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

Abstract

Sensory thalami are central sensory pathway stations for information processing. Their role for human cognition and perception, however, remains unclear. Recent evidence suggests a specific 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 text book view of the sensory thalamus is still that of a passive relay station (Squire et al., 2012), although 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, over the last two decades, 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 than the text book view implies (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). Specifically, the MGB showed significantly higher responses to an auditory speech recognition task than to control tasks, independent of attentional load (von Kriegstein et al., 2008, Díaz et al., 2012). The performance level in the auditory speech recognition task was significantly correlated with the task-dependent modulation in the MGB of the left hemisphere (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.), 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., 2007b, Mothe et al., 2006), and are not considered to be part of the first order (i.e., lemniscal) auditory pathway (Anderson et al., 2007a, 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).

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

Due to the relatively small size of human MGB (ca. 5×4×5 mm, Winer, 1984) and the spatial limitations of non-invasive imaging techniques, it was so far not possible to differentiate between the three major MGB divisions in order to test localization of this task-dependent modulation to the lemniscal part of the MGB. Here we therefore used ultra-high field functional magnetic resonance imaging (fMRI) at 7 Tesla, enabling high spatial resolution measurements (Duyn, 2012). The vMGB has a strong tonotopic organization (Calford, 1983, Rodrigues-Dagaeff et al., 1989, Anderson et al., 2007a) while the other two MGB subsections have only a weak tonotopic organization (i.e., broadly tuned neurons, Anderson and

Linden, 2011, Calford, 1983, Bartlett and Wang, 2011, Rodrigues-Dagaeff et al., 1989, Ohga et al., 2018). We planned to distinguish the vMGB based on its tonotopic organization as well as its topographic (i.e., ventral) location.

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

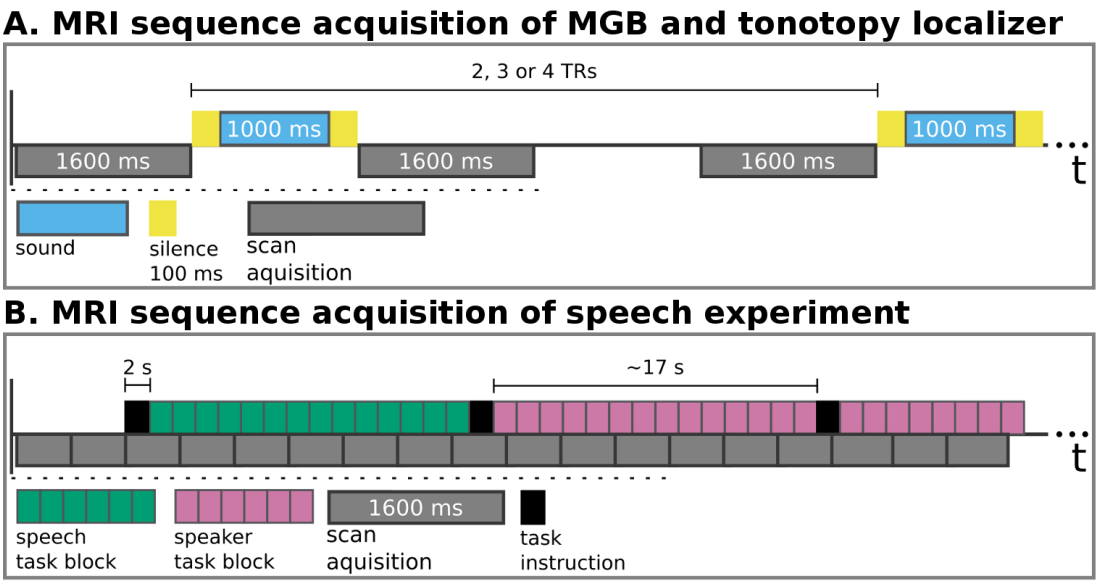

Figure 1. A. MRI sequence acquisition of MGB and tonotopy localizer. Stimuli ('sound') were presented in silence periods between scan acquisitions and jittered with 2, 3, or 4 TRs. TR: repetition time of volume acquisition. **B.** MRI sequence acquisition of the speech experiment. Each green or magenta 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.

In previous studies we found the task-dependent modulation for speech (i.e., higher response in the speech in contrast to a control task on the same stimulus material) in both the left and right MGB and a correlation of the task-dependent modulation with speech recognition performance only in the left MGB (Kriegstein et al., 2008, Díaz et al., 2012). We therefore hypothesized (i) a higher response to the speech than to the control (speaker) task in the tonotopically organized left and right vMGB, and (ii) a positive correlation between speech recognition performance and the task-dependent modulation for speech in the tonotopically organized left vMGB.

Results

Tonotopy localizer – replication of tonotopy in MGB

First, we aimed to replicate the MGB tonotopy reported previously by Moerel et al. (2015) with a larger participant sample. Participants listened to natural sounds (human voices, animal cries, tool sounds) in a fast event-related scheme during silent gaps of the clustered imaging technique (Moerel et al., 2015). Using a model that mimics peripheral sound processing (Chi et al., 2005), each sound was represented as a spectrogram. The resulting spectrograms were averaged over time and divided into ten equal bandwidths in octaves. Onsets for each bin were convolved with the hemodynamic response function and entered into the general linear model. Each voxel within each participant's left and right MGB 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.

Similar as in Moerel et al. (2015), we found two tonotopic gradients within the MGB in the group analysis. On visual inspection, one high frequency region in the middle of the MGB was flanked by gradually lower frequency components dorsally and ventrally (Figure 2A and B).

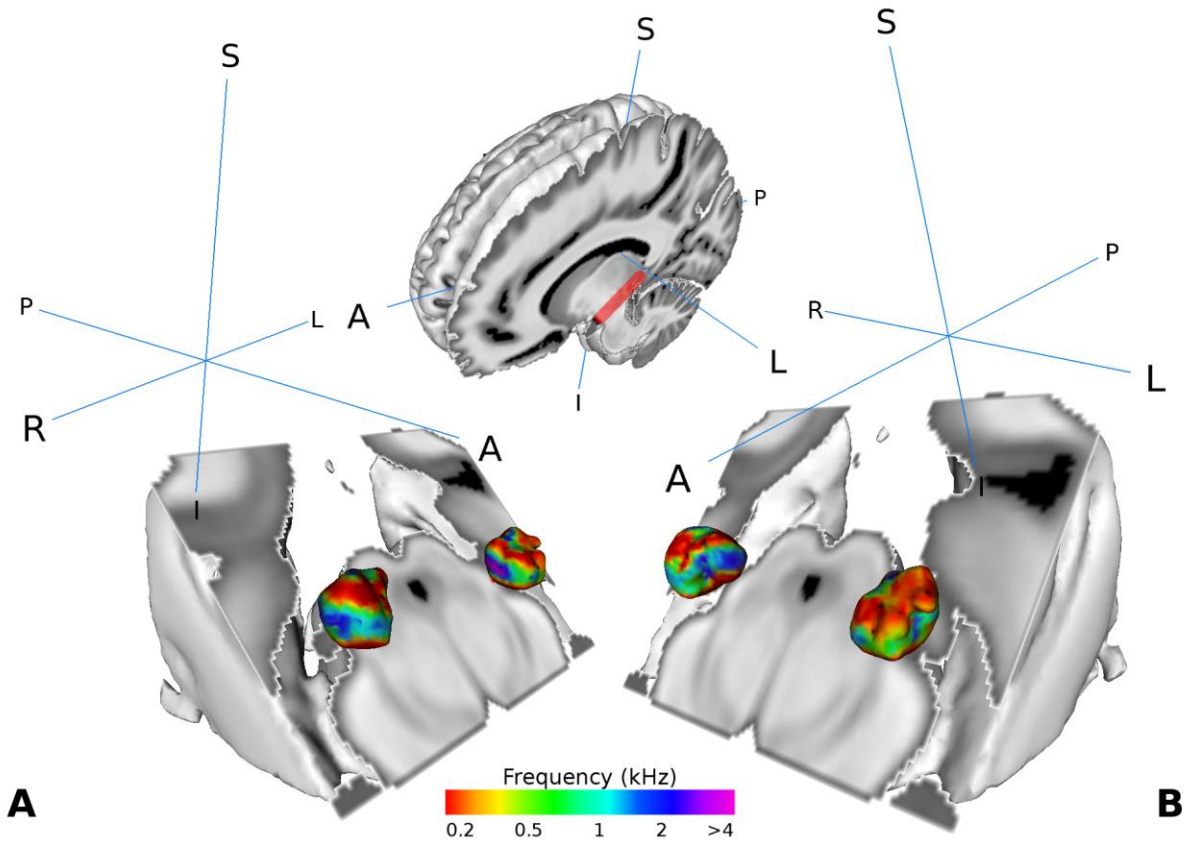

Figure 2. Visualization of the average tonotopy across participants (n = 28) found in the MGB using the tonotopic localizer. The half-brain image at the top shows the orientation (A: anterior, P: posterior, L: left, R: right, S: superior, I: inferior), and a cut through the brain with a red line denoting 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.

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°.

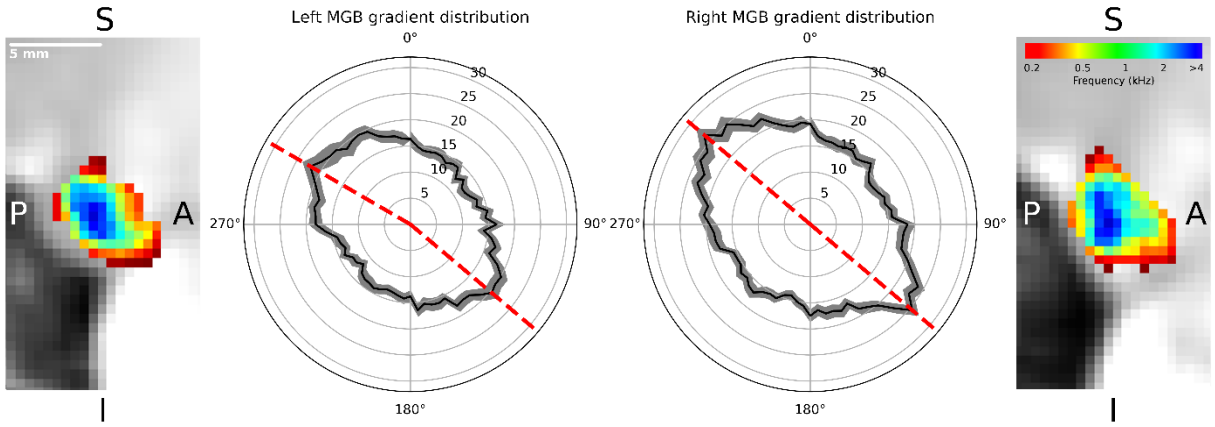

Figure 3. Distribution of gradients in a sagittal plane for ten slices averaged over participants (n = 28). The mean number of angle counts in 5° steps (black line with standard error of the mean in grey, numbers indicate counts) for the left MGB have maxima at 130° and 300° (red dashed lines). For the 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).

Tonotopy localizer - Localisation of vMGB

We used the high frequency components in the middle as a reference to subdivide the MGB volume into two regions per side (Figure 4 A and B). For the left MGB, gradient 1 is located ventrally and slightly medial compared to gradient 2, which is situated more anterior, dorsal, and lateral. For the right MGB we find similar locations: gradient 1 is 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).

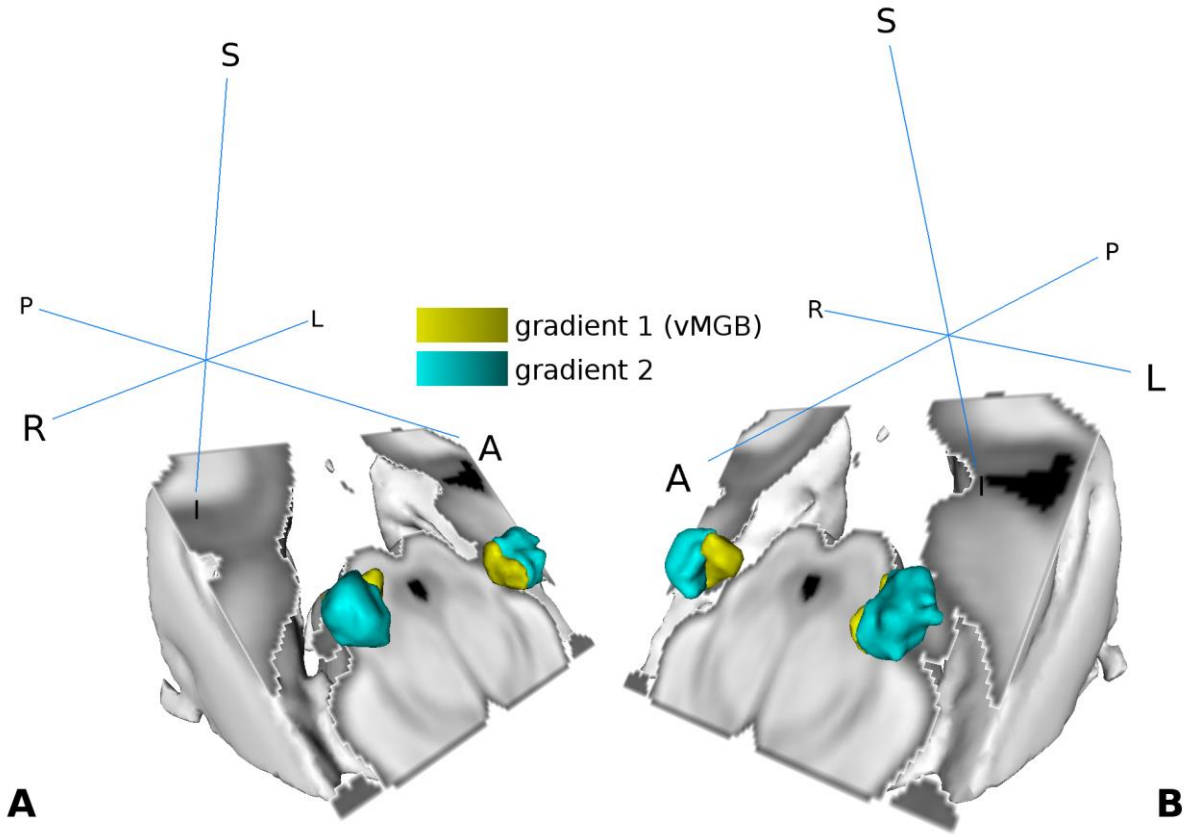

Figure 4. Visualization of the tonotopic gradients found in the MGB based on the tonotopic localizer (see Figure 2). **A.** Three dimensional rendering of the two tonotopic gradients (yellow: ventro-medial gradient 1, interpreted as vMGB, cyan: dorso-lateral gradient 2) in the left and right MGB. **B.** 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

Participants scored a mean hit rate in the speech task of 87.2% with 97% highest posterior density (HPD) interval [82.8%, 91.5%], and mean hit rate in the speaker task was 76.0% with a 97% HPD interval [70.6%, 81.0%]. The mean hit-rate was 11.2% higher in the speech task than in the speaker task with 97% HPD interval [7.6%, 15.0%] (Supplementary Figure 1 A and B).

fMRI

We tested our hypothesis that within the ventral tonotopic gradient (i.e., vMGB) there is a task-dependent modulation. Unexpectedly, we did not observe such a task-dependent modulation, i.e., a higher BOLD responses in the speech task in comparison to the speaker task (*Speech vs Speaker* contrast) in vMGB nor outside this MGB division (i.e., in gradient 2). We proceeded to test our final hypothesis, stating that a task-dependent modulation in the left vMGB correlates with the speech recognition scores 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 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 5 and 6]. In exploratory analyses we tested the correlation between the task-dependent modulation and the mean correct speech recognition scores across participants in the other left MGB subsection (i.e., gradient 2). There was no statistically significant effect [gradient 2: $(-12, -28, -5)$, $p = 0.223$ FWE, $T = 2.24$, β and 90% CI 1.03 [0.22, 1.83]]. In addition, the results were specific to the correlation with the *Speech vs Speaker* contrast. In a control analysis, we tested whether the contrast *Speech vs Speaker* correlated with the mean percent correct *speaker* recognition scores across participants, and found no suprathreshold voxels in vMGB (nor in gradient 2) even at a lenient statistical threshold (uncorrected $p < 0.05$).

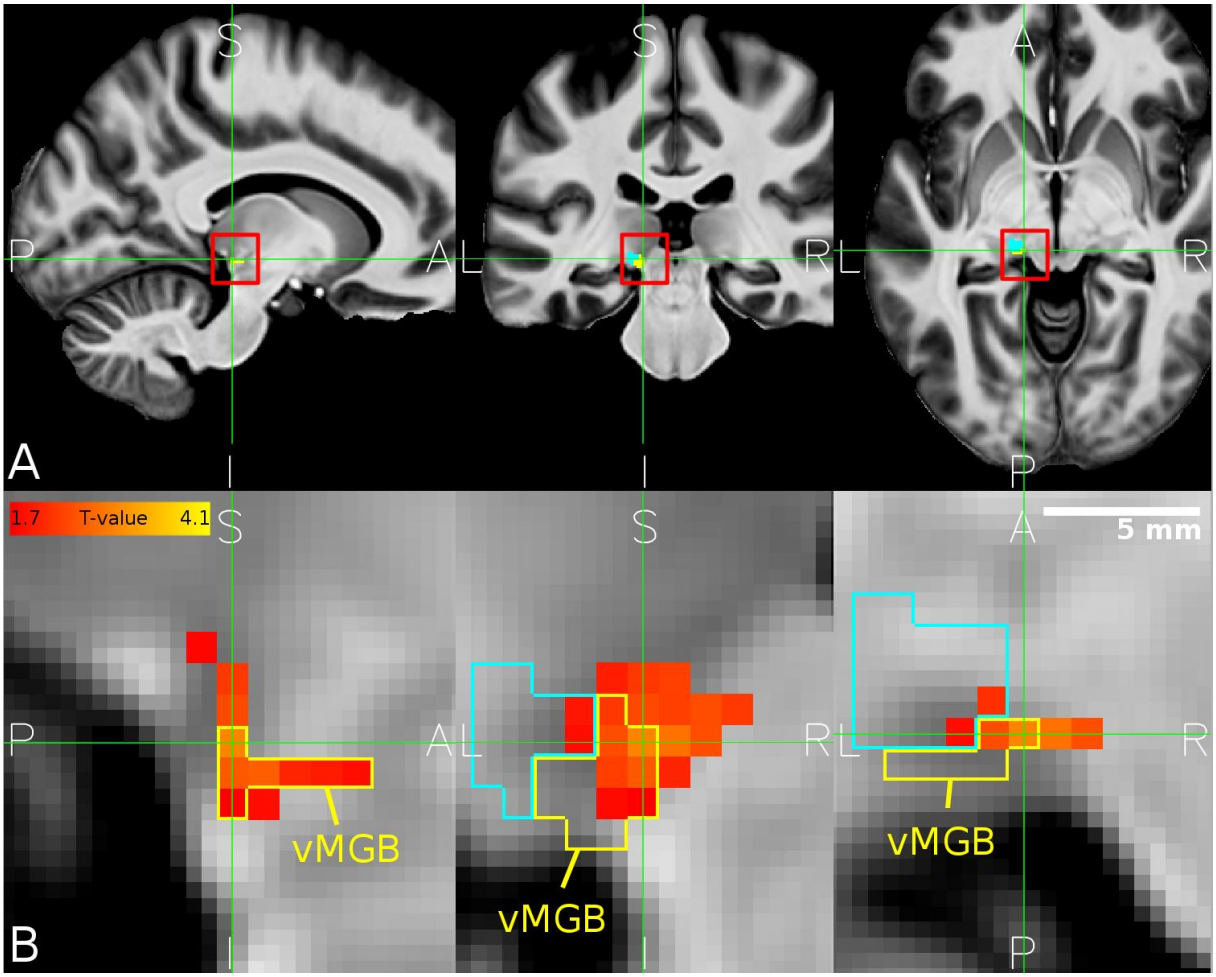

Figure 5. Overlap between MGB divisions and performance-related task-dependent modulation. A. The mean structural image across participants (n = 33) in MNI space. The red squares denote the approximate location of the left MGB and encompass the zoomed in view in B. **B.** Overlap of correlation between *Speech vs Speaker* contrast and the mean percent correct in the speech task (hot colour code) with 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).

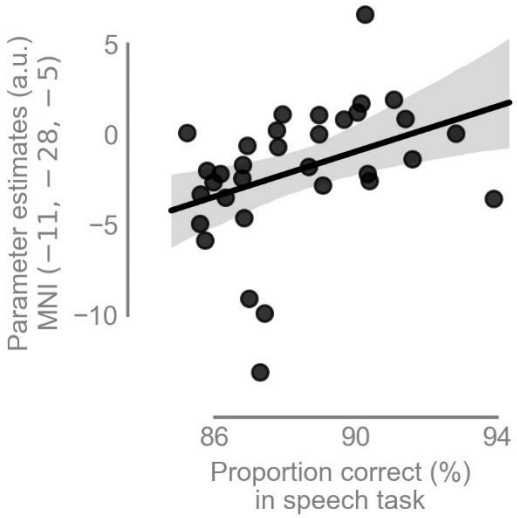

Figure 6. Task dependent modulation of left vMGB correlates with proportion correct responses in the speech task over participants (n = 33): the better the behavioral score in the speech task, the stronger the BOLD response in left vMGB (maximum statistic at MNI coordinate [-11, -28, -5]. The line represents the best fit with 97% bootstrapped confidence interval (gray shaded region).

Discussion

Using ultra-high field fMRI we here showed that the left auditory first-order sensory thalamus – the left ventral subdivision of the MGB (vMGB) – is involved in 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., 2007a, 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).

We localized the vMGB based on its tonotopic organization and location relative to other MGB divisions. The tonotopic organization of the ventral MGB has been observed in many species with the use of invasive techniques (Winer, 1984, Anderson et al., 2007a, 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 categorical Speech vs Speaker contrast in vMGB. The lack of a significant main effect of task (Speech vs Speaker) in the vMGB was surprising, as this categorical task-effect was observed in three previous experiments in participants with typical development (von Kriegstein et al., 2008, Díaz et al., 2012).

There are three potential explanations. 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 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 (Supplementary Figure S2), and in the previous study (Díaz et al., 2018) a correlation between MGB modulation and reading speed and comprehension scores has been found only in participants with developmental dyslexia, but not in neurotypical 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. However, the potential use of such cues did not seem to yield a behavioral benefit, as there was no correlation between the amount of task-dependent modulation and performance in the speaker task across participants.

The localization of the correlation between the speech vs speaker contrast and performance in the speech task to the vMGB confirmed our hypothesis that the left first-order thalamic nucleus – vMGB – is involved in speech recognition. The results can be explained neither by differences in stimulus input in the two conditions, as the same stimuli were heard in both tasks, nor by a correlation with general better task performance, as there was no correlation with the speaker task. They imply that the modulation of speech representations at the level of the primary sensory thalamus are important for speech recognition performance.

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 (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 Ojima and Rouiller, 2011), and alter the response properties of thalamic neurons (Andolina et al., 2007, Cudeiro and Sillito, 2006, Ergenzinger et al., 1998, Ghazanfar and Nicolelis, 2001, Sillito et al., 2006 [review] , Wang et al., 2018, Sillito et al., 1994, Krupa et al., 1999). In speech processing such a mechanism might be especially useful as the signal includes both rapid dynamics, and is predictable (e.g., due to co-articulation or learned statistical regularities in words Saffran, 2003). Furthermore, speech needs to be computed online often under suboptimal listening conditions. Building up accurate predictions within an internal generative model about fast sensory dynamics would result in more efficient processing when the perceptual system is confronted with taxing conditions such as fast stimulus presentation rates or background noise. We speculate that the correlation between speech task performance and task-dependent vMGB modulation might be a result of feedback from cerebral cortex areas. Feedback may emanate directly from auditory primary or association cortices, or indirectly via other structures such as the reticular nucleus with its inhibitory connections to the MGB (Rouiller and de Ribaupierre, 1985). Feedback cortico-thalamic projections from layer 6 in A1 to the vMGB, but also from association cortices

(Tschentscher et al., 2018), may modulate information ascending through the lemniscal pathway, rather than convey information to the ventral division (Lee, 2013, Llano and Sherman, 2001).

Although most of speech and language research focuses on cerebral cortex structures, investigating subcortical sensory contributions to speech perception is paramount to the development of a mechanistic understanding of how the human brain accomplishes speech recognition. The present study brings us a decisive step further in this direction by suggesting that the task-dependent modulation of the ventral subdivision of the medial geniculate body – the primary sensory auditory thalamus – is an important and specific contributor when we want to understand what is said.

Materials and Methods

Participants

The Ethics committee of the Medical Faculty, University of Leipzig, Germany approved the study. 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).

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.

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

Procedure

MGB and Tonotopy localizer. For the MGB localizer and the tonotopy localizer, participants listened to natural sounds (human voices, animal cries, tool sounds; Moerel et al., 2015). The MGB localizer consisted of one run where 84 natural sound stimuli were presented in random order and had a duration of 12:50 minutes. The tonotopy localizer consisted of six runs where 56 of the 84 natural sound stimuli from the MGB localizer were presented. The sounds were randomly chosen before the first run and the same 56 sounds were played in each run. Each run had a duration of 8:58 minutes. To ensure listener engagement, in both localizers the participants performed a 1-back task and pushed a button when two consecutive sounds were the same. This happened on average 5% of the time. Additionally, 5% of the trials contained no sound (null events). Within each run, sounds were randomly jittered at an interval of 2, 3, or 4 repetition times (TR) and presented in the middle of the silent gap of 1200 ms (Figure 1A). The MGB localizer was used as an independent functional identifier for the left and right

MGB. The resulting masks were then used to constrain the analyses of the tonotopy localizer to these regions of interest. In turn, the tonotopic regions of the MGB were used as masks in the speech experiment (see section Functional MRI Data Analysis).

Speech experiment. In the speech experiment (Figure 1C) participants listened to blocks of auditory VCV syllables, and were asked to perform two types of tasks: a speech task and a speaker task. In the speech task, participants reported via button press whether the current syllable was different from the previous one (1-back task). In the speaker task, participants reported via button press whether the current speaker was different from the previous one. Speakers within a block were either all male or all female. This was necessary to avoid that participants performed a gender discrimination task on some trials and a speaker 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 (for details refer to the Participant Training section of the Supplementary Information).

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).

Data Acquisition and Processing

MRI data were acquired using a Siemens Magnetom 7 T scanner (Siemens AG, Erlangen, Germany) with a Nova 32-channel head coil. Functional MRI data were acquired using echo planar imaging (EPI) sequences. We used a field of view (FoV) of 792 mm and partial coverage with 28 slices. This volume was oriented obliquely such that the slices encompassed the inferior colliculi (IC), the MGB and the superior temporal gyrus, running in parallel to the latter (Figure S3).

The MGB and tonotopy localizers had the following acquisition parameters: TR = 2800 ms (acquisition time TA = 1600 ms, silent gap: 1200 ms), TE = 22 ms, flip angle 65°, GRAPPA (Griswold et al., 2002) with acceleration factor 2, 33% phase oversampling, matrix size 120 x 120, FoV 792 mm x 792 mm, phase partial Fourier 6/8, voxel size $(1.1\ mm)^3$, interleaved acquisition, anterior to posterior phase-encode direction. We employed a clustered EPI technique allowing for stimulus presentations in quiet in a fast-event related design. Stimuli were presented during the silent gap. For the MGB localizer we acquired one run of 275 volumes (13:06 minutes). For the tonotopy localizer we acquired 187 volumes (9:01 minutes) per run with a total of six runs.

For the speech experiment, acquisition parameters were the same as for the localizers, with the exception of a shorter TR (1600 ms) due to continuous scanning (i.e., no silent gap), 320 volumes per run, and total length of acquisition per run of 8:30 minutes. Five runs were recorded for each participant. The acquisition parameters were similar to the protocol described by (Moerel et al., 2015), with the exception of a longer echo time and phase oversampling which eschewed front-back wrapping artifacts. A sample EPI is shown in Figure S4. The difference in echo time between our sequence and the one in Moerel et al. (2015) may have resulted in a lower signal-to-noise ratio in subcortical structures. However, as the MGB has a T2* value of ~33 ms the different echo times of 19 ms (our sequence) and 22 ms (Moerel et al., 2015) had little to no effect on the signal-to-noise ratio (Hollander et al., 2017).

During functional MRI data acquisition we also acquired physiological values (heart rate, and respiration rate) using a BIOPAC MP150 system (BIOPAC Systems Inc., Goleta, CA, USA). Structural images were recorded using an MP2RAGE (Marques et al., 2010) T1 protocol: 700 µm isotropic resolution, TE = 2.45ms, TR = 5000 ms, TI1 = 900 ms, TI2 = 2750 ms, flip angle 1 = 5°, flip angle 2 = 3°, FoV 224 mm × 224 mm, GRAPPA acceleration factor 2, duration 10:57 min. More details on the sequence protocols can be found in the Supplementary Information (Section 'Imaging Sequence Protocols').

Behavioral Data Analysis

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

To pool over participants and runs we modeled the correlation between intercepts and slopes in a logistic linear equation. For the model implementation and data analysis, we used PyMC3 (Salvatier et al., 2016) using a No-U-Turn Sampler (Hoffman and Gelman, 2011) with three parallel chains. Per chain we had 20,000 samples with 5,000 of these as warm-up. Only the latter 7,500 were used for posterior mean and highest posterior density (HPD) interval estimates. For details see the Supplementary Information (Section 'Statistical Analysis**Error! Reference source not found.** of Behavioral Data'). The difference in percent correct button presses between the speech and speaker task was calculated using the posterior densities averaged over consonant and vowel changes. The resulting distribution was averaged and the 97% HPD was calculated. If the 97% HPD is very large and/or contains the value zero then we can infer that there is no difference in responses between these two conditions.

Functional MRI Data Analysis

Preprocessing of MRI data

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

A template of all participants was created with ANTs (Avants et al., 2009) using the participants' MP2RAGE images, which was then registered to the MNI space using the same software package and the MNI152 $(0.5\ mm)^3$ voxel size template provided by FSL 5.0.8 (Smith et al., 2004). All MP2RAGE images were preprocessed with Freesurfer (Fischl et al., 2002, Fischl et al., 2004, Han and Fischl, 2007)

using the recon-all command to obtain boundaries between grey and white matter, which were later used in the functional to structural registration step.

The rest of the analysis was coded in nipype (Gorgolewski et al., 2011). A graphical overview of the nipype pipeline can be found in the Supplementary Information (Figure S5). Head motion and susceptibility distortion by movement interaction of functional runs were corrected using the Realign and Unwarp method (Andersson et al., 2001) in SPM 12 after which outlier runs were detected using ArtifactDetect[1] (composite threshold of translation and rotation: 1; intensity Z-threshold: 3; global threshold: 8). Coregistration matrices for realigned functional runs per participant were computed based on each participant's structural image using Freesurfer's BBregister function (register mean EPI image to T1, option `'–init-header'` was specified in order to preserve the origin of the manual alignment of structural and functional data). Warping using coregistration matrices (after conversion to ITK coordinate system) and resampling to 1 mm isovoxel was performed using ANTs. Before model creation we smoothed the data in SPM12 using a 1 mm kernel at full-width half-maximum.

Physiological data

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

Statistical analysis of speech experiment

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

[1] https://www.nitrc.org/projects/artifact_detect/

Speech vs Speaker at the first level using restricted maximum likelihood (REML) as implemented in SPM 12.

After estimation, the contrasts were registered to the MNI structural template of all participants using a two-step registration in ANTs (see also Figure S5). First, a quick registration was performed on the whole head using rigid, affine and diffeomorphic (using Symmetric Normalization: SyN) transformations and the mutual information similarity metric. Second, the high quality registration was 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.

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 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 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; v. 3.6.0; Yushkevich et al., 2006). Volume size and center of mass (COM) for each gradient are listed in Table 1**Error! Reference source not found.**.

Significance testing

We used small volume corrections (SVC) to test for significant voxels for *Speech vs Speaker* as well as the correlation of *Speech vs Speaker* 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 and left vMGB for the second hypothesis motivated by findings in previous studies (von Kriegstein et al., 2008, Díaz et al., 2012).

References

589 ALCÁNTARA, J. I., WEISBLATT, E. J. L., MOORE, B. C. J. & BOLTON, P. F. 2004. Speech-in-noise perception
590 in high-functioning individuals with autism or Asperger's syndrome. *Journal of Child Psychology*
591 *and Psychiatry,* 45**,** 1107-1114.
592 ANDERSON, L. A., CHRISTIANSON, G. B. & LINDEN, J. F. 2009. Stimulus-Specific Adaptation Occurs in the
593 Auditory Thalamus. *Journal of Neuroscience,* 29**,** 7359-7363.
594 ANDERSON, L. A. & LINDEN, J. F. 2011. Physiological differences between histologically defined
595 subdivisions in the mouse auditory thalamus. *Hearing Research,* 274**,** 48-60.
596 ANDERSON, L. A., WALLACE, M. N. & PALMER, A. R. 2007a. Identification of subdivisions in the medial
597 geniculate body of the guinea pig. *Hearing Research,* 228**,** 156-167.
598 ANDERSON, L. A., WALLACE, M. N. & PALMER, A. R. 2007b. Identification of subdivisions in the medial
599 geniculate body of the guinea pig. *Hearing Research,* 228**,** 156-167.
600 ANDERSSON, J. L., HUTTON, C., ASHBURNER, J., TURNER, R. & FRISTON, K. 2001. Modeling geometric
601 deformations in EPI time series. *NeuroImage,* 13**,** 903-19.
602 ANDOLINA, I. M., JONES, H. E., WANG, W. & SILLITO, A. M. 2007. Corticothalamic feedback enhances
603 stimulus response precision in the visual system. *Proceedings of the National Academy of*
604 *Sciences,* 104**,** 1685-1690.
605 ANTUNES, F. M. & MALMIERCA, M. S. 2011. Effect of auditory cortex deactivation on stimulus-specific
606 adaptation in the medial geniculate body. *J Neurosci,* 31**,** 17306-16.
607 AVANTS, B. B., TUSTISON, N. & SONG, G. 2009. Advanced normalization tools (ANTS). *Insight j,* 2**,** 1-35.
608 BARON-COHEN, S., WHEELWRIGHT, S., SKINNER, R., MARTIN, J. & CLUBLEY, E. 2001. The Autism-
609 Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism,
610 Malesand Females, Scientists and Mathematicians. *Journal of Autism and Developmental*
611 *Disorders,* 31**,** 5-17.
612 BARTLETT, E. L., SADAGOPAN, S. & WANG, X. 2011. Fine frequency tuning in monkey auditory cortex and
613 thalamus. *Journal of Neurophysiology,* 106**,** 849-859.
614 BARTLETT, E. L. & WANG, X. 2011. Correlation of neural response properties with auditory thalamus
615 subdivisions in the awake marmoset. *Journal of Neurophysiology,* 105**,** 2647-2667.
616 BENESTY, J., SONDHI, M. M. & HUANG, Y. 2007. *Springer Handbook of Speech Processing*, Springer Berlin
617 Heidelberg.
618 BORDI, F. & LEDOUX, J. E. 1994. Response properties of single units in areas of rat auditory thalamus
619 that project to the amygdala. *Experimental Brain Research,* 98**,** 261-274.
620 CALFORD, M. B. 1983. The parcellation of the medial geniculate body of the cat defined by the auditory
621 response properties of single units. *Journal of Neuroscience,* 3**,** 2350-2364.
622 CAMARILLO, L., LUNA, R., NÁCHER, V. & ROMO, R. 2012. Coding perceptual discrimination in the
623 somatosensory thalamus. *Proceedings of the National Academy of Sciences,* 109**,** 21093-21098.
624 CHANDRASEKARAN, B., HORNICKEL, J., SKOE, E., NICOL, T. & KRAUS, N. 2009. Context-dependent
625 encoding in the human auditory brainstem relates to hearing speech in noise: Implications for
626 developmental dyslexia. *Neuron,* 64**,** 311-319.
627 CHANDRASEKARAN, B., KRAUS, N. & WONG, P. C. M. 2011. Human inferior colliculus activity relates to
628 individual differences in spoken language learning. *Journal of Neurophysiology,* 107**,** 1325-1336.
629 CHI, T., RU, P. & SHAMMA, S. A. 2005. Multiresolution spectrotemporal analysis of complex sounds. *The*
630 *Journal of the Acoustical Society of America,* 118**,** 887-906.
631 CRUIKSHANK, S. J., KILLACKEY, H. P. & METHERATE, R. 2001. Parvalbumin and calbindin are differentially
632 distributed within primary and secondary subregions of the mouse auditory forebrain.
633 *Neuroscience,* 105**,** 553-569.
634 CUDEIRO, J. & SILLITO, A. M. 2006. Looking back: corticothalamic feedback and early visual processing.
635 *Trends in Neurosciences,* 29**,** 298-306.

636 DÍAZ, B., BLANK, H. & VON KRIEGSTEIN, K. 2018. Task-dependent modulation of the visual sensory
637 thalamus assists visual-speech recognition. *NeuroImage,* 178**,** 721-734.
638 DÍAZ, B., HINTZ, F., KIEBEL, S. J. & KRIEGSTEIN, K. V. 2012. Dysfunction of the auditory thalamus in
639 developmental dyslexia. *Proceedings of the National Academy of Sciences,* 109**,** 13841-13846.
640 DUYN, J. H. 2012. The future of ultra-high field MRI and fMRI for study of the human brain. *NeuroImage,*
641 62**,** 1241-1248.
642 ELLIOTT, T. M. & THEUNISSEN, F. E. 2009. The Modulation Transfer Function for Speech Intelligibility.
643 *PLOS Computational Biology,* 5**,** e1000302.
644 ERGENZINGER, E. R., GLASIER, M. M., HAHM, J. O. & PONS, T. P. 1998. Cortically induced thalamic
645 plasticity in the primate somatosensory system. *Nature Neuroscience,* 1**,** 226-229.
646 FISCHL, B., SALAT, D. H., BUSA, E., ALBERT, M., DIETERICH, M., HASELGROVE, C., VAN DER KOUWE, A.,
647 KILLIANY, R., KENNEDY, D., KLAVENESS, S., MONTILLO, A., MAKRIS, N., ROSEN, B. & DALE, A. M.
648 2002. Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the
649 Human Brain. *Neuron,* 33**,** 341-355.
650 FISCHL, B., SALAT, D. H., VAN DER KOUWE, A. J. W., MAKRIS, N., SÉGONNE, F., QUINN, B. T. & DALE, A.
651 M. 2004. Sequence-independent segmentation of magnetic resonance images. *NeuroImage,* 23**,**
652 S69-S84.
653 FISHER, R. A. 1915. Frequency Distribution of the Values of the Correlation Coefficient in Samples from
654 an Indefinitely Large Population. *Biometrika,* 10**,** 507-521.
655 FOSTER, K. H., GASKA, J. P., NAGLER, M. & POLLEN, D. A. 1985. Spatial and temporal frequency
656 selectivity of neurones in visual cortical areas V1 and V2 of the macaque monkey. *The Journal of*
657 *Physiology,* 365**,** 331-363.
658 FRIEDERICI, A. D. & GIERHAN, S. M. 2013. The language network. *Current Opinion in Neurobiology,* 23**,**
659 250-254.
660 FRISTON, K. 2005. A theory of cortical responses. *Philosophical Transactions of the Royal Society of*
661 *London B: Biological Sciences,* 360**,** 815-836.
662 FRISTON, K. & KIEBEL, S. 2009. Predictive coding under the free-energy principle. *Philosophical*
663 *Transactions of the Royal Society of London B: Biological Sciences,* 364**,** 1211-1221.
664 GALABURDA, A. M., MENARD, M. T. & ROSEN, G. D. 1994. Evidence for aberrant auditory anatomy in
665 developmental dyslexia. *Proceedings of the National Academy of Sciences,* 91**,** 8010-8013.
666 GAUDRAIN, E., LI, S., BAN, V. S. & PATTERSON, R. D. 2009. The Role of Glottal Pulse Rate and Vocal Tract
667 Length in the Perception of Speaker Identity. *Interspeech-2009***,** 148-151.
668 GHAZANFAR, A. A. & NICOLELIS, M. A. L. 2001. Feature Article: The Structure and Function of Dynamic
669 Cortical and Thalamic Receptive Fields. *Cerebral Cortex,* 11**,** 183-193.
670 GIRAUD, A.-L., LORENZI, C., ASHBURNER, J., WABLE, J., JOHNSRUDE, I., FRACKOWIAK, R. &
671 KLEINSCHMIDT, A. 2000. Representation of the Temporal Envelope of Sounds in the Human
672 Brain. *Journal of Neurophysiology,* 84**,** 1588-1598.
673 GONZALEZ-LIMA, F. & CADA, A. 1994. Cytochrome oxidase activity in the auditory system of the mouse:
674 A qualitative and quantitative histochemical study. *Neuroscience,* 63**,** 559-578.
675 GORGOLEWSKI, K., BURNS, C. D., MADISON, C., CLARK, D., HALCHENKO, Y. O., WASKOM, M. L. & GHOSH,
676 S. S. 2011. Nipype: a flexible, lightweight and extensible neuroimaging data processing
677 framework in python. *Front Neuroinform,* 5**,** 13.
678 GRISWOLD, M. A., JAKOB, P. M., HEIDEMANN, R. M., NITTKA, M., JELLUS, V., WANG, J., KIEFER, B. &
679 HAASE, A. 2002. Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic*
680 *Resonance in Medicine,* 47**,** 1202-1210.
681 GROEN, W. B., VAN ORSOUW, L., HUURNE, N. T., SWINKELS, S., VAN DER GAAG, R.-J., BUITELAAR, J. K. &
682 ZWIERS, M. P. 2009. Intact Spectral but Abnormal Temporal Processing of Auditory Stimuli in
683 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.

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.

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.

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].

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.

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

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.

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.

SQUIRE, L., BERG, D., BLOOM, F. E., DU LAC, S., GHOSH, A. & SPITZER, N. C. 2012. *Fundamental neuroscience*, Academic Press.

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., DIAZ, B. & VON KRIEGSTEIN, K. 2018. Reduced structural connectivity between left auditory thalamus and the motion-sensitive planum temporale in developmental dyslexia. *arXiv:1811.11658 [q-bio]*.

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.

826 VON KRIEGSTEIN, K., PATTERSON, R. D. & GRIFFITHS, T. D. 2008. Task-dependent modulation of medial
827 geniculate body is behaviorally relevant for speech recognition. *Curr Biol,* 18**,** 1855-9.
828 WANG, W., ANDOLINA, I. M., LU, Y., JONES, H. E. & SILLITO, A. M. 2018. Focal Gain Control of Thalamic
829 Visual Receptive Fields by Layer 6 Corticothalamic Feedback. *Cerebral Cortex,* 28**,** 267-280.
830 WANG, X., LU, T., BENDOR, D. & BARTLETT, E. 2008. Neural coding of temporal information in auditory
831 thalamus and cortex. *Neuroscience,* 154**,** 294-303.
832 WANG, Y., ZHANG, J., ZOU, J., LUO, H. & DING, N. Prior Knowledge Guides Speech Segregation in Human
833 Auditory Cortex. *Cerebral Cortex*.
834 WERNICKE, C. 1874. *Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer*
835 *Basis*, Cohn.
836 WHITE, S., FRITH, U., MILNE, E., ROSEN, S., SWETTENHAM, J. & RAMUS, F. 2006. A double dissociation
837 between sensorimotor impairments and reading disability: A comparison of autistic and dyslexic
838 children. *Cognitive Neuropsychology,* 23**,** 748-761.
839 WINER, J. A. 1984. The human medial geniculate body. *Hearing Research,* 15**,** 225-247.
840 WINER, J. A., KELLY, J. B. & LARUE, D. T. 1999. Neural architecture of the rat medial geniculate body.
841 *Hearing Research,* 130**,** 19-41.
842 WINER, J. A., MILLER, L. M., LEE, C. C. & SCHREINER, C. E. 2005. Auditory thalamocortical transformation:
843 structure and function. *Trends in Neurosciences,* 28**,** 255-263.
844 WINER, J. A. & PRIETO, J. J. 2001. Layer V in cat primary auditory cortex (AI): Cellular architecture and
845 identification of projection neurons. *Journal of Comparative Neurology,* 434**,** 379-412.
846 YUSHKEVICH, P. A., PIVEN, J., HAZLETT, H. C., SMITH, R. G., HO, S., GEE, J. C. & GERIG, G. 2006. User-
847 guided 3D active contour segmentation of anatomical structures: Significantly improved
848 efficiency and reliability. *NeuroImage,* 31**,** 1116-1128.

849