior, dorsal, and lateral. For the right MGB we found similar locations: gradient 1 was more ventral and medial compared to gradient 2. The center of mass (COM) and the volume for each region is summarized in Table 1. Based on the tonotopy and its ventral location (Morel et al., 1997, Bartlett and Wang, 2011) we considered gradient 1 to represent the vMGB (Moerel et al., 2015).

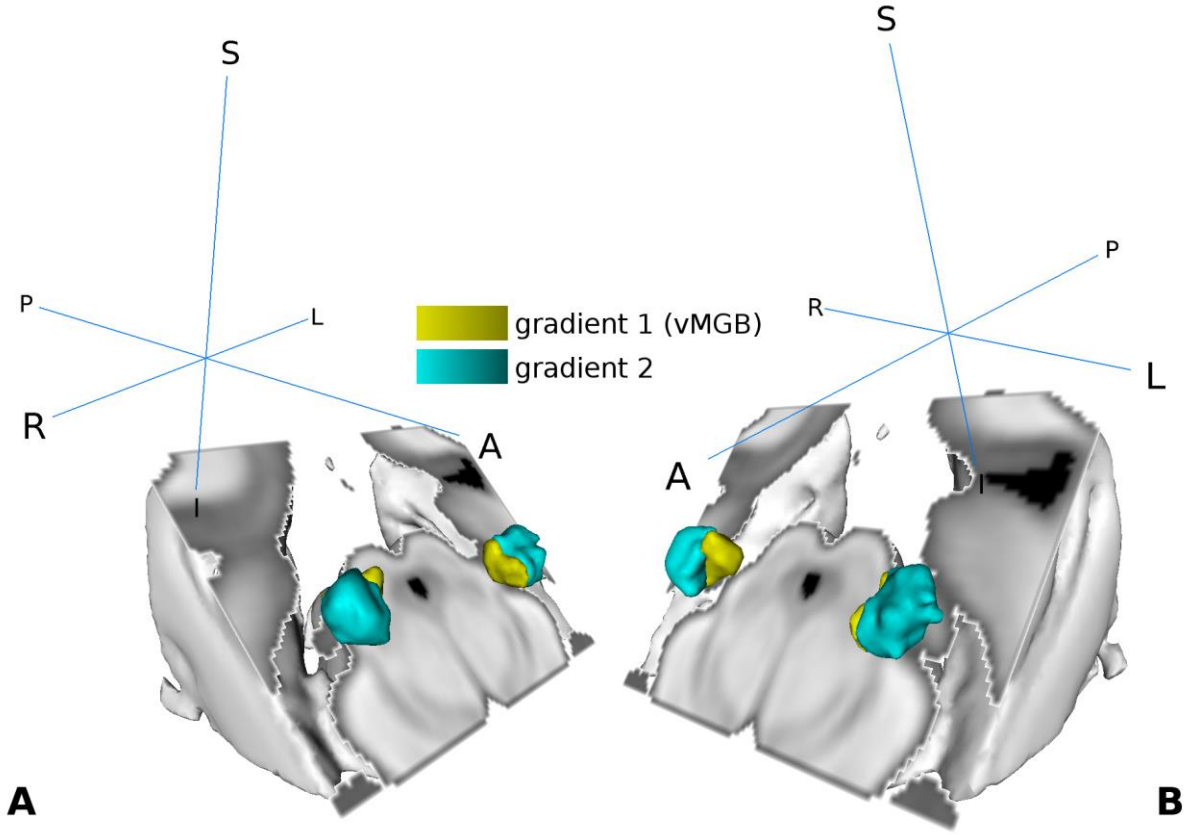

Figure 4. Visualization of the tonotopic gradients found in the MGB based on the tonotopic localizer (see Figure 2). **A.** Three dimensional rendering of the two tonotopic gradients (yellow: ventro-medial gradient 1, interpreted as vMGB, cyan: dorso-lateral gradient 2) in the left and right MGB. **B.** Same as in A with a different orientation. Orientation is the same as in Figure 2; crosshairs denote orientation.

Table 1. Center of mass (COM) and volume of each MGB mask used in the analysis.

Mask	COM (MNI coordinates mm)	Volume (mm^3)
Left Gradient 1 (ventro-medial)	(-12.7, -26.9, -6.3)	37.38
Left Gradient 2 (dorso-lateral)	(-14.8, -25.9, -5.4)	77.38
Right Gradient 1 (ventro-medial)	(12.7, -27.6, -4.4)	45.00
Right Gradient 2 (dorso-lateral)	(14.7, -25.8, -4.3)	67.38

Speech experiment

Behavioral results

Participants scored a mean hit rate in the speech task of 0.872 with 97% highest posterior density (HPD) interval [0.828, 0.915], and a mean hit rate in the speaker task of 0.760 with a 97% HPD interval [0.706, 0.810] (Figure 5; Figure 5-figure supplement 1). The mean hit-rate was 0.112 higher in the speech task than in the speaker task with 97% HPD interval [0.760, 0.150].

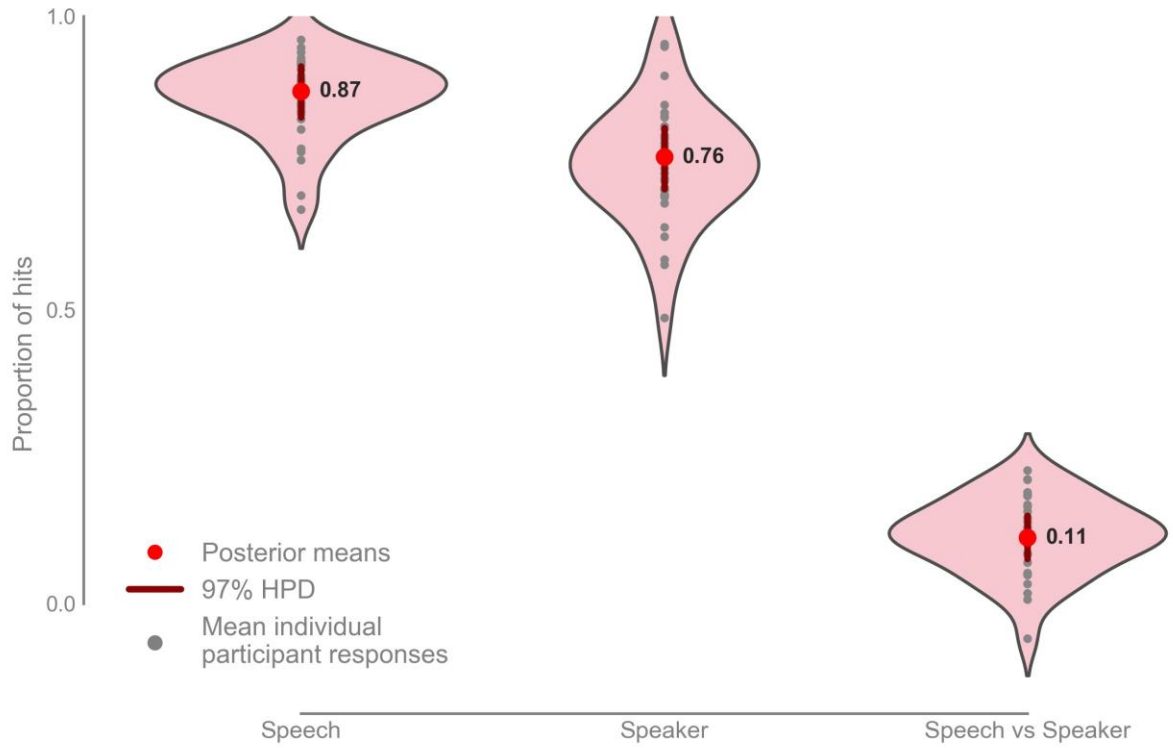

Figure 5. Mean proportion of correct button presses for the speech and speaker task behavioral scores, as well as the difference between the speech and speaker task (n = 33). Mean speech task: 0.872 with 97% HPD [0.828, 0.915], mean speaker task: 0.760 with 97% HPD [0.706, 0.800], mean speech vs speaker task: 0.112 with 97% HPD [0.760, 0.150]. Raw data provided in the Source Data File.

An analysis of reaction times showed that there was no speed-accuracy trade-off for the two tasks (Figure 5-figure supplement 2).

fMRI results

Using the fMRI data of the speech experiment we tested our two hypotheses. The first hypothesis was that within the ventral tonotopic gradient (i.e., vMGB) there is a task-dependent modulation (i.e., higher

responses for the speech than the speaker task). Unexpectedly, there was no evidence for higher BOLD response in the speech task in comparison to the speaker task (speech vs speaker contrast) in vMGB.

Our second hypothesis was that there is a positive correlation between speech recognition performance and the amount of task-dependent modulation for speech (i.e., speech vs speaker contrast) in the left vMGB across participants. As expected, there was a significant correlation between the speech vs speaker contrast and mean percent correct speech recognition scores across participants in the left vMGB [MNI coordinate: $(-11, -28, -5)$; SVC for vMGB $p = 0.04$ FWE, $T = 2.97, r = 0.46$ using Fisher's T to r transform (Fisher, 1915); parameter estimate (β) and 90% CI 0.82 [0.36, 1.27]; Figure 6 and 7]. The correlation coefficient r=0.46 (R^2=0.21) is considered to represent a large effect (Cohen, 1988 p´. 80 p´. 80), explaining 21% of the variance of either variable when linearly associated with the variance in the other.

To check whether potential outliers were driving the correlation, we excluded those data points that were two standard deviations away from the parameter estimate mean. One data point was outside this threshold. The re-calculated correlation was very similar to the one with all data points ($T = 2.92$, $p = 0.038$, $r = 0.46$), indicating that the correlation was robust to outlier removal.

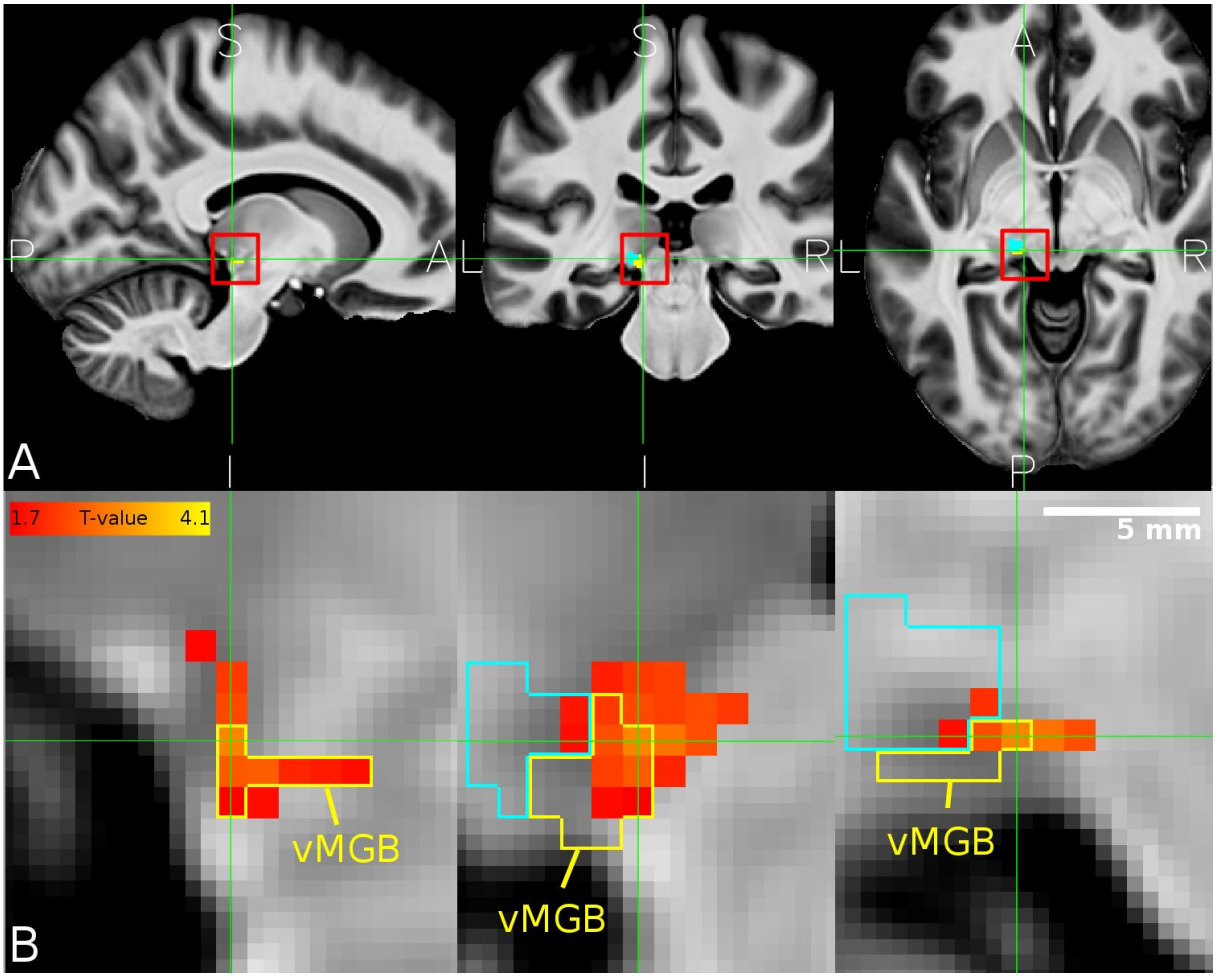

Figure 6. Overlap between MGB divisions and the behaviourally relevant task-dependent modulation. A. The mean structural image across participants (n = 33) in MNI space. The red squares denote the approximate location of the left MGB and encompass the zoomed in view in B. **B.** Overlap of correlation between the speech vs speaker contrast and the mean percent correct in the speech task (hot color code) across participants within the left vMGB (yellow). The tonotopic gradient 2 is shown in cyan. Panels correspond to sagittal, coronal, and axial slices (P: posterior, A: anterior, S: superior, I: inferior, L: left, R: right). Crosshairs point to the significant voxel using SVC in the vMGB mask (MNI coordinate -11, -28, -5).

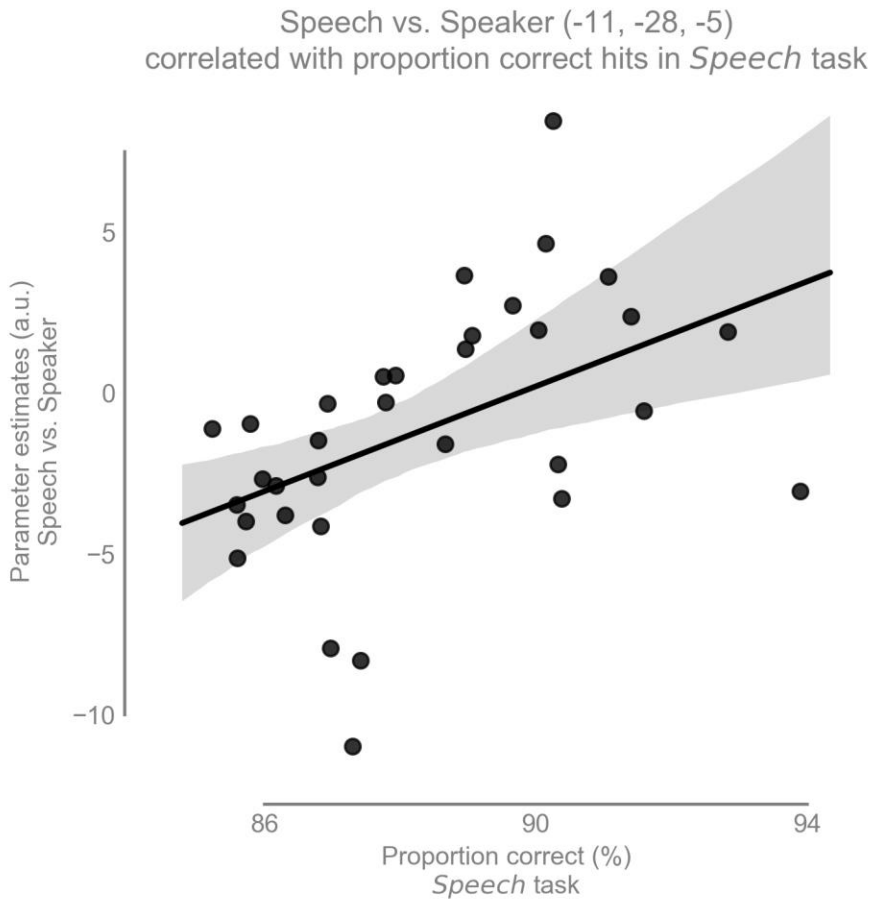

Figure 7. Task dependent modulation of left vMGB correlates with proportion correct responses in the speech task over participants (n = 33): the better the behavioral score in the speech task, the stronger the BOLD response difference between speech and speaker task in the left vMGB (maximum statistic at MNI coordinate [-11, -28, -5]. The line represents the best fit with 97% bootstrapped confidence interval (gray shaded region).

Meta-analysis of the main effect of task (speech vs speaker contrast)

We performed a random effects meta-analysis to test whether the (non-significant) effect of the main effect of task in the present study (i.e., speech vs speaker task contrast) was different from other studies that have reported a significant task-dependent MGB modulation for speech. We included five studies in the meta-analysis that each contained a speech task vs control task contrast: two experiments from (von Kriegstein et al., 2008), the data from the control participants of (Díaz et al., 2012), the result of a recent study (Mihai et al., 2019), and the result of the current study. The meta-analysis yielded an overall large effect size of d=0.85 [0.06, 1.65], p=0.036 (Figure 8; see Figure 8-figure supplement 1 for a meta-analysis including only the four published studies). The analysis showed that four of the five experiments had a

positive medium to large effect and only the current study had a very small insignificant negative effect. The confidence intervals from the current study also do not overlap with those of the other studies. In terms of equivalence testing (Schuirmann, 1987), this means that the result of the speech task vs control task contrast of the current study is different from the other studies compared here. We detail potential reasons for this difference in the task-dependent modulation between the studies in the discussion.

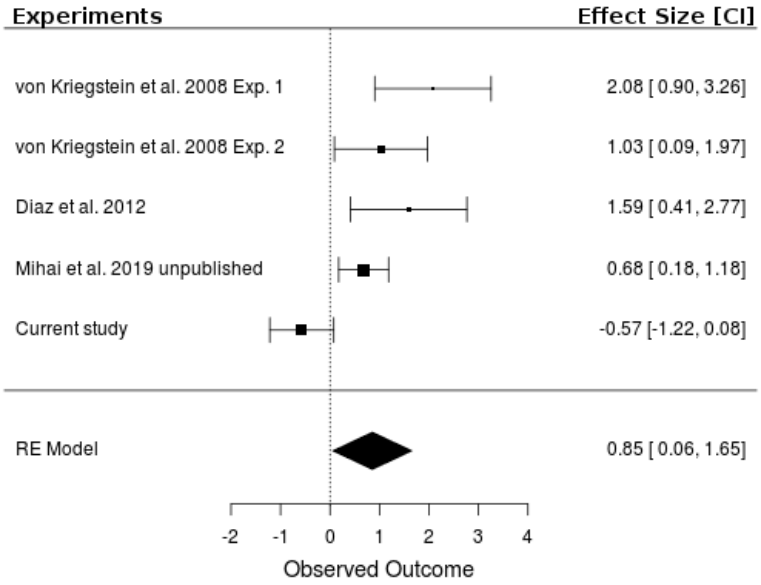

Figure 8. Meta-analysis of five experiments that investigated the difference between a speech and a non-speech control task. Experiment 1 of (von Kriegstein et al., 2008) tested a speech task vs loudness task contrast (n=16). All other experiments included a speech task vs speaker task contrast (i.e., Experiment 2 of (von Kriegstein et al., 2008) (n=17), the control participants of (Díaz et al., 2012) (n=14), the result of a recent experiment (Mihai et al., 2019) (n=17) as well as the current study (n=33)). The meta-analysis yielded an overall large effect size of d=0.85 [0.06, 1.65], p=0.036. 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.

Exploratory analyses

Specificity of the behaviorally relevant task-dependent modulation for speech. In exploratory control analyses we checked whether we could test for a specificity of the correlation between the task-dependent modulation for speech (i.e., the speech task vs speaker task contrast) and the speech recognition behavior across participants. This was, however, not possible. Performance in the speech and the speaker tasks was significantly correlated (r=0.77, p<0.001). Accordingly, there was no difference in the correlation between the contrast speech vs speaker task and the speech recognition scores and the correlation of the contrast speech vs speaker with the speaker recognition score (z=-0.717, p=0.474). The results indicated that although there is a behaviorally relevant task-dependent modulation in the vMGB for speech, we currently do not know whether it is specific to speech recognition abilities.

Specificity of the behaviorally relevant task-dependent modulation to the vMGB. In an exploratory analysis we checked whether the correlation between the task-dependent modulation for speech and the speech recognition behavior across participants was specific to the vMGB in contrast to other MGB subsection. There was no correlation (speech vs speaker task correlated with speech recognition performance across participants) in the left MGB – gradient 2 (Supplementary File 1). The comparison of the two correlations (i.e., for left vMGB and left MGB-gradient 2) showed that they are significantly different ($z=4.466$, $p<0.0001$). Thus this is a first indication that the behaviorally relevant task-dependent modulation for speech is specific to the left vMGB.

Exploration of other subcortical areas. In exploratory analyses we tested the main effect of task (speech vs speaker contrast) and the correlation of the speech vs speaker contrast with behavioral performance in the speech task in several areas outside the vMGB, for which we did not have an a priori hypothesis. There were no significant effects in any other regions (Supplementary File 1).

Discussion

Using ultra-high field fMRI we showed that it is the left auditory first-order sensory thalamus – the left ventral subdivision of the MGB (vMGB) – that is modulated during speech recognition. The vMGB is the primary sensory pathway nucleus of the auditory thalamus and transmits input to the cerebral cortex (Winer, 1984, Anderson et al., 2007b, Bartlett et al., 2011, Bordi and LeDoux, 1994, Calford, 1983, Moerel et al., 2015). The present results imply that, when decoding speech, higher order cortical areas modify representations of the sensory input in the primary sensory thalamus and that such modification is relevant for speech recognition abilities. These results are a further indication that speech recognition might only be fully understood if dynamic cortico-thalamic interactions are taken into account (Klostermann, 2013, Kriegstein et al., 2008). The results are based on the test of two equally weighted hypotheses: a main effect and a correlation. Although we found no significant main effect in the vMGB (nor in any other subregion of the MGB), the correlation between the task-dependent modulation and behavioral performance was, as hypothesized, significant within the left vMGB.

We localized the vMGB based on its tonotopic organization and location relative to other MGB divisions. The tonotopic organization of the vMGB has been observed in many species with the use of invasive techniques (Winer, 1984, Anderson et al., 2007b, Bartlett et al., 2011, Bordi and LeDoux, 1994, Calford, 1983) and non-invasively in six human participants using ultra-high field fMRI (Moerel et al., 2015). Similar to Moerel et al. (2015) we here also identified two tonotopic gradients. Moerel et al. (2015) attributed

the ventral gradient to the ventral MGB, and the other gradient cautiously to the tonotopically organized lateral posterior thalamic nucleus (Pol), which is part of the non-lemniscal system (Jones, 1985). The Pol is also tonotopically organized with sharp tuning curves similar to the vMGB (Imig and Morel, 1985). Gradient 2 in our study is, however, larger than gradient 1. Thus, gradient 2 might also represent a composite of several nuclei that are in close proximity to the MGB (Bartlett and Wang, 2011) such as the Pol and potentially the suprageniculate, which has a preference for high frequencies (Bordi and LeDoux, 1994) (for a detailed thalamic atlas see Morel et al., 1997). Furthermore, the weak tonotopy of the dMGB or mMGB might also contribute to gradient 2. Another interpretation of the two tonotopic gradients is that the vMGB in humans might include two tonotopic maps, i.e., that frequency gradient 1 and 2 are both part of the vMGB. Two tonotopic gradients have been found in the rat vMGB (Shiramatsu et al., 2016), but not consistently in other species (Hackett et al., 2011, Horie et al., 2013, Tsukano et al., 2017). The volume of the two gradients, however, speaks against the possibility of two tonotopic maps in human vMGB. That is, the two gradients make up already ca. 100 mm^3 and reported whole MGB volumes based on characterization in post-mortem human brains are between ca. 40-120 mm^3 (Rademacher et al., 2002, Moro et al., 2015). Thus while gradient 1 can be clearly attributed to the vMGB, due to its tonotopic gradient and ventral location, the nature of the second frequency gradient remains an open question.

Based on previous findings (von Kriegstein et al., 2008; Diaz et al., 2012), we expected significant responses for the speech vs speaker task contrast in vMGB. The lack of a significant main effect of task (speech vs speaker) in the vMGB was surprising, and the meta-analyses showed that the null-finding was indeed different from categorical task-effects (speech vs loudness tasks and speech vs speaker tasks) observed in other experiments in participants with typical development (von Kriegstein et al., 2008, Díaz et al., 2012, Mihai et al., 2019).

There are three potential explanations for the difference in the results for the main effect of task between the present and previous studies. First, the speaker task was more difficult to perform (indicated by the lower behavioral score during the speaker vs speech task and subjective reports of the participants), which may have led to higher BOLD responses for the more difficult task. However, this explanation is unlikely as previous studies with matched performance across tasks (Díaz et al., 2012, von Kriegstein et al., 2008 experiment 2) and also studies where the control task was more difficult than the speech task (von Kriegstein et al., 2008 experiment 1) have found a task-dependent MGB modulation for the speech task.

Second, we employed a liberal threshold in choosing participants based on their reading speed and comprehension scores (lower fourth of the mean and higher, i.e., 26%-100%). Participants who scored

lower on this test might also show a lower task-dependent modulation of the MGB (see Díaz et al., 2018). However, we find this explanation unlikely as those participants with lower reading score showed a broad (low to high) BOLD-response spectrum (Figure 7-figure supplement 1), and in the previous study (Díaz et al., 2018) a correlation between the MGB task-dependent modulation and reading speed and comprehension scores has been found only in participants with developmental dyslexia, but not in typically developed controls.

A third explanation is that we used unmanipulated natural voices from different speakers. In the previous studies different speaker voices were synthesized from one original voice to differ only in two key voice-identity parameters, i.e., the acoustic effect of the vocal tract length and the fundamental frequency (f0) (von Kriegstein et al., 2008, Díaz et al., 2012, Gaudrain et al., 2009). Vocal tract length and f0 are relatively stable acoustic cues that do not vary greatly over time in contrast to the highly dynamic cues (e.g., formant transitions, voice onset times, stops, Kent et al., 1992) that are most important for signaling phonemes and are used for speech recognition. However, dynamic cues, such as pitch periodicity, segmental timings, and prosody can also be used for speaker identification (Benesty et al., 2007). In the present experiment, which included natural voices, participants might have also used fast changing cues for speaker identity recognition, particularly because the task was difficult. Since dynamic cues are essential for speech recognition, using dynamic cues in a speaker task would render the two tasks less different. Thus, MGB modulation might also have played a role in performing the speaker task.

The localization of the correlation between the speech vs speaker contrast and performance in the speech task to the vMGB confirmed our hypothesis that it is the first-order thalamic nucleus – vMGB – that is involved in speech recognition. The correlation indicated that those participants on the lower side of the task-dependent modulation spectrum, as given by the speech vs speaker contrast, have lower proportion of hits in the speech task and those participants on the higher side of the task-dependent modulation spectrum, show higher proportion of hits in the speech task and are thus better at speech recognition. The variability of proportion of hits across participants was between 85.2% and 93.9% which translates to an odds ratio of 2.66 (i.e., those participants with higher task-dependent modulation are ~2.5 times more likely to have more hits compared to misses in the speech task). This variability is meaningful for a speech recognition task; for example, it has been shown that the hearing impaired perform 5-10% lower on speech recognition compared to normal hearing participants (Panouillères and Möttönen, 2018, Gordon-Salant and Fitzgibbons, 1993).

In the present study, the results cannot be explained by differences in stimulus input in the two conditions, as the same stimuli were heard in both tasks. We, however, cannot exclude with the present data set that, in general, participants who are better task-performers have a higher task-dependent modulation of the vMGB for speech. It is therefore still an open question whether the task-dependent modulation is specific to speech recognition behavior. A previous study in the visual modality, gave first evidence that the task-dependent modulation of the visual thalamus (LGN) for visual speech is specific for predictable dynamic information such as visual speech in contrast to unpredictable dynamic information (Díaz et al., 2018). However a similar study in the auditory modality is so far missing.

What kind of mechanism could be represented by the correlation between task-dependent modulation of the vMGB and speech recognition performance? Experimental and theoretical accounts of brain function emphasize the importance of an anatomical cortical and subcortical hierarchy that is organized according to the timescale of complex stimuli in the natural environment (Davis and Johnsrude, 2007, Giraud et al., 2000, Kiebel et al., 2008, Wang et al., 2008). In brief, it is assumed that levels closer to the sensory input encode faster dynamics of the stimulus than levels further away from the sensory input. In accordance with this view, the MGB (as well as the visual first-order thalamus [LGN]; Hicks et al., 1983) is tuned to high frequencies of temporal modulation (ca. 16 Hz in human MGB;Giraud et al., 2000) in relation to their associated primary sensory cortical areas (Giraud et al., 2000, Wang et al., 2008, Foster et al., 1985). For humans, the optimized encoding of relatively fast dynamics; e.g., at the phoneme level, is critical for speech recognition and communication (Shannon et al., 1995a, Tallal et al., 1996, Tallal and Piercy, 1975). Many important speech components like formant transitions, voice onset times, or stops are on very fast time scales of 100 ms or less (Hayward, 2000). Additionally, the sound envelope described by relatively fast temporal modulations (1-10 Hz in quiet environments, 10-50 Hz in noisy environments) is important for speech recognition (Elliott and Theunissen, 2009, Shannon et al., 1995b). The Bayesian brain hypothesis proposes that the brain uses internal dynamic models of its environment to predict the trajectory of the sensory input hypothesis (Knill and Pouget, 2004, Friston and Kiebel, 2009, Friston, 2005, Kiebel et al., 2008). In accordance with this hypothesis, we have previously suggested that slower dynamics encoded by auditory cortical areas (Giraud et al., 2000, Wang et al., 2008) provide predictions about input arriving at lower levels of the temporal-anatomic hierarchy (Kiebel et al., 2008, von Kriegstein et al., 2008). In this view, these dynamic predictions modulate the response properties of the first-order sensory thalamus to optimize the early stages of speech recognition. For example, in non-human animals cortico-thalamic projections outnumber thalamo-cortical projections (reviewed in reviewed in Ojima and Rouiller, 2011), and alter the response properties of thalamic neurons (Andolina et al., 2007, Cudeiro and

394 Sillito, 2006, Ergenzinger et al., 1998, Ghazanfar and Nicolelis, 2001, Sillito et al., 2006 [review] , Wang et
395 al., 2018a, Sillito et al., 1994, Krupa et al., 1999). In speech processing such a mechanism might be
396 especially useful as the signal includes both rapid dynamics, and is predictable (e.g., due to co-articulation
397 or learned statistical regularities in words Saffran, 2003). We suggest that higher-level regularities yield
398 the predictions for lower-level details of the stimulus. While, for example, words are represented at the
399 level of the cerebral cortex (Huth et al., 2016, Davis and Johnsrude, 2003, Price et al., 2005), the
400 predictions about the most likely components of the word are percolating down the hierarchy. In this
401 view, sensory thalamus structures do not represent word level information but they are tuned by the
402 cerebral cortex areas to expect the detailed trajectory of the speech signal that is most likely given the
403 expectations generated at the cerebral cortex level. Speech needs to be computed online often under
404 suboptimal listening conditions. Building up accurate predictions within an internal generative model
405 about fast sensory dynamics would result in more efficient processing when the perceptual system is
406 confronted with taxing conditions such as fast stimulus presentation rates or background noise. We
407 speculate that the correlation between speech task performance and task-dependent vMGB modulation
408 might be a result of feedback from cerebral cortex areas. Feedback may emanate directly from auditory
409 primary or association cortices, or indirectly via other structures such as the reticular nucleus with its
410 inhibitory connections to the MGB (Rouiller and de Ribaupierre, 1985). Feedback cortico-thalamic
411 projections from layer 6 in A1 to the vMGB, but also from association cortices such as the motion sensitive
412 part of the planum temporale (Tschentscher et al., 2019), may modulate information ascending through
413 the lemniscal pathway, rather than convey information to the ventral division (Lee, 2013, Llano and
414 Sherman, 2001).

415 Although most of speech and language research focuses on cerebral cortex structures, investigating
416 subcortical sensory contributions to speech perception is paramount to the development of a mechanistic
417 understanding of how the human brain accomplishes speech recognition. The present study brings us a
418 decisive step further in this direction by suggesting that it is the *primary* sensory auditory thalamus that
419 shows a behaviorally relevant task-dependent modulation for speech recognition.

Materials and Methods

Participants

The Ethics committee of the Medical Faculty, University of Leipzig, Germany approved the study (protocol number 273/14-ff). We recruited 33 participants using 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), native German-speakers, had a mean age and standard deviation (SD) of 24.9 ± 2.5 years, and included 23 females. 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) with a threshold equal to and below 25 dB (Madsen Micromate 304, GN Otometrics, Denmark). To exclude possible undiagnosed 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 were 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 and standard deviation: 69.9% ± 19.5%, lowest mean score: 36%). 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 and standard deviation: 16.2 ± 4.8; cutoff: 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) and has overlapping symptoms with dyslexia (White et al., 2006). Participants received monetary compensation for participating in the study.

Experiments

We performed three different functional MRI measurements: the speech experiment (n=33), a MGB localizer (n=33), and a tonotopy localizer (n = 28, 18 females, age 24.8 ± 5.0 years). Each experiment was performed once.

Stimuli

MGB and Tonotopy localizer. The stimuli for the MGB localizer and the tonotopy localizer consisted of 84 and 56 natural sounds, respectively, sampled at 16 kHz at 32 bit, and included samples of human speech, animal cries and tool sounds (these were the same as described in Moerel et al., 2015). The stimuli had a duration of 1000 ms, were ramped with 10 ms linear slopes, and had equalized root-mean-square levels.

450 *Speech experiment*. The speech experiment stimuli consisted of 448 vowel-consonant-vowel (VCV)
451 syllables with an average duration and SD of 803 ± 105 ms. These were spoken by three female and three
452 male speakers (mean age and SD 27.7 ± 3.3 years) unfamiliar to the participants, and were recorded with
453 a video camera (Canon Legria HFS10, Canon, Japan) and a Røde NTG-1 Microphone (Røde Microphones,
454 Silverwater, NSW, Australia) connected to a pre-amplifier (TubeMP Project Series, Applied Research and
455 Technology, Rochester, NY, USA) in a sound-attenuated room. The sampling rate was 48 kHz at 16 bit.
456 Auditory stimuli were cut and flanked by Hamming windows of 15 ms at the beginning and end, converted
457 to mono, and root-mean-square equalized using Python 3.6 (Python Software Foundation,
458 www.python.org).

459 **Procedure**

460 *MGB and Tonotopy localizer*. For the MGB localizer and the tonotopy localizer, participants listened to
461 natural sounds (human voices, animal cries, tool sounds; Moerel et al., 2015). The MGB localizer consisted
462 of one run where 84 natural sound stimuli were presented in random order and had a duration of 12:50
463 minutes. The tonotopy localizer consisted of six runs where 56 of the 84 natural sound stimuli from the
464 MGB localizer were presented. The sounds were randomly chosen before the first run and the same 56
465 sounds were played in each run. Each run had a duration of 8:58 minutes. To ensure listener engagement,
466 in both localizers the participants performed a 1-back task and pushed a button when two consecutive
467 sounds were the same. This happened on average 5% of the time. Additionally, 5% of the trials contained
468 no sound (null events). Within each run, sounds were randomly jittered at an interval of 2, 3, or 4
469 repetition times (TR) and presented in the middle of the silent gap of 1200 ms (Figure 1A). The MGB
470 localizer was used as an independent functional identifier for the left and right MGB. The resulting masks
471 were then used to constrain the analyses of the tonotopy localizer to these regions of interest. In turn,
472 the tonotopic regions of the MGB were used as masks in the speech experiment (see section Functional
473 MRI Data Analysis).

474 *Speech experiment*. In the speech experiment (Figure 1C) participants listened to blocks of auditory VCV
475 syllables, and were asked to perform two types of tasks: a speech task and a speaker task. In the speech
476 task, participants reported via button press whether the current syllable was different from the previous
477 one (1-back task). In the speaker task, participants reported via button press whether the current speaker
478 was different from the previous one. Speakers within a block were either all male or all female. This was
479 necessary to avoid that participants performed a gender discrimination task on some trials and a speaker
480 identity task on other trials. Task instructions were presented for two seconds prior to 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 1B). Each block contained 14 stimuli. Each stimulus presentation was followed by 400 ms of silence. Within one block both syllables and speakers changed six or seven times. The average length of a block and SD was 17.0 ± 0.9 seconds. Counterbalancing of the stimulus material for the two tasks was achieved by presenting each block twice: once with the instruction to perform the speech task and once with the instruction to perform the speaker task. Besides the factor “task”, the experiment included another factor. That is, blocks had either only vowel or only consonant changes. While this factor is included in the analysis, it is irrelevant for addressing the current research question.

The experiment was divided into five runs with a duration of 8:30 minutes per run. Each of the four condition blocks (speech vowel change, speaker vowel change, speech consonant change, speaker consonant change) were presented six times in pseudo-randomized order. The last stimulus presentation in the run was followed by 30 s of no stimulation. Participants were allowed to rest for one minute between runs. To familiarize participants with speakers’ voices and to ensure they understood the task, they performed two initial training runs outside the MRI-scanner: one run for speaker familiarization, and one for experiment familiarization (detailed below).

The experiments were programmed and presented using Presentation (v17.1, NeuroBehavioral Systems, Berkley, CA, USA) in Windows XP and delivered through an MrConfon amplifier and earbuds linked to the transducers via air tubes (manufactured 2008, MrConfon GmbH, Magdeburg, Germany).

Participant training

The participant training consisted of a speaker familiarization and an experiment familiarization.

The speaker familiarization 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 stimuli. Participants heard a pseudo-random choice of the same stimuli used in the experiment. Each block contained one speaker-identity only. Participants were alerted to the onset of a new speaker-identity block with the phrase “Andere/r Sprecher/in” (German for “Another Speaker”). Participants listened to the voices with the instruction to memorize the speaker’s voice. In the following test phase participants were presented with four blocks of nine trials that each contained syllable pairs spoken by the same or different speaker from the ones they were trained on. Participants were asked to indicate

510 whether the speakers of the two syllables were the same or different by pressing keypad buttons “1” for
511 yes and “2” for no. Participants received visual feedback for correct (green flashing German word for
512 correct: “Richtig”) and incorrect (red flashing German word for incorrect: “Falsch”) answers. If participants
513 scored below 85%, they repeated the speaker familiarization training.

514 The experiment familiarization consisted of one 8:30 min long run of the fMRI speech experiment (Figure
515 1B/C). Stimuli were randomly chosen from the same stimulus material used in the experiment. If
516 participants scored below 85% across tasks in the experiment familiarization run they repeated the
517 experiment familiarization training.

518 The training (speaker and experiment familiarization) took place within the same testing-session as the
519 pre-tests (audiometry, reading comprehension, and AQ questionnaire), and was repeated half an hour
520 prior to the fMRI experiment, which took place at a later date.

521 Data Acquisition and Processing

522 MRI data were acquired using a Siemens Magnetom 7 T scanner (Siemens Healthineers, Erlangen,
523 Germany) with a Nova 32-channel head coil (Nova Medical, Wilmington MA, USA). Functional MRI data
524 were acquired using echo planar imaging (EPI) sequences. We used a field of view (FoV) of 132 x 132 mm
525 and partial coverage with 28 slices. This volume was oriented obliquely such that the slices encompassed
526 the inferior colliculi (IC), the MGB and the superior temporal gyrus, running in parallel to the latter (Figure
527 6-figure supplement 1).

528 The MGB and tonotopy localizers had the following acquisition parameters: TR = 2800 ms (acquisition
529 time TA = 1600 ms, silent gap: 1200 ms), TE = 22 ms, flip angle 65°, GRAPPA (Griswold et al., 2002) with
530 acceleration factor 2, 33% phase oversampling, matrix size 120 x 120, FoV 132 mm x 132 mm, phase partial
531 Fourier 6/8, voxel size 1.1 mm x 1.1 mm x 1.1 mm, interleaved acquisition, anterior to posterior phase-
532 encode direction. We employed a clustered EPI technique allowing for stimulus presentations in quiet in
533 a fast-event related design. Stimuli were presented during the silent gap. For the MGB localizer we
534 acquired one run of 275 volumes (13:06 minutes). For the tonotopy localizer we acquired 187 volumes
535 (9:01 minutes) per run with a total of six runs.

536 For the speech experiment, acquisition parameters were the same as for the localizers, with the exception
537 of a shorter TR (1600 ms) due to continuous scanning (i.e., no silent gap), 320 volumes per run, and total
538 length of acquisition per run of 8:30 minutes. Five runs were recorded for each participant. The acquisition

539 parameters were similar to the protocol described by (Moerel et al., 2015), with the exception of a longer
540 echo time and phase oversampling which eschewed front-back wrapping artifacts. An example EPI is
541 shown in Figure 6-figure supplement 2. The difference in echo time between our sequence and the one
542 in Moerel et al. (2015) may have resulted in a lower signal-to-noise ratio in subcortical structures.
543 However, as the MGB has a T2* value of ~33 ms the different echo times of 19 ms (our sequence) and 22
544 ms (Moerel et al., 2015) had little to no effect on the applied general linear model (Hollander et al., 2017).

545 During functional MRI data acquisition we also acquired physiological values (heart rate, and respiration
546 rate) using a BIOPAC MP150 system (BIOPAC Systems Inc., Goleta, CA, USA). Structural images were
547 recorded using an MP2RAGE (Marques et al., 2010) T1 protocol: 700 µm isotropic resolution, TE = 2.45ms,
548 TR = 5000 ms, TI1 = 900 ms, TI2 = 2750 ms, flip angle 1 = 5°, flip angle 2 = 3°, FoV 224 mm × 224 mm,
549 GRAPPA acceleration factor 2, duration 10:57 min.

550 **Behavioral Data Analysis**

551 Button presses were modeled using a binomial logistic regression which predicts the probability of correct
552 button presses based on four independent variables (speech task, vowel change; speech task, consonant
553 change; speaker task, vowel change; speaker, task consonant change) in a Bayesian framework
554 (McElreath, 2015).

555 To pool over participants and runs we modeled the correlation between intercepts and slopes. For the
556 model implementation and data analysis, we used PyMC3 (Salvatier et al., 2016) using a No-U-Turn
557 Sampler (Hoffman and Gelman, 2011) with three parallel chains. Per chain we had 20,000 samples with
558 5,000 of these as warm-up. Only the latter 7,500 were used for posterior mean and highest posterior
559 density (HPD) interval estimates. The difference in percent correct button presses between the speech
560 and speaker task was calculated using the posterior densities averaged over consonant and vowel
561 changes. The resulting distribution was averaged and the 97% HPD was calculated. If the posterior
562 probability distribution does not strongly overlap zero, then there was a detectable difference between
563 conditions (McElreath, 2015).

564 The predictors included in the behavioral data model were: Task (1 = speech, 0 = speaker), and Syllable
565 Change (1 = vowel, 0 = consonant). We also included the two way interaction of Task and Syllable Change.
566 Because data were collected across participants and runs it is reasonable to include random effects for
567 both of these in the logistic model, albeit we expected little to no difference in performance between
568 runs. The model had the following likelihood and linear models:

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

$$\text{logit}(p_{i,j}) = A_{i,j} + B_{S,i,j} + B_{V,i,j} + B_{SV,i,j}, \text{ for } i = 1, \dots, I; j = 1, \dots, J$$

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

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

$$B_{V,i,j} = \beta_V + \beta_{V,participant[i]} + \beta_{V,\,\text{run}[j]}$$

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

$$\begin{bmatrix} \alpha_{participant} \\ \beta_{S,\text{participant}} \\ \beta_{V,\text{participant}} \\ \beta_{SV,\text{participant}} \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \alpha \\ \beta_S \\ \beta_V \\ \beta_{SV} \end{bmatrix}, \mathbf{S}_{participant} \right)$$

$$\begin{bmatrix} \alpha_{\text{run}} \\ \beta_{S,\text{run}} \\ \beta_{V,\text{run}} \\ \beta_{SV,\text{run}} \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \alpha \\ \beta_S \\ \beta_V \\ \beta_{SV} \end{bmatrix}, \mathbf{S}_{\text{run}} \right)$$

$$\mathbf{S_{participant}} = \begin{pmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta_S} & 0 & 0 \\ 0 & 0 & \sigma_{\beta_S} & 0 \\ 0 & 0 & 0 & \sigma_{\beta_S V} \end{pmatrix} \mathbf{R_{participant}} \begin{pmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta_S} & 0 & 0 \\ 0 & 0 & \sigma_{\beta_S} & 0 \\ 0 & 0 & 0 & \sigma_{\beta_S V} \end{pmatrix}$$

$$\mathbf{S_{run}} = \begin{pmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta_S} & 0 & 0 \\ 0 & 0 & \sigma_{\beta_S} & 0 \\ 0 & 0 & 0 & \sigma_{\beta_S V} \end{pmatrix} \mathbf{R_{run}} \begin{pmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta_S} & 0 & 0 \\ 0 & 0 & \sigma_{\beta_S} & 0 \\ 0 & 0 & 0 & \sigma_{\beta_S V} \end{pmatrix}$$

$$\alpha \sim \text{Normal}(0,5)$$

$$\beta_S \sim \text{Normal}(0,5)$$

$$\beta_V \sim \text{Normal}(0,5)$$

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

$$(\sigma_{\text{participant}}, \sigma_{\text{run}}) \sim \text{HalfCauchy}(2)$$

$$\sigma_{corr,participant} \sim \text{HalfCauchy}(2)$$

$$\sigma_{corr,run} \sim \text{HalfCauchy}(2)$$

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

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

where *I* is the number of subjects and *J* is the number of runs. The model is compartmentalized into sub-models for the intercepts and each slope. $A_{i,j}$ is the sub-model for the intercept for observations *i,j*. Similarly, $B_{S,i,j}$, $B_{V,i,j}$ and $B_{SV,i,j}$ are the sub-models for the speech-speaker slope, vowel-consonant slope and the interaction slope; $\boldsymbol{S_{participant}}$ and $\boldsymbol{S_{run}}$ are the covariance matrices, $\boldsymbol{R_{participant}}$ and $\boldsymbol{R_{run}}$ are the priors for the correlation matrices between the intercepts and slopes modeled as a LKJ probability density for participants and runs (Lewandowski et al., 2009). Informative priors for the intercept (α) and additional coefficients (e.g. β_S), random effects for subject and run ($\beta_{S,participant}$, $\beta_{S,run}$), and multivariate priors for subjects and runs identify the model by constraining the position of $p_{i,j}$ to reasonable values.

To check for correlations in the performance between tasks, we calculated the Pearson's product moment (Cohen, 1988 p. 75) on the proportion of hits in the speech and speaker task using Python 3.6.

We analyzed reaction times using a linear model in a Bayesian framework. Reaction times were mean centered and the priors on reaction times were modeled as T-distributions. We calculated the mean difference in reaction times between the speech and the speaker task to check for a speed-accuracy trade-off. The model is described below:

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

$$p_{i,j} = A_{i,j} + B_{S,i,j} + B_{V,i,j} + B_{SV,i,j}, \text{for } i = 1, \dots, I; j = 1, \dots, J$$

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

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

$$B_{V,i,j} = \beta_V + \beta_{V,participant[i]} + \beta_{V,\,\text{run}[j]}$$

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

$$\begin{bmatrix} \alpha_{participant} \\ \beta_{S,\text{participant}} \\ \beta_{V,\text{participant}} \\ \beta_{SV,\text{participant}} \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \alpha \\ \beta_S \\ \beta_V \\ \beta_{SV} \end{bmatrix}, \mathbf{S}_{participant} \right)$$

$$\begin{bmatrix} \alpha_{\text{run}} \\ \beta_{S,\text{run}} \\ \beta_{V,\text{run}} \\ \beta_{SV,\text{run}} \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \alpha \\ \beta_S \\ \beta_V \\ \beta_{SV} \end{bmatrix}, \mathbf{S}_{\text{run}} \right)$$

$$\mathbf{S}_{\mathbf{participant}} = \begin{pmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta_S} & 0 & 0 \\ 0 & 0 & \sigma_{\beta_S} & 0 \\ 0 & 0 & 0 & \sigma_{\beta_S V} \end{pmatrix} \mathbf{R}_{\mathbf{participant}} \begin{pmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta_S} & 0 & 0 \\ 0 & 0 & \sigma_{\beta_S} & 0 \\ 0 & 0 & 0 & \sigma_{\beta_S V} \end{pmatrix}$$

$$\mathbf{S}_{\mathbf{run}} = \begin{pmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta_S} & 0 & 0 \\ 0 & 0 & \sigma_{\beta_S} & 0 \\ 0 & 0 & 0 & \sigma_{\beta_S V} \end{pmatrix} \mathbf{R}_{\mathbf{run}} \begin{pmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta_S} & 0 & 0 \\ 0 & 0 & \sigma_{\beta_S} & 0 \\ 0 & 0 & 0 & \sigma_{\beta_S V} \end{pmatrix}$$

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

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

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

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

$$\left(\sigma_{\text{participant}}, \sigma_{\text{run}}\right) \sim \text{HalfCauchy}(2)$$

$$\sigma_{corr,participant} \sim \text{HalfCauchy}(2)$$

$$\sigma_{corr,run} \sim \text{HalfCauchy}(2)$$

$$\mathbf{R}_{\mathbf{participant}} \sim LKJcorr\left(4, \sigma_{corr,participant}\right)$$

$$\mathbf{R}_{\mathbf{run}} \sim LKJcorr\left(4, \sigma_{corr,run}\right)$$

621 where I is the number of subjects and J is the number of runs. The model is compartmentalized into sub-
622 models for the intercepts and each slope. $A_{i,j}$ is the sub-model for the intercept for observations i,j.
623 Similarly, $B_{S,i,j}$, $B_{V,i,j}$ and $B_{SV,i,j}$ are the sub-models for the speech-speaker slope, vowel-consonant slope and
624 the interaction slope; $\boldsymbol{S_{participant}}$ and $\boldsymbol{S_{run}}$ are the covariance matrices, $\boldsymbol{R_{participant}}$ and $\boldsymbol{R_{run}}$ are the
625 priors for the correlation matrices between the intercepts and slopes modeled as a LKJ probability density
626 for participants and runs (Lewandowski et al., 2009). Informative priors for the intercept (α) and additional
627 coefficients (e.g. β_S), random effects for subject and run ($\beta_{S,participant}$, $\beta_{S,run}$), and multivariate priors
628 for subjects and runs identify the model by constraining the position of $p_{i,j}$ to reasonable values.

629

630 ## Functional MRI Data Analysis

631 ### *Preprocessing of MRI data*

632 The partial coverage by the 28 slices and the lack of a whole brain EPI measurement resulted in
633 coregistration difficulties of functional and structural data. As a workaround, the origin (participant space
634 coordinate [0, 0, 0]) of all EPI and MP2RAGE images were manually set to the anterior commissure using
635 SPM 12. Furthermore, to deal with the noise surrounding the head in MP2RAGE images, these were first
636 segmented using SPM's new segment function (SPM 12, version 12.6906, Wellcome Trust Centre for
637 Human Neuroimaging, UCL, UK, http://www.fil.ion.ucl.ac.uk/spm) running on Matlab 8.6 (The Mathworks
638 Inc., Natick, MA, USA). The resulting grey and white matter segmentations were summed and binarized
639 to remove voxels that contain air, scalp, skull and cerebrospinal fluid from structural images using the
640 ImCalc function of SPM.

641 A template of all participants was created with ANTs (Avants et al., 2009) using the participants' MP2RAGE
642 images, which was then registered to the MNI space using the same software package and the MNI152
643 (0.5 mm)^3 voxel size template provided by FSL 5.0.8 (Smith et al., 2004). All MP2RAGE images were
644 preprocessed with Freesurfer (Fischl et al., 2002, Fischl et al., 2004, Han and Fischl, 2007) using the recon-
645 all command to obtain boundaries between grey and white matter, which were later used in the
646 functional to structural registration step.

647 The rest of the analysis was coded in nipype (Gorgolewski et al., 2011). A graphical overview of the nipype
648 pipeline can be found in Figure 6-figure supplement 3. Head motion and susceptibility distortion by
649 movement interaction of functional runs were corrected using the Realign and Unwarp method

650 (Andersson et al., 2001) in SPM 12 after which outlier volumes were detected using ArtifactDetect
651 (composite threshold of translation and rotation: 1; intensity Z-threshold: 3; global threshold: 8;
652 https://www.nitrc.org/projects/artifact_detect/). Coregistration matrices for realigned functional runs
653 per participant were computed based on each participant's structural image using Freesurfer's BBregister
654 function (register mean EPI image to T1, option '—init-header' was specified in order to preserve the
655 origin of the manual alignment of structural and functional data). Warping using coregistration matrices
656 (after conversion to ITK coordinate system) and resampling to 1 mm isovoxel was performed using ANTs.
657 Before model creation we smoothed the data in SPM12 using a 1 mm kernel at full-width half-maximum.

658 ***Physiological data***

659 Physiological data (heart rate and respiration rate) were processed by the PhysIO Toolbox (Kasper et al.,
660 2017) to obtain Fourier expansions of each, in order to enter these into the design matrix (see statistical
661 analyses sections below).

662 ***Statistical analysis of speech experiment***

663 Models were set up in SPM using the native space data for each participant. The design matrix included
664 three cardiac and four respiratory regressors, six realignment parameters, and a variable number of
665 outlier regressors from the ArtifactDetect step, depending on how many outliers were found in each run.
666 These regressors of no interest were also used in the models of the other two experiments (MGB and
667 tonotopy localizer). Since participants provided a response only for the target stimulus changes and not
668 for each stimulus presentation, we modeled these to eschew a potential sensory-motor confound as 0.5
669 for hit, –0.5 for miss and 0.0 for everything else. If more than one syllable presentation took place within
670 one volume acquisition, the values within this volume were averaged. The speech experiment had a total
671 of five modeled conditions, which were convolved with the hemodynamic response function (HRF):
672 speech task/vowel change, speech task/consonant change, speaker task/vowel change, speaker
673 task/consonant change, and task instruction. Parameter estimates were computed for the contrast
674 speech vs speaker at the first level using restricted maximum likelihood (REML) as implemented in SPM
675 12.

676 After estimation, the contrasts were registered to the MNI structural template of all participants using a
677 two-step registration in ANTs (see also Figure 6-figure supplement 3). First, a quick registration was
678 performed on the whole head using rigid, affine and diffeomorphic (using Symmetric Normalization: SyN)
679 transformations and the mutual information similarity metric. Second, the high quality registration was
680 confined to a rectangular prism mask encompassing the left and right MGB, and IC only. This step used

affine and SyN transformations and mean squares and neighborhood cross correlation similarity measures, respectively. We performed the registration to MNI space for all experiments by linearly interpolating the contrast images using the composite transforms from the high quality registration.

We used a random effects (RFX) analysis to compute the speech vs speaker contrast across participants to test our first hypothesis that the MGB response is modulated by this contrast. To do this we took the first level contrasts across participants and entered them into an RFX model to be estimated using REML. Based on the results of previous experiments (Díaz et al., 2012, von Kriegstein et al., 2008), we expected a result for the categorical speech vs speaker contrast in the left and right MGB. Our second hypothesis was that the proportion of correct button presses in the speech task correlates with the responses elicited by the speech vs speaker contrast over participants in the left MGB only (von Kriegstein et al., 2008, Díaz et al., 2012). We thus computed the RFX correlation between the speech vs speaker contrast and the proportion of correct button presses in the speech task across participants. This was implemented using the behavioral percent correct scores for the speech task as a covariate of interest for each participant in the SPM RFX model. We used an equivalent procedure to test for correlation between the speech vs speaker task contrast and the proportion of correct button presses in the speaker task across participants. To formally test for a difference between these two correlations we performed a comparison using a freely available online tool (comparingcorrelations.org) (Diedenhofen and Musch, 2015). This tool computes the z-score and p-value of the difference between two correlations based on (Hittner et al., 2003).

Meta-analysis of the main effect of task (speech vs speaker task contrast)

The lack of statistical significance for the speech vs speaker contrast raised the question whether the overall effect is different from the ones reported previously (Díaz et al., 2012, 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 task-dependent modulation of the MGB. We included five studies in the meta-analysis that included a speech task vs control task contrast: two experiments from (von Kriegstein et al., 2008), the data from the control participants of (Díaz et al., 2012), the result of a recent study (Mihai et al., 2019), and the result of the current study. Effect sizes and standard errors were entered into a random effects model that was estimated with maximum likelihood using JASP 0.9 (jasp-stats.org).

Statistical analysis of the MGB localizer

For the MGB localizer we used a stick function convolved with the HRF to model each presented sound. Null events were not modeled, as well as repeated sounds, to avoid a sensory-motor confound through the button-press. The data were modeled according to Perrachione and Ghosh (2013) where repetition (TR = 2.8 s) and acquisition times (TA = 1.6 s) were modeled separately. The contrast *Sound vs Silence* was computed for each participant. The inference across participants was modeled using the first level contrasts in a second-level RFX analysis for the group. Significant voxels (see Section Masks below) in the left and right MGB found in the RFX analysis for the contrast sound vs silence were used as a mask for the tonotopy localizer.

Statistical analysis of the tonotopy localizer

For the tonotopy localizer we followed a similar approach as Moerel et al. (2015). The sounds were first processed through the NSL toolbox (Chi et al., 2005) which mimics the spectral transformation of sounds passing through the cochlea to the midbrain. This frequency representation includes a bank of 128 overlapping bandpass filters equally spaced on a log frequency axis (180-7040 Hz; range 5.3 octaves). The resulting spectrograms were averaged over time. To reduce overfitting we divided the tonotopic axis into 12 equal bandwidths in octaves and averaged the model's output within these regions. The MrConfon headphones guarantee a linear frequency response up to 4 kHz, thus only the first 10 bins were used in the analysis, which resulted in 10 frequency bins for each sound file. The frequency model consisted of a vector of values corresponding to the frequency representations per sound. Since each sound had a frequency representation the final model is a matrix $W = [S \times F]$, where S is the number of sounds and F the number of features per sound. The predictors were z-scored across bins since low frequencies have more energy and would thus be more strongly represented compared to high frequencies (Moerel et al., 2015). The matrix was convolved with the hemodynamic response function and its components (i.e., the 10 frequency bins) were used as regressors of interest in the design matrix of SPM. In addition, we included the same regressors of no-interest as in the design matrix for the speech experiment (i.e., six respiratory regressors, six realignment parameters, and a variable number of outlier regressors from the ArtifactDetect step, depending on how many outliers were found). Parameter estimates were calculated for each frequency bin at the first level in native space.

Masks

MGB localizer: We created masks using all voxels from the second level MGB localizer analysis for the contrast sound vs silence (family-wise error [FWE] corrected $p < 0.001$) constrained within a $r = 5$ mm

sphere centered at the voxel with the statistical maximum in the left and right MGB. We chose such a stringent p-value due to the strong effect and the multitude of above threshold voxels found within and around the left and right MGB. This procedure excluded all voxels which were clearly too far away from the structural boundaries of the MGB as seen in the MP2RAGE MNI template, yet still within the cluster, to be considered part of the MGB. These masks were inverse transformed per participant from MNI space to participant space using ANTs. Above threshold voxels (uncorrected $p < 0.05$) within the transformed masks were extracted, for each participant, from the MGB localizer *Sound vs Silence* contrast. These masks were then used to define each participant's tonotopy with the tonotopy localizer.

Tonotopy localizer: Each voxel within each participant's left and right MGB and IC localizer mask was labeled according to the frequency bin to which it responded strongest, i.e., which had the highest parameter estimate (Moerel et al., 2015). Thus, voxels would have values from 1-10 corresponding to the frequency bin that they best represented. This resulted in a map of frequency distributions from low to high frequencies in the left and right MGB for each participant. To create masks at the group level, these tonotopic maps were registered to MNI space using ANTs and averaged across participants.

To evaluate the tonotopic representations in the MGB and IC in a similar way as Moerel et al. (2015), we visually inspected the direction which showed the strongest tonotopy. This was a dorsal-lateral to ventral-medial gradient that was most visible in a sagittal view. We thus rotated and resliced the individual maps around the z-axis by $90°$, which placed the sagittal view in the x-y plane. In this plane we calculated gradient directions in 10 adjacent slices, ensuring a representative coverage of the tonotopic pattern. A cut at $90°$ captured both low and high frequency areas. Histograms in 5° steps were calculated for each slice. The histograms of the gradients were then averaged first over slices per participant, followed by an average over participants. Based on the atlas by (Morel et al., 1997) and findings of MGB subdivision in awake primates (Bartlett and Wang et al., 2011) we parcellated the resulting frequency gradients as distinct regions. Voxels that represented the highest frequency were chosen as the boundary within each slice. Voxels above this boundary corresponded to one region, and those below this boundary to the other region. The regions were drawn in each slice using ITKSnap (v. 3.6.0; Yushkevich et al., 2006). Volume size and center of mass (COM) for each gradient are listed in Table 1.

For the IC, we created functional masks of IC responses from the MGB localizer experiment (Sound vs Silence, uncorrected $p < 0.05$) and constrained these to the anatomical volumes of the IC. Frequency distribution maps were calculated per participant in MNI space and averaged. Gradient directions were calculated from the mean tonotopy maps in three different slices. Histograms were then averaged and

plotted. As the IC was not part of the main objective of this manuscript, we report it in the Supplementary File 1 and Figure 2-figure supplement 1.

Unthresholded t-maps of contrasts of interest, tonotopy maps of the MGB and IC, as well as vMGB masks are available on neurovault (https://neurovault.org/collections/CSGJQQDW/).

Significance testing

We used small volume corrections (SVC) to test for significant voxels for the contrast speech vs speaker task as well as the correlation of speech vs speaker task with the behavioral proportion correct scores in the speech task (significance defined as $p < 0.05$ FWE corrected for the region of interest). We tested bilaterally using the vMGB masks described above for the first hypothesis (main effect of task) and left vMGB for the second hypothesis (correlation between speech recognition performance and main effect of task) motivated by findings in previous studies (von Kriegstein et al., 2008, Díaz et al., 2012).

References

ALCÁNTARA, J. I., WEISBLATT, E. J. L., MOORE, B. C. J. & BOLTON, P. F. 2004. Speech-in-noise perception in high-functioning individuals with autism or Asperger's syndrome. *Journal of Child Psychology and Psychiatry,* 45**,** 1107-1114.

ANDERSON, L. A., CHRISTIANSON, G. B. & LINDEN, J. F. 2009. Stimulus-Specific Adaptation Occurs in the Auditory Thalamus. *Journal of Neuroscience,* 29**,** 7359-7363.

ANDERSON, L. A. & LINDEN, J. F. 2011. Physiological differences between histologically defined subdivisions in the mouse auditory thalamus. *Hearing Research,* 274**,** 48-60.

ANDERSON, L. A., WALLACE, M. N. & PALMER, A. R. 2007a. Identification of subdivisions in the medial geniculate body of the guinea pig. *Hearing Research,* 228**,** 156-167.

ANDERSON, L. A., WALLACE, M. N. & PALMER, A. R. 2007b. Identification of subdivisions in the medial geniculate body of the guinea pig. *Hearing Research,* 228**,** 156-167.

ANDERSSON, J. L., HUTTON, C., ASHBURNER, J., TURNER, R. & FRISTON, K. 2001. Modeling geometric deformations in EPI time series. *NeuroImage,* 13**,** 903-19.

ANDOLINA, I. M., JONES, H. E., WANG, W. & SILLITO, A. M. 2007. Corticothalamic feedback enhances stimulus response precision in the visual system. *Proceedings of the National Academy of Sciences,* 104**,** 1685-1690.

ANTUNES, F. M. & MALMIERCA, M. S. 2011. Effect of auditory cortex deactivation on stimulus-specific adaptation in the medial geniculate body. *J Neurosci,* 31**,** 17306-16.

AVANTS, B. B., TUSTISON, N. & SONG, G. 2009. Advanced normalization tools (ANTS). *Insight j,* 2**,** 1-35.

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. *Journal of Autism and Developmental Disorders,* 31**,** 5-17.
BARTLETT, E. L., SADAGOPAN, S. & WANG, X. 2011. Fine frequency tuning in monkey auditory cortex and thalamus. *Journal of Neurophysiology,* 106**,** 849-859.
BARTLETT, E. L. & WANG, X. 2011. Correlation of neural response properties with auditory thalamus subdivisions in the awake marmoset. *Journal of Neurophysiology,* 105**,** 2647-2667.
BAUMANN, S., GRIFFITHS, T. D., SUN, L., PETKOV, C. I., THIELE, A. & REES, A. 2011. Orthogonal representation of sound dimensions in the primate midbrain. *Nature Neuroscience,* 14**,** 423-425.
BENESTY, J., SONDHI, M. M. & HUANG, Y. 2007. *Springer Handbook of Speech Processing*, Springer Berlin Heidelberg.
BORDI, F. & LEDOUX, J. E. 1994. Response properties of single units in areas of rat auditory thalamus that project to the amygdala. *Experimental Brain Research,* 98**,** 261-274.
CALFORD, M. B. 1983. The parcellation of the medial geniculate body of the cat defined by the auditory response properties of single units. *Journal of Neuroscience,* 3**,** 2350-2364.
CAMARILLO, L., LUNA, R., NÁCHER, V. & ROMO, R. 2012. Coding perceptual discrimination in the somatosensory thalamus. *Proceedings of the National Academy of Sciences,* 109**,** 21093-21098.
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.
CHANDRASEKARAN, B., KRAUS, N. & WONG, P. C. M. 2011. Human inferior colliculus activity relates to individual differences in spoken language learning. *Journal of Neurophysiology,* 107**,** 1325-1336.
CHI, T., RU, P. & SHAMMA, S. A. 2005. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America,* 118**,** 887-906.
COHEN, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates.
CRUIKSHANK, S. J., KILLACKEY, H. P. & METHERATE, R. 2001. Parvalbumin and calbindin are differentially distributed within primary and secondary subregions of the mouse auditory forebrain. *Neuroscience,* 105**,** 553-569.
CUDEIRO, J. & SILLITO, A. M. 2006. Looking back: corticothalamic feedback and early visual processing. *Trends in Neurosciences,* 29**,** 298-306.
DAVIS, M. H. & JOHNSRUDE, I. S. 2003. Hierarchical Processing in Spoken Language Comprehension. *The Journal of Neuroscience,* 23**,** 3423-3431.
DAVIS, M. H. & JOHNSRUDE, I. S. 2007. Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research,* 229**,** 132-147.
DE MARTINO, F., MOEREL, M., VAN DE MOORTELE, P.-F., UGURBIL, K., GOEBEL, R., YACOUB, E. & FORMISANO, E. 2013. Spatial organization of frequency preference and selectivity in the human inferior colliculus. *Nature Communications,* 4**,** 1386.
DÍAZ, B., BLANK, H. & VON KRIEGSTEIN, K. 2018. Task-dependent modulation of the visual sensory thalamus assists visual-speech recognition. *NeuroImage,* 178**,** 721-734.
DÍAZ, B., HINTZ, F., KIEBEL, S. J. & KRIEGSTEIN, K. V. 2012. Dysfunction of the auditory thalamus in developmental dyslexia. *Proceedings of the National Academy of Sciences,* 109**,** 13841-13846.
DIEDENHOFEN, B. & MUSCH, J. 2015. cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLOS ONE,* 10**,** e0121945.
DUYN, J. H. 2012. The future of ultra-high field MRI and fMRI for study of the human brain. *NeuroImage,* 62**,** 1241-1248.
ELLIOTT, T. M. & THEUNISSEN, F. E. 2009. The Modulation Transfer Function for Speech Intelligibility. *PLOS Computational Biology,* 5**,** e1000302.
ERGENZINGER, E. R., GLASIER, M. M., HAHM, J. O. & PONS, T. P. 1998. Cortically induced thalamic plasticity in the primate somatosensory system. *Nature Neuroscience,* 1**,** 226-229.

FISCHL, B., SALAT, D. H., BUSA, E., ALBERT, M., DIETERICH, M., HASELGROVE, C., VAN DER KOUWE, A., KILLIANY, R., KENNEDY, D., KLAVENESS, S., MONTILLO, A., MAKRIS, N., ROSEN, B. & DALE, A. M. 2002. Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. *Neuron,* 33**,** 341-355.

FISCHL, B., SALAT, D. H., VAN DER KOUWE, A. J. W., MAKRIS, N., SÉGONNE, F., QUINN, B. T. & DALE, A. M. 2004. Sequence-independent segmentation of magnetic resonance images. *NeuroImage,* 23**,** S69-S84.

FISHER, R. A. 1915. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika,* 10**,** 507-521.

FOSTER, K. H., GASKA, J. P., NAGLER, M. & POLLEN, D. A. 1985. Spatial and temporal frequency selectivity of neurones in visual cortical areas V1 and V2 of the macaque monkey. *The Journal of Physiology,* 365**,** 331-363.

FRIEDERICI, A. D. & GIERHAN, S. M. 2013. The language network. *Current Opinion in Neurobiology,* 23**,** 250-254.

FRISTON, K. 2005. A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences,* 360**,** 815-836.

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.

GALABURDA, A. M., MENARD, M. T. & ROSEN, G. D. 1994. Evidence for aberrant auditory anatomy in developmental dyslexia. *Proceedings of the National Academy of Sciences,* 91**,** 8010-8013.

GAUDRAIN, E., LI, S., BAN, V. S. & PATTERSON, R. D. 2009. The Role of Glottal Pulse Rate and Vocal Tract Length in the Perception of Speaker Identity. *Interspeech-2009***,** 148-151.

GHAZANFAR, A. A. & NICOLELIS, M. A. L. 2001. Feature Article: The Structure and Function of Dynamic Cortical and Thalamic Receptive Fields. *Cerebral Cortex,* 11**,** 183-193.

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.

GONZALEZ-LIMA, F. & CADA, A. 1994. Cytochrome oxidase activity in the auditory system of the mouse: A qualitative and quantitative histochemical study. *Neuroscience,* 63**,** 559-578.

GORDON-SALANT, S. & FITZGIBBONS, P. J. 1993. Temporal Factors and Speech Recognition Performance in Young and Elderly Listeners. *Journal of Speech, Language, and Hearing Research,* 36**,** 1276-1285.

GORGOLEWSKI, K., BURNS, C. D., MADISON, C., CLARK, D., HALCHENKO, Y. O., WASKOM, M. L. & GHOSH, S. S. 2011. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front Neuroinform,* 5**,** 13.

GRISWOLD, M. A., JAKOB, P. M., HEIDEMANN, R. M., NITTKA, M., JELLUS, V., WANG, J., KIEFER, B. & HAASE, A. 2002. Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic Resonance in Medicine,* 47**,** 1202-1210.

GROEN, W. B., VAN ORSOUW, L., HUURNE, N. T., SWINKELS, S., VAN DER GAAG, R.-J., BUITELAAR, J. K. & ZWIERS, M. P. 2009. Intact Spectral but Abnormal Temporal Processing of Auditory Stimuli in Autism. *Journal of Autism and Developmental Disorders,* 39**,** 742-750.

HACKETT, T. A., BARKAT, T. R., O'BRIEN, B. M. J., HENSCH, T. K. & POLLEY, D. B. 2011. Linking Topography to Tonotopy in the Mouse Auditory Thalamocortical Circuit. *Journal of Neuroscience,* 31**,** 2983-2995.

HACKETT, T. A., STEPNIEWSKA, I. & KAAS, J. H. 1998. Thalamocortical connections of the parabelt auditory cortex in macaque monkeys. *The Journal of Comparative Neurology,* 400**,** 271-286.

HAN, X. & FISCHL, B. 2007. Atlas Renormalization for Improved Brain MR Image Segmentation Across Scanner Platforms. *IEEE Transactions on Medical Imaging,* 26**,** 479-486.

HAYNES, J. D., DEICHMANN, R. & REES, G. 2005. Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus. *Nature,* 438**,** 496-9.

HAYWARD, K. 2000. *Experimental phonetics,* Harlow, Eng. ; New York, Longman.

HICKOK, G. & POEPPEL, D. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience,* 8**,** 393-402.

HICKS, T. P., LEE, B. B. & VIDYASAGAR, T. R. 1983. The responses of cells in macaque lateral geniculate nucleus to sinusoidal gratings. *The Journal of Physiology,* 337**,** 183-200.

HITTNER, J. B., MAY, K. & SILVER, N. C. 2003. A Monte Carlo Evaluation of Tests for Comparing Dependent Correlations. *The Journal of General Psychology,* 130**,** 149-168.

HOFFMAN, M. D. & GELMAN, A. 2011. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *arXiv:1111.4246 [cs, stat]*.

HOLLANDER, G. D., KEUKEN, M. C., ZWAAG, W. V. D., FORSTMANN, B. U. & TRAMPEL, R. 2017. Comparing functional MRI protocols for small, iron-rich basal ganglia nuclei such as the subthalamic nucleus at 7 T and 3 T. *Human Brain Mapping,* 38**,** 3226-3248.

HORIE, M., TSUKANO, H., HISHIDA, R., TAKEBAYASHI, H. & SHIBUKI, K. 2013. Dual compartments of the ventral division of the medial geniculate body projecting to the core region of the auditory cortex in C57BL/6 mice. *Neuroscience Research,* 76**,** 207-212.

HUTH, A. G., DE HEER, W. A., GRIFFITHS, T. L., THEUNISSEN, F. E. & GALLANT, J. L. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature,* 532**,** 453-458.

IMIG, T. J. & MOREL, A. 1985. Tonotopic organization in ventral nucleus of medial geniculate body in the cat. *Journal of Neurophysiology,* 53**,** 309-340.

JONES, E. G. 1985. *The thalamus*, Plenum Press, New York.

KASPER, L., BOLLMANN, S., DIACONESCU, A. O., HUTTON, C., HEINZLE, J., IGLESIAS, S., HAUSER, T. U., SEBOLD, M., MANJALY, Z.-M., PRUESSMANN, K. P. & STEPHAN, K. E. 2017. The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data. *Journal of Neuroscience Methods,* 276**,** 56-72.

KENT, R. D., READ, C. & KENT, R. D. 1992. *The acoustic analysis of speech*, Singular Publishing Group San Diego.

KIEBEL, S. J., DAUNIZEAU, J. & FRISTON, K. J. 2008. A Hierarchy of Time-Scales and the Brain. *PLOS Computational Biology,* 4**,** e1000209.

KLOSTERMANN, F. 2013. Functional roles of the thalamus for language capacities. *Frontiers in Systems Neuroscience,* 7.

KNILL, D. C. & POUGET, A. 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences,* 27**,** 712-719.

KRIEGSTEIN, K. V., DOGAN, Ö., GRÜTER, M., GIRAUD, A.-L., KELL, C. A., GRÜTER, T., KLEINSCHMIDT, A. & KIEBEL, S. J. 2008. Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences,* 105**,** 6747-6752.

KRUPA, D. J., GHAZANFAR, A. A. & NICOLELIS, M. A. L. 1999. Immediate thalamic sensory plasticity depends on corticothalamic feedback. *Proceedings of the National Academy of Sciences,* 96**,** 8200-8205.

LEE, C. C. 2013. Thalamic and cortical pathways supporting auditory processing. *Brain and Language,* 126**,** 22-28.

LEE, C. C. & SHERMAN, S. M. 2012. Intrinsic modulators of auditory thalamocortical transmission. *Hearing Research,* 287**,** 43-50.

LEE, C. C. & WINER, J. A. 2011. Convergence of thalamic and cortical pathways in cat auditory cortex. *Hearing Research,* 274**,** 85-94.

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.

LLANO, D. A. & SHERMAN, S. M. 2001. Evidence for nonreciprocal organization of the mouse auditory thalamocortical-corticothalamic projection systems. *Journal of Comparative Neurology,* 507**,** 1209-1227.

MALMIERCA, M. S., ANDERSON, L. A. & ANTUNES, F. M. 2015. The cortical modulation of stimulus-specific adaptation in the auditory midbrain and thalamus: a potential neuronal correlate for predictive coding. *Frontiers in Systems Neuroscience,* 9.

MARQUES, J. P., 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.

MCALONAN, K., CAVANAUGH, J. & WURTZ, R. H. 2008. Guarding the gateway to cortex with attention in visual thalamus. *Nature,* 456**,** 391-4.

MCELREATH, R. 2015. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* [Online]. [Accessed].

MIHAI, P. G., TSCHENTSCHER, N. & KRIEGSTEIN, K. V. 2019. The role of the ventral MGB in speech in noise comprehension. *bioRxiv***,** 646570.

MOEREL, M., DE MARTINO, F., UĞURBIL, K., YACOUB, E. & FORMISANO, E. 2015. Processing of frequency and location in human subcortical auditory structures. *Scientific Reports,* 5.

MOREL, A., MAGNIN, M. & JEANMONOD, D. 1997. Multiarchitectonic and stereotactic atlas of the human thalamus. *The Journal of Comparative Neurology,* 387**,** 588-630.

MOREST, D. K. 1964. The neuronal architecture of the medial geniculate body of the cat. *Journal of Anatomy,* 98**,** 611-630.1.

MORO, S. S., KELLY, K. R., MCKETTON, L., GALLIE, B. L. & STEEVES, J. K. E. 2015. Evidence of multisensory plasticity: Asymmetrical medial geniculate body in people with one eye. *NeuroImage: Clinical,* 9**,** 513-518.

MOTHE, L. A. D. L., BLUMELL, S., KAJIKAWA, Y. & HACKETT, T. A. 2006. Thalamic connections of the auditory cortex in marmoset monkeys: Core and medial belt regions. *Journal of Comparative Neurology,* 496**,** 72-96.

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.

O'CONNOR, D. H., FUKUI, M. M., PINSK, M. A. & KASTNER, S. 2002. Attention modulates responses in the human lateral geniculate nucleus. *Nat Neurosci,* 5**,** 1203-9.

OHGA, S., TSUKANO, H., HORIE, M., TERASHIMA, H., NISHIO, N., KUBOTA, Y., TAKAHASHI, K., HISHIDA, R., TAKEBAYASHI, H. & SHIBUKI, K. 2018. Direct Relay Pathways from Lemniscal Auditory Thalamus to Secondary Auditory Field in Mice. *Cerebral Cortex*.

OJIMA, H. & ROUILLER, E. M. 2011. Auditory Cortical Projections to the Medial Geniculate Body. *In:* WINER, J. A. & SCHREINER, C. E. (eds.) *The Auditory Cortex.* Boston, MA: Springer US.

OLDFIELD, R. C. 1971. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia,* 9**,** 97-113.

PANOUILLÈRES, M. T. N. & MÖTTÖNEN, R. 2018. Decline of auditory-motor speech processing in older adults with hearing loss. *Neurobiology of Aging,* 72**,** 89-97.

PERRACHIONE, T. K. & GHOSH, S. S. 2013. Optimized design and analysis of sparse-sampling fMRI experiments. *Brain Imaging Methods,* 7**,** 55.

PRICE, C., THIERRY, G. & GRIFFITHS, T. 2005. Speech-specific auditory processing: where is it? *Trends in Cognitive Sciences,* 9**,** 271-276.

RADEMACHER, J., BÜRGEL, U. & ZILLES, K. 2002. Stereotaxic Localization, Intersubject Variability, and Interhemispheric Differences of the Human Auditory Thalamocortical System. *NeuroImage,* 17**,** 142-160.

RAUSCHECKER, J. P. & SCOTT, S. K. 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience,* 12**,** 718-724.

RODRIGUES-DAGAEFF, C., SIMM, G., DE RIBAUPIERRE, Y., VILLA, A., DE RIBAUPIERRE, F. & ROUILLER, E. M. 1989. Functional organization of the ventral division of the medial geniculate body of the cat: Evidence for a rostro-caudal gradient of response properties and cortical projections. *Hearing Research,* 39**,** 103-125.

ROUILLER, E. M. & 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.

SAALMANN, Y. B. & KASTNER, S. 2011. Cognitive and perceptual functions of the visual thalamus. *Neuron,* 71**,** 209-23.

SAFFRAN, J. R. 2003. Statistical Language Learning: Mechanisms and Constraints. *Current Directions in Psychological Science,* 12**,** 110-114.

SALVATIER, J., WIECKI, T. V. & FONNESBECK, C. 2016. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science,* 2**,** e55.

SCHNEIDER, W., ENNEMOSER, M. & SCHLAGMÜLLER, M. 2007. *Lesegeschwindigkeits- und -verständnistest für die Klassen 6 - 12: LGVT 6 - 12*, Hogrefe.

SCHUIRMANN, D. J. 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics,* 15**,** 657–680.

SHANNON, R. V., ZENG, F.-G., KAMATH, V., WYGONSKI, J. & EKELID, M. 1995a. Speech Recognition with Primarily Temporal Cues. *Science,* 270**,** 303-304.

SHANNON, R. V., ZENG, F.-G., KAMATH, V., WYGONSKI, J. & EKELID, M. 1995b. Speech Recognition with Primarily Temporal Cues. *Science,* 270**,** 303-304.

SHERMAN, S. M. & GUILLERY, R. W. 1998. On the actions that one nerve cell can have on another: Distinguishing “drivers” from “modulators”. *Proceedings of the National Academy of Sciences,* 95**,** 7121-7126.

SHERMAN, S. M. & GUILLERY, R. W. 2006. *Exploring the Thalamus and Its Role in Cortical Function*, MIT Press.

SHIRAMATSU, T. I., TAKAHASHI, K., NODA, T., KANZAKI, R., NAKAHARA, H. & TAKAHASHI, H. 2016. Microelectrode mapping of tonotopic, laminar, and field-specific organization of thalamo-cortical pathway in rat. *Neuroscience,* 332**,** 38-52.

SILLITO, A. M., CUDEIRO, J. & JONES, H. E. 2006. Always returning: feedback and sensory processing in visual cortex and thalamus. *Trends in Neurosciences,* 29**,** 307-316.

SILLITO, A. M., JONES, H. E., GERSTEIN, G. L. & WEST, D. C. 1994. Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex. *Nature,* 369**,** 479-482.

SMITH, S. M., JENKINSON, M., WOOLRICH, M. W., BECKMANN, C. F., BEHRENS, T. E. J., JOHANSEN-BERG, H., BANNISTER, P. R., DE LUCA, M., DROBNJAK, I., FLITNEY, D. E., NIAZY, R. K., SAUNDERS, J., VICKERS, J., ZHANG, Y., DE STEFANO, N., BRADY, J. M. & MATTHEWS, P. M. 2004. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage,* 23**,** S208-S219.

TALLAL, P., MILLER, S. L., BEDI, G., BYMA, G., WANG, X., NAGARAJAN, S. S., SCHREINER, C., JENKINS, W. M. & MERZENICH, M. M. 1996. Language Comprehension in Language-Learning Impaired Children Improved with Acoustically Modified Speech. *Science,* 271**,** 81-84.

TALLAL, P. & PIERCY, M. 1975. Developmental aphasia: The perception of brief vowels and extended stop consonants. *Neuropsychologia,* 13**,** 69-74.

TSCHENTSCHER, N., RUISINGER, A., BLANK, H., DÍAZ, B. & KRIEGSTEIN, K. V. 2019. Reduced Structural Connectivity Between Left Auditory Thalamus and the Motion-Sensitive Planum Temporale in Developmental Dyslexia. *Journal of Neuroscience,* 39**,** 1720-1732.

TSUKANO, H., HORIE, M., OHGA, S., TAKAHASHI, K., KUBOTA, Y., HISHIDA, R., TAKEBAYASHI, H. & SHIBUKI, K. 2017. Reconsidering Tonotopic Maps in the Auditory Cortex and Lemniscal Auditory Thalamus in Mice. *Frontiers in Neural Circuits,* 11.

VASQUEZ-LOPEZ, S. A., WEISSENBERGER, Y., LOHSE, M., KEATING, P., KING, A. J. & DAHMEN, J. C. 2017. Thalamic input to auditory cortex is locally heterogeneous but globally tonotopic. *Elife,* 6**,** e25141.

VON KRIEGSTEIN, K., PATTERSON, R. D. & GRIFFITHS, T. D. 2008. Task-dependent modulation of medial geniculate body is behaviorally relevant for speech recognition. *Curr Biol,* 18**,** 1855-9.

WANG, W., ANDOLINA, I. M., LU, Y., JONES, H. E. & SILLITO, A. M. 2018a. Focal Gain Control of Thalamic Visual Receptive Fields by Layer 6 Corticothalamic Feedback. *Cerebral Cortex,* 28**,** 267-280.

WANG, X., LU, T., BENDOR, D. & BARTLETT, E. 2008. Neural coding of temporal information in auditory thalamus and cortex. *Neuroscience,* 154**,** 294-303.

WANG, Y., ZHANG, J., ZOU, J., LUO, H. & DING, N. 2018b. Prior Knowledge Guides Speech Segregation in Human Auditory Cortex. *Cerebral Cortex*.

WERNICKE, C. 1874. *Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis*, Cohn.

WHITE, S., FRITH, U., MILNE, E., ROSEN, S., SWETTENHAM, J. & RAMUS, F. 2006. A double dissociation between sensorimotor impairments and reading disability: A comparison of autistic and dyslexic children. *Cognitive Neuropsychology,* 23**,** 748-761.

WINER, J. A. 1984. The human medial geniculate body. *Hearing Research,* 15**,** 225-247.

WINER, J. A., KELLY, J. B. & LARUE, D. T. 1999. Neural architecture of the rat medial geniculate body. *Hearing Research,* 130**,** 19-41.

WINER, J. A., MILLER, L. M., LEE, C. C. & SCHREINER, C. E. 2005. Auditory thalamocortical transformation: structure and function. *Trends in Neurosciences,* 28**,** 255-263.

WINER, J. A. & PRIETO, J. J. 2001. Layer V in cat primary auditory cortex (AI): Cellular architecture and identification of projection neurons. *Journal of Comparative Neurology,* 434**,** 379-412.

YUSHKEVICH, P. A., PIVEN, J., HAZLETT, H. C., SMITH, R. G., HO, S., GEE, J. C. & GERIG, G. 2006. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage,* 31**,** 1116-1128.