A neural mechanism for recognizing speech spoken by different speakers

Jens Kreitewolf, Etienne Gaudrain, Katharina von Kriegstein

A neural mechanism for recognizing speech spoken by different speakers

Jens Kreitewolf [a,*], Etienne Gaudrain [b,c], Katharina von Kriegstein [a,d]

[a] Max Planck Institute for Human Cognitive and Brain Sciences, Max Planck Research Group Neural Mechanisms of Human Communication, D-04103 Leipzig, Germany

[b] University of Groningen, University Medical Center Groningen, Department of Otorhinolaryngology/Head and Neck Surgery, 9700 RB Groningen, Netherlands

[c] University of Groningen, Graduate School of Medical Sciences, Research School of Behavioural and Cognitive Neurosciences, 9713 GZ Groningen, Netherlands

[d] Humboldt University of Berlin, Psychology Department, D-12489 Berlin, Germany

* Corresponding author at: Max Planck Institute for Human Cognitive and Brain Sciences, MPRG Neural Mechanisms of Human Communication, Stephanstr. 1a, 04103 Leipzig, Germany. Fax: +49 341 9940 2499. E-mail address: kreitewolf@cbs.mpg.de (J. Kreitewolf).

Abstract

Understanding speech from different speakers is a sophisticated process, particularly because the same acoustic parameters convey important information about both the speech message and the person speaking. How the human brain accomplishes speech recognition under such conditions is unknown.

One view is that speaker information is discarded at early processing stages and not used for understanding the speech message. An alternative view is that speaker information is exploited to improve speech recognition. Consistent with the latter view, previous research identified functional interactions between the left- and the right-hemispheric superior temporal sulcus/gyrus, which process speech- and speaker-specific vocal tract parameters, respectively. Vocal tract parameters are one of the two major acoustic features that determine both speaker identity and speech message (phonemes). Here, using functional magnetic resonance imaging (fMRI), we show that a similar interaction exists for glottal fold parameters between the left and right Heschl's gyri. Glottal fold parameters are the other main acoustic feature that determines speaker identity and speech message (linguistic prosody).

The findings suggest that interactions between left- and right-hemispheric areas are specific to the processing of different acoustic features of speech and speaker, and that they represent a general neural mechanism when understanding speech from different speakers.

Keywords: fMRI, Glottal fold, Heschl's gyrus, Linguistic prosody, Voice

Introduction

The same speech message can be acoustically very different depending on who is speaking (e.g., Peterson and Barney, 1952). Nevertheless, the human brain shows remarkable robustness to speaker-related variations despite the fact that the same acoustic parameters convey important information for speech understanding as well as for speaker recognition (reviewed in Obleser and Eisner, 2009; Pisoni, 1997). Glottal pulse rate (GPR) (Figs. 1A/B, green), for instance, which is the result of movements of the glottal folds, signals whether an utterance is a statement or a question (i.e., linguistic prosody) and determines the voice height of a speaker. To date, it is an open question how the human brain accomplishes robust speech recognition under conditions where information about speech and speaker is encoded in the same parameter (like it is the case for GPR).

For many years, neuroscientific research on speech recognition has been performed separately from work on speaker recognition, either implicitly or explicitly assuming that these are two independent processes (reviewed in Belin et al., 2004; Hickok and Poeppel, 2007; Pisoni, 1997; Scott and Johnsrude, 2003). However, several findings from behavioral (reviewed in Cutler et al., 2010; Nusbaum and Magnuson, 1997; Nygaard, 2005) and neuroimaging

studies (e.g., Chandrasekaran et al., 2011; Kaganovich et al., 2006; Wong et al., 2004) showed that there are strong interdependencies between speech and speaker recognition and that even non-speech contexts can shift phoneme categorization (Laing et al., 2012). Recent fMRI work has suggested that speech recognition in the context of changing speakers relies on functional interactions between left- and right-hemispheric areas processing specific acoustic features of speech and speaker (von Kriegstein et al., 2010). In that study, speech stimuli were resynthesized to evoke speaker changes by variations of vocal tract parameters (Fig. 1A, blue), which, similar to glottal fold parameters, affect both the perceived identity of the speaker (Fig. 1B, bottom right) and parts of the speech message (i.e., phonemes in the case of vocal tract parameters) (Fig. 1B, top right) (Gaudrain et al., 2009; Lavner et al., 2000; Smith and Patterson, 2005). However, it remained unclear whether interactions between specific areas in the right and left hemispheres are restricted to vocal tract parameters and to the task of phoneme recognition. Here, we investigated whether such interactions also occur when speakers differ in their glottal fold parameters and during a task that involves recognizing aspects of the speech message that are determined by glottal fold parameters (i.e., linguistic prosody). Finding a similar interaction for speech- and speaker-specific glottal fold parameters would be important since it would suggest that such interactions are not only restricted to one acoustic parameter in speech but represent a general feature of how the brain deals with acoustic speaker variability during speech processing.

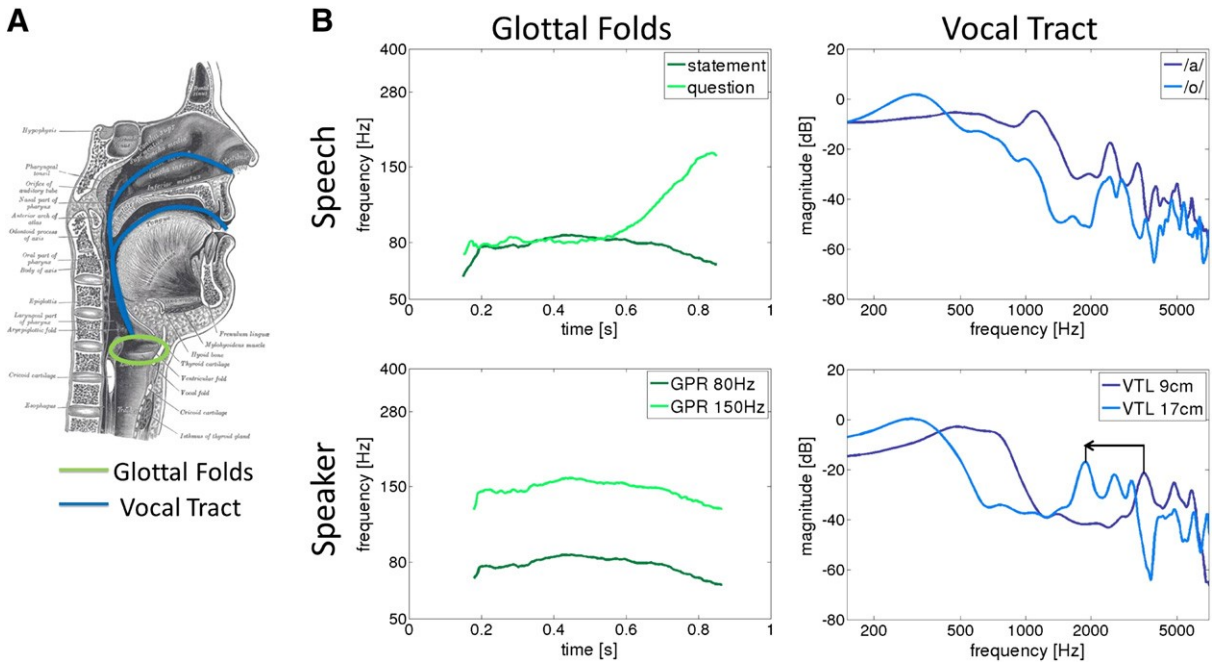


Fig. 1. A. Sagittal section through a human head and neck. Green circle, glottal folds; blue lines, extension of the vocal tract from glottal folds to tip of the nose and lips. B. The plots represent the contribution of glottal fold (left column) and vocal tract parameters (right column) to speech (top row) as well as speaker recognition (bottom row). For glottal fold parameters (left column), frequency is plotted against time on a semi-logarithmic scale. Dynamic variations of glottal pulse rate (GPR) over the course of an utterance determine linguistic prosody (such as whether the speech signal is a question or a statement) (top left). The fundamental frequency (*f*0) contour (i.e., pitch trajectory) of a question is rising at the end of the utterance, whereas the *f*0 contour of a statement is falling. The average GPR over time (bottom left), in contrast, provides information about the voice height (i.e., voice pitch) of the speaker which can be used for speaker recognition (Gaudrain et al., 2009; Lavner et al., 2000). For a higher-pitched voice, the *f*0 contour shifts towards higher frequencies. For vocal tract parameters (right column), magnitude is plotted against frequency; frequency is plotted on a logarithmic scale. Dynamic variations of the vocal tract (i.e., movement of the articulators) determine which speech sound is uttered by producing a different pattern of formants (i.e., peaks) in the spectral envelope (top right). In contrast, the anatomic features of the vocal tract, such as the vocal tract length, determine the timbre of the voice. For a longer vocal tract, formant positions are shifted towards lower frequency values (as indicated by the arrow; bottom right).

We employed an fMRI design in which participants recognized linguistic prosody from speakers who differed only in their average GPR (Fig. 2A; 'prosody task/GPR change'). We used syllables spoken by a single speaker and selectively manipulated their average GPR to induce a perception of speaker change (Gaudrain et al., 2009; Lavner et al., 2000). We will call this 'GPR change' in the following. Furthermore, the syllables were resynthesized with pitch trajectories typical of either question or statement intonation (i.e., with rising or falling pitch) to test

recognition of linguistic prosody. We used sophisticated vocoder software (Kawahara et al., 2008) to ensure that the speaker changes as well as the linguistic prosody was determined by GPR information only, while controlling for all other acoustic parameters. Stimuli were concatenated into sequences of six syllables, and after each syllable sequence, blood oxygen level-dependent (BOLD) responses were measured using fMRI. Participants were asked to report whether or not a presented syllable had a different linguistic prosody than the previous syllable (1-back prosody task); concomitantly, speakers changed in average GPR (GPR change) (Fig. 2A). In this condition, both prosody information and speaker information were encoded by the same anatomically defined acoustic parameter, namely GPR. In order to differentiate between questions and statements in this condition, participants had to disentangle speech- and speaker-specific GPR information; that is, GPR variation over the course of the syllable for prosody and average GPR for speaker identity. As control conditions, the experiment also included syllable sequences in which speaker changes were induced by a manipulation of vocal tract length instead of GPR (VTL change) (Fig. 1B, bottom right; Fig. 2B), and a control task in which participants had to report whether or not a presented syllable was spoken by a different speaker than the previous syllable (1-back speaker task) (Fig. 2). Importantly, the same syllable sequences were presented in the prosody and control tasks. In summary, the experiment had a 2 × 2 factorial design with the factors task (proso dy vs. speaker task) and speaker change (GPR change vs. VTL change). This means that the prosody task was performed while speakers changed in either average GPR (prosody task/GPR change; Fig. 2A, top left) or VTL (prosody task/VTL change; Fig. 2A, top right). The speaker task required to focus on changes in speaker identity that were either solely induced by changes in average GPR (speaker task/GPR change; Fig. 2A, bottom left) or changes in VTL (speaker task/VTL change; Fig. 2A, bottom right). Since the aim of this study was to localize brain regions involved in recognition of GPR-based linguistic prosody from speakers who differ in average GPR, the contrast of interest was defined by the task × speaker change interaction ([(prosody task / GPR change − speaker task / GPR change) − (prosody task / VTL change − speaker task / VTL change)]; Fig. 2A). The rationale behind this procedure was to ensure that the observed BOLD response is specific to the recognition of GPR-based linguistic prosody when speakers differ in average GPR. We employed another type of speaker change (i.e., VTL change) and another task (i.e., speaker task) to control for the possibility that the BOLD response reflects a general activity increase only due to GPR-induced speaker changes or only due to the prosody task.

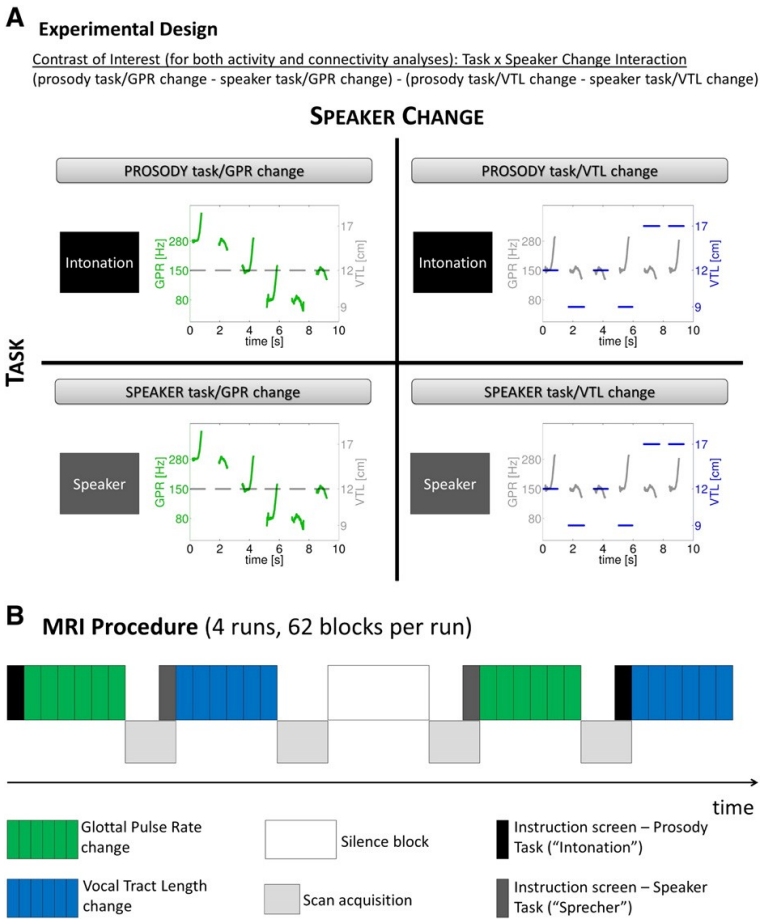


Fig. 2. Experimental design (A) and MRI procedure (B). A. The study was a 2 × 2 factorial design with the factors task [prosody task (top) vs. speaker task (bottom)] and speaker change [GPR change (left) vs. VTL change (right)]. Each of the four cells of the panel shows an example syllable block within the respective condition. At the start of each block, participants saw a task instruction screen indicating that they had to perform either a prosody task ("Intonation") or a speaker task ("Speaker") on the following stimuli. For GPR change, each of three different GPR values (green) is presented twice within a block while VTL is fixed (gray). For VTL change, each of three different VTL values (blue) is presented twice within a block while the speaker's GPR is fixed (gray). The same syllable blocks are presented during the prosody and the speaker task. For the activity analyses, the contrast of interest was the task × speaker change interaction: (prosody task / GPR change − speaker task / GPR change) − (prosody task / VTL change − speaker task / VTL change). This interaction was also used for the psychological variable in the connectivity analyses (see the section "Connectivity analysis (psychophysiological interaction)"). B. The experiment comprised four experimental conditions (prosody task/GPR change, prosody task/VTL change, speaker task/GPR change, speaker task/VTL change) and a silence condition, all of which were presented in blocks (48 blocks/ condition). Each block started with a task instruction screen which was followed by a sequence of six auditory syllables (syllable onsets are represented by black lines within colored boxes). To avoid masking of the auditory stimuli, brain volumes were acquired only after presentation of the auditory stimuli.

We hypothesized that (i) right Heschl's gyrus, which is known to process glottal fold parameters, deals with GPR-induced speaker changes during recognition of linguistic prosody; and that (ii) right Heschl's gyrus is functionally connected to its homologous area in the left hemisphere when participants recognize linguistic prosody from speakers who differ in their glottal fold parameters. These hypotheses were based on two strands of evidence from previous work. First, a previous study (von Kriegstein et al., 2010) showed that right posterior STG/STS deals with speaker changes during speech recognition when both types of information are determined by vocal tract parameters. Additionally, functional connectivity analyses showed that this right posterior STG/STS region interacted with a homologous area in the left posterior STG/STS during this process. Our hypotheses mirror the findings of this previous study with the critical difference that we did not expect an involvement of STG/STS in the processing of GPR-based linguistic prosody and speaker changes but rather an involvement of Heschl's gyrus. This expectation is based on a large body of literature that showed Heschl's gyrus to be involved in processing GPR and other types of pitchevoking stimuli (Griffiths et al., 2010; Kumar et al., 2011; Patterson et al., 2002; Penagos et al., 2004; Puschmann et al., 2010; Wong et al., 2008; for a recent review, see Griffiths and Hall, 2012). In combination with previous results on vocal tract parameters (von Kriegstein et al., 2010), finding that left- and right-hemispheric areas interact with each other when recognizing linguistic prosody from speakers who differ in GPR would provide strong evidence that the brain, in a general fashion, uses such interactions to cope with speaker changes in speech recognition and that such a mechanism might be especially useful when both speech and speaker information are encoded by the same acoustic parameter.

Materials and methods

Participants

Seventeen adults [9 females; mean age 26.1 years; age range 22–34 years; all right-handed as assessed with the Edinburgh questionnaire (Oldfield, 1971)] participated in the study. None of the participants had any history of neurological or psychiatric disorder. None of the participants was trained in a tone language, was a professional musician, or had prior experience with the stimuli used in this study. All participants reported having normal hearing, and they all had normal structural MRI brain scans. Written informed consent was collected from all participants according to procedures approved by the Research Ethics Committee of the University of Leipzig. Participants were paid after completing the experiment.

Stimuli

The stimuli were based on 16 consonant-vowel syllables recorded from one male speaker. The speaker's VTL was estimated to be 15.35 cm (Fitch and Giedd, 1999), and his average GPR (defined as the geometric mean fundamental frequency, *f* 0, across all syllables) was 128.85 Hz. Recordings were made with a cardioid condenser microphone (RØDE NT55, Silverwater, Australia) in a sound-attenuating chamber (IAC-I200 series, Winchester, UK) with a resolution of 16 bits and at a sampling rate of 44.1 kHz. The speaker intonated all of the syllables once in the form of a statement (i.e., with a falling *f* 0 contour) and once in the form of a question (i.e., with a rising *f* 0 contour).

From these original recordings, two syllables (one statement and one question) were selected to serve as templates (see Fig. 1B, top left). Their *f* 0 contours were extracted using vocoder software (Kawahara et al., 2008), and all of the original syllables were resynthesized with new *f* 0 contours derived from these templates. Thus, two new stimuli were created per syllable: a question version with rising *f*0 contour, and a statement version with falling *f*0 contour. This resynthesis ensured that the question and statement versions of each syllable were identical in all aspects apart from their *f*0 contours, and that participants had to rely solely on variations in GPR when differentiating between questions and statements.

Finally, the syllables were resynthesized for a second time to simulate speakers with three different GPR values (80, 150, 280 Hz; i.e., corresponding to steps of 87% increase) and three different VTL values (9, 12, 17 cm; i.e., corresponding to steps of 33% and 45% increase, respectively). These step sizes have been found to be critical for the perception of different speakers rather than changes in the voice characteristics of a single speaker (Gaudrain et al., 2009). Additionally, participants' reports and behavioral results (see the section "Recognizing linguistic prosody is more difficult when speakers differ in GPR than when they differ in VTL") confirmed that the chosen GPR and VTL values were perceived as different speakers. For GPR modifications, *f*0 contours of each syllable were shifted along the frequency axis so that the geometric mean *f*0 in the first 400 ms of that syllable matched the GPR target values. This was done because 400 ms after sound onset marked the time point at which *f*0 contours of question and statement templates diverged. As a result, question and statement syllables started with the same average GPR and then diverged to signal question vs. statement intonation (Fig. 1B, top left).

Experimental design

The experiment was a 2 × 2 factorial design with the factors task (prosody task vs. speaker task) and speaker change (GPR change vs. VTL change) (Fig. 2A). These four experimental conditions and an additional silence condition were presented in blocks of 10 s (Fig. 2B). The order of blocks was randomized with the restriction that a maximum of three repetitions of each condition was allowed. In experimental condition blocks, sequences consisting of six syllables were presented (Fig. 2A). Half of the syllables within each sequence were intonated as questions and the other half were intonated as statements (linguistic prosody). Please note that linguistic prosody varied in all four conditions (Fig. 2A). Each syllable lasted between 955 ms and 1105 ms. The syllables were separated by a brief interstimulus interval (between 562 ms and 712 ms); the total duration of syllable plus interstimulus interval was fixed at 1667 ms. Participants had to respond in the period from 300 ms after the onset of the current syllable to 300 ms after the onset of the following syllable (i.e., within a time window of 1667 ms). The order of syllables was randomized with the restriction that a change from question to statement or vice versa occurred at least twice within each syllable sequence.

Half of the syllable sequences included GPR-induced speaker changes; the other half included VTL-induced speaker changes. In syllable sequences in which GPR changed, each of three GPR values was shown twice while all stimuli had the same VTL value. The fixed VTL value was chosen randomly with the restriction that all VTL values were presented an equal number of times within the experiment. In syllable sequences in which VTL changed, each of the three VTL values was shown twice while all stimuli had the same GPR value. The fixed GPR value was chosen randomly with the restriction that all GPR values were presented an equal number of times within the experiment. Speaker changes in both conditions (GPR change, VTL change) were presented in random order. The changes in speaker and prosody (i.e., question vs. statement intonation) were independent from each other.

Before each sequence, participants received a visual instruction (shown for 1 s) to perform either a prosody task ("Intonation", English: intonation) or a speaker task ("Sprecher", English: speaker). In the prosody task, participants indicated via button press whether or not the prosody of the current syllable was different from the previous one. In the speaker task, participants indicated via button press whether or not the speaker of the current syllable was different from the previous one. The assignment of buttons to same- vs. different-responses was counterbalanced across participants. Each sequence (with a specific stimulus combination) always occurred twice, once in the prosody task and once in the speaker task. Therefore, the exact same sequences were heard during both tasks. Furthermore, the ratio of same-/ different-trials was identical in both tasks.

Data acquisition

MRI data were obtained using a 3 T Siemens Tim Trio MR scanner (Siemens Medical Systems, Erlangen, Germany). The auditory stimuli were delivered using MR-compatible headphones, MR confon OPTIME 1 (MR confon GmbH, Magdeburg, Germany). During the fMRI session, participants wore flat frequency-response earplugs (ER20; Etymotic Research, Inc., Elk Grove Village, IL, USA) to attenuate scanner noise. Before the fMRI session started, sounds were adjusted to a comfortable hearing level for each participant separately. Participants were instructed to make responses with the index and the middle finger of their right hand on a custom-made two-button response box. Task instructions and fixation cross during auditory presentations were delivered using a LCD projector (PLC-XP50L, SANYO, Tokyo, Japan) which could be viewed via a mirror located above the head coil. Presentation of stimuli, recording of participants' responses and synchronization of the experiment with the MR scanner was accomplished using Cogent 2000 (http:// www.vislab.ucl.ac.uk/cogent_2000.php). To avoid masking of the auditory stimuli by scanner noise (Gaab et al., 2007; Hall et al., 1999), the gradient-echo planar images (EPIs) were acquired at the end of each block (42 slices; flip angle, 90°; acquisition bandwidth, 116 kHz; time to echo, 30 ms; 2 mm slice thickness; 1 mm interslice gap; in-plane resolution, 3 × 3 mm; alignment with the anterior commissure–posterior commissure plane; cardiac triggering). Due to cardiac triggering, there was a variable scan repetition time (TR) (mean = 12,486 ms; standard deviation = 365 ms). It has been previously shown (Zhang et al., 2006) that T1 correction can be bypassed by using TRs that are sufficiently long (i.e., in the range of 10 s) to allow for nearly full T1 relaxation and a combination of sparse temporal sampling and cardiac gating is commonly applied without T1 correction (e.g., Backes and van Dijk, 2002; Griffiths et al., 2001; Schönwiesner et al., 2007; von Kriegstein et al., 2007). We, therefore, did not correct for different T1 weighting which would be otherwise indicated for shorter non-constant TRs. The 42 transverse slices of each brain volume were acquired in ascending order and covered the entire brain. For each volume, the task instruction was presented during the fifteen slice acquisitions (i.e., 1 s) immediately before the auditory stimulation started. The experiment included 248 brain volumes for each participant (4 runs of 62 volumes each). Each run lasted approximately 12 min. The first two volumes of each run were discarded. Thus, there were 48 volumes for each of the four experimental conditions plus 48 volumes for the silence condition.

Geometric distortions were characterized by a B0 field-map scan. The field-map scan consisted of a gradient-echo readout (24 echoes, interecho time 0.95 ms) with a standard 2D phase encoding. The B0 field was obtained by a linear fit to the unwrapped phases of all odd echoes. Structural images were acquired with a T1-weighted 3D MP-RAGE (magnetization-prepared rapid gradient echo) sequence with selective water excitation and linear phase encoding. Magnetization preparation consists of a non-selective inversion pulse. The imaging parameters were: TI = 650 ms; repetition time of the total sequence cycle, TR = 1300 ms; repetition time of the gradient-echo kernel (snapshot FLASH), TR,A = 10 ms; TE = 3.93 ms; alpha = 10°; bandwidth = 130 Hz/pixel (i.e., 67 kHz total); image matrix = 256 × 240; FOV = 256 mm × 240 mm; slab thickness = 192 mm; 128 partitions; 95% slice

resolution; sagittal orientation; spatial resolution =1 mm × 1 mm × 1.5 mm; 2 acquisitions. To avoid aliasing, oversampling was performed in the read direction (head–foot).

Data analysis

The behavioral data were analyzed using MATLAB (version 7.11, MathWorks, USA) for descriptive statistics and PASW Statistics 18.0 (SPSS Inc., Chicago, IL, USA) for inferential statistics. Imaging data were analyzed with the statistical parametric mapping package (SPM8; Wellcome Trust Centre for Neuroimaging; http://www.fil.ion.ucl.ac.uk/spm/). Standard spatial preprocessing procedures were used [realignment and unwarp, normalization to MNI standard stereotactic space using the T1 scan of each participant, smoothing with an isotropic Gaussian filter of 8 mm at FWHM, and high-pass filtering at 128 s (Friston et al., 2007)]. Geometric distortions due to susceptibility gradients were corrected by an interpolation procedure based on the B0 field-map.

Activity analysis

Statistical parametric maps were generated by modeling the evoked hemodynamic response for the different conditions as boxcars convolved with a synthetic hemodynamic response function in the context of the general linear model (Friston et al., 2007). Population-level inferences concerning BOLD signal changes between conditions of interest were based on a random-effects model that estimated the secondlevel *t* statistic at each voxel.

To test our first hypothesis, we investigated the BOLD response elicited by the task × speaker change interaction in the following direction: (prosody task / GPR change − speaker task / GPR change) − (prosody task / VTL change − speaker task / VTL change). This ensured that the observed BOLD response is specific to the recognition of GPRbased linguistic prosody when speakers differ in average GPR and does not simply reflect a general activity increase due to any kind of speaker change during the prosody task or any kind of task demand during GPR-induced speaker changes. To descriptively show the direction of the interaction (for a discussion, see Poldrack and Mumford, 2010), we extracted and plotted the parameter estimates at the voxel that showed the statistical maximum in the interaction contrast. This was done by contrasting each experimental condition against the silence condition and extracting the first eigenvariate in each of these contrasts at the statistical maximum of the interaction. To rule out the possibility that brain activation was due to differences in performance level, we performed a second SPM analysis in which behavioral scores of the interaction contrast were entered into the second-level analysis as a covariate. The behavioral scores were calculated using the performance (i.e., proportion correct) of each of the participants in the respective conditions: (prosody task / GPR change − speaker task / GPR change) − (prosody task / VTL change − speaker task / VTL change).

For completeness, we also investigated BOLD responses associated with the main effects of prosody (prosody task – speaker task) and speaker task (speaker task – prosody task) as well as the simple main effects of prosody task for GPR change (prosody task / GPR change – speaker task / GPR change) and VTL change (prosody task / VTL change – speaker task / VTL change) and the conjunction of both contrasts (prosody task / GPR change – speaker task / GPR change ∩ prosody task / VTL change – speaker task / VTL change). Furthermore, we also performed the interaction contrast in the opposite direction [(prosody task / VTL change – speaker task / / VTL change) – (prosody task / GPR change – speaker / GPR change)]. For all analyses, we entered behavioral scores into the second-level analysis as a covariate using the proportion correct values of each participant in the respective conditions.

Connectivity analysis (psychophysiological interaction)

A PPI analysis (Friston et al., 1997) served to address our second hypothesis. As the seed region, we selected the area in right Heschl's gyrus that showed the maximum statistic in the interaction contrast: (prosody task / GPR change – speaker task / GPR change) – (prosody task / VTL change – speaker task / VTL change). For each participant, the first eigenvariate from a sphere with a radius of 4 mm around the group mean was extracted (PPI-seed region). PPI regressors were created using routines implemented in SPM8. The psychological variable was the interaction contrast in the following direction: (prosody task / GPR change – speaker task / GPR change) – (prosody task / VTL change – speaker task / VTL change) (Fig. 2A). PPI regressor, psychological variable and first eigenvariate were entered in a design matrix at the single-subject level. Population-level inferences about

BOLD signal changes were based on a random-effects model that estimated the second-level statistic at each voxel using a one-sample *t* test. To rule out the possibility that the connectivity results were influenced by differences in performance level, we performed a second PPI analysis. As in the activity analysis, behavioral scores of the interaction contrast (i.e., the psychological variable in the PPI) were entered into the second-level analysis as a covariate.

Significance thresholds for fMRI data

Responses were considered significant if they were present at $p < 0.05$ familywise error (FWE) corrected for the region of interest or at $p < 0.05$ FWE-corrected at whole-brain level. The regions of interest for the interaction contrast in the activity analysis were the three anatomical divisions of Heschl's gyrus (TE1.1, TE1.0, and TE1.2; (Morosan et al., 2001) in the right hemisphere. The regions of interest for the connectivity analysis were the homologous regions in the left hemisphere. To locate the divisions of Heschl's gyrus, we used the anatomical templates provided by the anatomy toolbox in SPM. In the text, we only refer to activations that conform to the significance criteria. For completeness and as an overview for interested readers, we additionally display all results at $p < 0.001$ uncorrected and a minimum cluster size of 5 voxels in the tables. For display purposes only, the activation and connectivity patterns in Figs. 3 and 4 are shown at $p < 0.005$ uncorrected.

Results

Recognizing linguistic prosody is more difficult when speakers differ in GPR than when they differ in VTL

The mean performance in the four experimental conditions was 76.22% correct (prosody task/GPR change), 84.44% correct (speaker task/GPR change), 79.00% correct (prosody task/VTL change), and 81.30% correct (speaker task/VTL change). A two-way repeated measures analysis of variance (ANOVA) on mean performance with the factors task (prosody task vs. speaker task) and speaker change (GPR change vs. VTL change) revealed that the main effects of task ($F_{(1,16)} = 3.55$; $p = 0.078$) and speaker change ($F_{(1,16)} = 0.06$; $p = 0.81$) were not significant. However, there was a significant interaction between the two factors ($F_{(1,16)} = 12.44$; $p = 0.003$). To investigate the cause of this behavioral interaction, we performed paired-samples t-tests comparing the performance in prosody and speaker task for GPR change and VTL change separately. This revealed that the prosody task was significantly more difficult than the speaker task when GPR varied ($t(16) = 2.75$; $p = 0.014$), but not when VTL varied ($t(16) = 0.81$; $p = 0.43$). Finally, the behavioral results demonstrated that mean performance in the speaker task was significantly above chance level (50%) for both GPR change ($t(16) = 14.76$; $p < 0.001$) and VTL change ($t(16) = 16.58$; $p < 0.001$). This indicates that participants perceived the chosen GPR and VTL values as different speakers. We refrained from analyzing response times since the critical time windows for solving the prosody and speaker task differed and, therefore, did not allow for a fair comparison between tasks: the prosody task could only be solved at the beginning of the *f*0 drift around 400 ms after sound onset (see the section "Stimuli"), while relevant information to solve the speaker task was already available at sound onset.

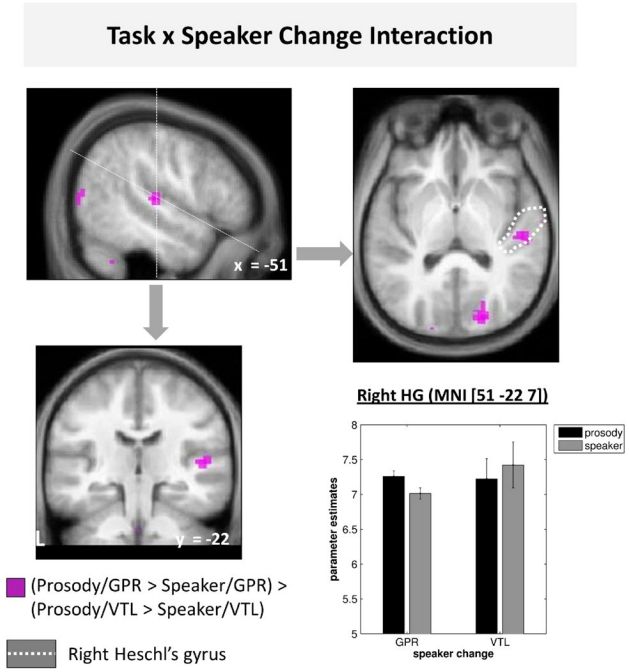


Fig. 3. BOLD response associated with the interaction between task and speaker change (prosody task / GPR change − speaker task / GPR change) − (prosody task / VTL change − speaker task / VTL change) (magenta). The marked area in the horizontal section (top right) contains right Heschl's gyrus (HG). For display purposes only, BOLD responses are shown at $p < 0.005$ uncorrected. The posterior and cerebellar activation patterns did not survive FWE correction (cf. Table 1). Right Heschl's gyrus was the only area showing BOLD responses that conformed to the significance criteria (see the section "Significance thresholds for fMRI data"). The plot shows the parameter estimates at the MNI coordinate of x = 51, y = −22, and z = 7 (i.e., the statistical maximum of the interaction). Bars represent means across participants, error bars represent 95%-CIs (Morey, 2008).

Right Heschl's gyrus is modulated by the recognition of linguistic prosody from speakers who differ in glottal fold parameters

To test our first hypothesis – that right Heschl's gyrus is involved in recognizing linguistic prosody when concurrently dealing with GPR-induced speaker changes – we tested the following interaction: (prosody task / GPR change − speaker task / GPR change) − (prosody task / VTL change − speaker task/VTL change). In accordance with the hypothesis, we found enhanced BOLD responses in right Heschl's gyrus (Fig. 3; MNI coordinates: x = 51, y = −22, z = 7; Z = 3.42). Small volume analyses showed an involvement of posteromedial (TE1.1) and central (TE1.0) portions of Heschl's gyrus (FWE-corrected $p = 0.041$ for TE1.1, and $p = 0.03$ for TE1.0). Anterolateral Heschl's (TE1.2) was not significantly activated in the interaction contrast (FWE-corrected $p = 0.29$). No other area showed BOLD responses that conformed to the significance criteria. To check whether this response is specific to the right hemisphere, we tested whether a similar small volume correction would produce significant results in any subdivision of left Heschl's gyrus. None of the subdivisions of left Heschl's gyrus showed a significant activation in this analysis even at a lenient threshold ($p = 0.32$ for TE1.0; $p = 0.58$ for TE1.1; $p = 0.59$ for TE1.2; all FWE corrected). For information purposes only, Table 1 lists areas that showed BOLD responses at $p < 0.001$ uncorrected. The plot in Fig. 3 descriptively shows that the interaction at the statistical maximum within right Heschl's gyrus was in the hypothesized direction. The interaction contrast performed in the opposite direction [i.e., (prosody task / VTL change − speaker task / VTL change) − (prosody task / GPR change − speaker / GPR change)] did not yield any significant effects. The results cannot be explained by differential degrees of task difficulty in the respective conditions: a separate analysis in which behavioral performance was added as a covariate to the interaction contrast still showed significant BOLD responses in posteromedial and central portions of right Heschl's gyrus (FWEcorrected $p = 0.049$ for TE1.1, and $p = 0.039$ for TE1.0; right TE1.2 was not significant, FWE-corrected $p = 0.31$). Moreover, the results cannot be explained by stimulus differences because the same stimuli were presented in the two tasks.

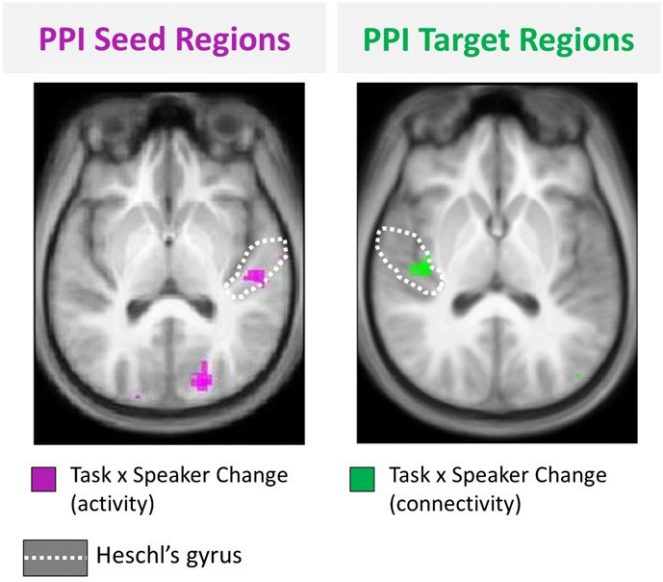


Fig. 4. Functional connectivity (PPI) between right and left Heschl's gyri. The target region identified by the PPI analysis (task × speaker change, connectivity) is shown in green [MNI coordinates: x = −39, y = −22, z = 7; Z = 3.05]. The seed region in right Heschl's gyrus is displayed in magenta as the group mean. For each participant, the eigenvariate was extracted at the location of the group mean with a sphere of 4 mm radius. The marked areas contain left and right Heschl's gyri. For display purposes only, the activation and connectivity patterns are shown at $p < 0.005$ uncorrected. Apart from left Heschl's gyrus no other area showed connectivity to the seed region that conformed to the significance criteria (see the section "Significance thresholds for fMRI data").

Functional connectivity between the right and the left Heschl's gyrus is increased during processing of linguistic prosody from speakers who differ in glottal fold parameters

To test our second hypothesis – that the area in right Heschl's gyrus that deals with GPR-induced changes during recognition of linguistic prosody is functionally connected to a homologous area in the left hemisphere – we performed a PPI analysis. The analysis revealed that activity in right Heschl's gyrus (seed region; Fig. 4 magenta) has a stronger functional connection to left Heschl's gyrus (target region; Fig. 4 green) during the prosody task when speakers differ in GPR than when they differ in VTL (prosody task / GPR change − speaker task / GPR change) − (prosody task / VTL change − speaker task/VTL change). The connectivity analysis showed that it is the posteromedial part of left Heschl's gyrus (TE1.1) (FWE-corrected $p = 0.049$) which is functionally connected to the seed region located in right TE1.1 and right TE1.0. Left TE1.0 and TE1.2 were not significantly connected to the seed region (FWE-corrected $p = 0.31$ for left TE1.0 and $p = 0.42$ for left TE1.2). Since we used the interaction contrast as the psychological variable, this means that there is a specific connectivity increase to linguistic prosody recognition in the context of speakers who change in GPR, but not if speakers change in VTL or if there is another task context (i.e., speaker recognition). There were no other areas significantly connected to the seed region in the right Heschl's gyrus. Clusters present at $p < 0.001$ uncorrected are displayed in Table 2 for information purposes only. The functional connectivity between right and left Heschl's gyri cannot be explained by differential degrees of task difficulty: a separate analysis in which behavioral performance was added as a covariate confirmed that the posteromedial part of left Heschl's gyrus (TE1.1) is functionally connected to the seed region in the right Heschl's gyrus when recognizing linguistic prosody from speakers who differ in GPR (FWE-corrected $p = 0.048$). Again, no significant connectivity was found between the seed region in the right Heschl's gyrus and left TE1.0 and TE1.2 (FWE-corrected $p = 0.27$ for left TE1.0 and FWE-corrected $p = 0.45$ for left TE1.2).

Main effects of prosody and speaker recognition

For completeness, we also analyzed the activity main effects of prosody and speaker tasks (i.e., the contrasts of prosody task – speaker task, and speaker task – prosody task). We also computed the simple main effects of prosody task when speakers changed in GPR (prosody task / GPR change – speaker task / GPR change) and in VTL (prosody task / VTL change – speaker task / VTL change) as well as the conjunction of both contrasts (prosody task / GPR change – speaker task / GPR change ∩ prosody task / VTL change – speaker task / VTL change). Recognition of linguistic prosody as compared to speaker recognition was associated with activity in frontal and parietal regions irrespective of whether speakers changed in glottal fold or vocal tract parameters (Suppl. Fig. 1). Additionally, cerebellar and thalamic regions responded more strongly to the prosody than to the

speaker task ($p < 0.05$ FWE corrected, Inline Supplementary Table S1). There were no enhanced BOLD responses to prosody as compared to speaker recognition in Heschl's gyrus even at a very lenient significance threshold ($p < 0.05$ uncorrected). Speaker recognition in contrast to recognition of linguistic prosody elicited significantly stronger activation of hippocampal, frontal, and parietal areas ($p < 0.05$ FWE corrected, Inline Supplementary Table S2).

Table 1 — **Activity results.** This table presents the activity results as revealed by the interaction contrast: (prosody task / GPR change speaker task / GPR change) (prosody task / VTL change speaker task / VTL change). In this analysis, behavioral performance was added as a covariate. In addition to results within right Heschl's gyrus that conformed to the significance criteria (see the section "Significance thresholds for fMRI data"), the table shows activity in areas that was present at $p < 0.001$ uncorrected with a cluster threshold of at least 5 voxels for information purposes only. Anatomical labels were based on the anatomy toolbox in SPM for the ROI analysis and the Harvard-Oxford cortical structural atlas in FSLview for the whole-brain analysis.

ROI analysis ($p < 0.05$ FWE-corrected for ROI)					
Region	t-Value	p-Value	MNI coordinates		
		(FWE corrected)	x	y	z
Right central Heschl's (TE1.0)	3.87	0.039	51	−19	7
Right posteromedial Heschl's (TE1.1)	3.61	0.049	48	−22	7
Whole-brain analysis ($p < 0.001$ uncorrected and $k \geq 5$)					
	k	t-Value	p-Value	MNI coordinates	
			(Uncorrected)	x	y	z
Right cuneus	25	4.49	0.0002	18	−79	34

Table 2 — **Connectivity results.** This table presents the connectivity results as revealed by the PPI analysis with right Heschl's gyrus as the seed region and the interaction contrast (prosody task / GPR change − speaker task / GPR change) − (prosody task / VTL change − speaker task / VTL change) as the psychological variable. In this analysis, behavioral performance was added as a covariate. In addition to results within left Heschl's gyrus that conformed to the significance criteria (see the section "Significance thresholds for fMRI data"), the table shows functional connectivity to areas that was present at $p < 0.001$ uncorrected with a cluster threshold of 5 voxels for information purposes only. Anatomical labels were based on the anatomy toolbox in SPM for the ROI analysis and the Harvard-Oxford cortical structural atlas in FSLview for the whole-brain analysis.

ROI analysis ($p < 0.05$ FWE-corrected for ROI)						
Region	t-Value		p-Value	MNI coordinates		
			(FWE corrected)	x	y	z
Left posteromedial Heschl's (TE1.1)	3.66		0.049	−39	−22	7
Whole-brain analysis ($p < 0.001$ uncorrected and $k \geq 5$)						
	k	t-Value	p-Value	MNI coordinates		
				x	y	z
Right central Heschl's (TE1.0)	24	6.26	0.000008	42	−16	10
Right lingual gyrus	28	5.28	0.00005	15	−64	−5
Left fusiform gyrus	5	5.06	0.00007	−36	−49	−8
Left cerebellum	9	4.78	0.0001	−3	−55	−29

Inline Supplementary Tables S1 and S2 can be found at the end of this manuscript.

Discussion

The present study showed that regions in right Heschl's gyrus are specifically activated when participants recognized linguistic prosody in conditions where prosody information and speaker identity information were intermingled in the same acoustic parameter, namely GPR. Additionally, there was an increased functional connectivity between right and left Heschl's gyri during recognition of linguistic prosody when speakers differ in glottal fold parameters. At the behavioral level, recognizing linguistic prosody was more difficult when speakers differ in glottal fold parameters in contrast to when speakers differ in vocal tract parameters. These findings provide strong evidence for interdependencies between speech- and speaker-specific processes and suggest that robust speech recognition in the context of changing speakers relies on a neural mechanism that exploits interactions between areas in the left and right hemispheres that process acoustic features of speech and speaker. The findings are unexpected under the view that speech and speaker parameters are processed independently, but they are in full agreement with the alternative view that processing of speech and speaker parameters interact on the neural and behavioral level (von Kriegstein et al., 2010; reviewed in, Nusbaum and Magnuson, 1997; Nygaard, 2005).

The fMRI findings confirm our hypotheses that (i) right Heschl's gyrus is involved in the recognition of linguistic prosody when speakers differ in glottal fold parameters and that (ii) right and left Heschl's gyri are functionally connected when recognizing linguistic prosody from speakers who differ in GPR. Both the activity and connectivity findings for GPR processing in Heschl's gyrus exactly mirrored the findings for vocal tract parameters in STG/STS of a previous study (von Kriegstein et al., 2010). The findings of the present study were located in the central and medial portions (TE1.0 and TE1.1) of Heschl's gyrus. Although pitch processing is often associated with anterolateral Heschl's gyrus (Griffiths et al., 2001; Gutschalk et al., 2004; Patterson et al., 2002; Penagos et al., 2004; Puschmann et al., 2010; von Kriegstein et al., 2010), recent studies suggest that medial parts are also involved in the processing of pitch information and that pitch analysis might critically rely on interactions between medial and lateral Heschl's regions (Griffiths et al., 2010; Kumar et al., 2011; Wong et al., 2008; for a recent review, see Griffiths and Hall, 2012).

There are findings showing that task demands rather than stimulus features can affect the lateralization of prosody processing (e.g., Luks et al., 1998; Plante et al., 2002) and that working memory load can lead to activation of auditory cortex (Brechmann et al., 2007). In our study, these potential factors were controlled for and are unlikely to account for our fMRI findings. Indeed, both recognition of prosody and speaker were tested using 1-back tasks on exactly the same stimulus material; contrasting both tasks therefore controls for working memory processes related to 1-back tasks. Furthermore, by using an interaction analysis of a 2×2 factorial design we can control for lateralization due to the prosody task alone; right Heschl's gyrus is specifically involved when the prosody task interacts with GPR speaker changes, but it is not generally involved in prosody as compared to speaker recognition (Supp. Fig. 1). To control for performance level differences, we modeled participants' performance as covariates of no interest in both the activity and connectivity analyses. Note that the results of the analyses with and without covariates were qualitatively the same, indicating that behavioral differences did not have much influence on the activity levels in Heschl's gyrus. Furthermore, although there was a trend to significance for the main effect of task in the behavioral results, there was no activation in Heschl's gyrus for the main effect of task even at an extremely lenient ($p < 0.05$ uncorrected) threshold. Some activation in the main effect of task would have been expected if Heschl's gyrus activity is modulated by task difficulty. Additionally, previous studies that were designed to investigate task difficulty in pitch discrimination tasks, did not find activation in Heschl's gyrus associated with high task difficulty levels. One study showed that increasing task difficulty modulates activity in parietal and insular regions but not auditory cortex (Rinne et al., 2009). Another study reported that an area in right STG is rather negatively correlated with increased task difficulty during pitch discrimination (Reiterer et al., 2005).

The behavioral results of the present study are consistent with previous findings showing that speech recognition is influenced by the acoustic characteristics of speakers. A number of behavioral findings suggested that attributes of a speaker's voice are perceived and memorized along with the speech message (Bradlow et al., 1999; Palmeri et al., 1993; Pisoni, 1993) and that knowledge about the vocal characteristics of a speaker enhances comprehension of what is said (Best et al., 2008; Bradlow and Pisoni, 1999; Kitterick et al., 2010; Levi et al., 2011; Nygaard and Pisoni, 1998; Nygaard et al., 1994; Remez et al., 2009; Yonan and Sommers, 2000). Conversely, understanding

speech in the context of speaker changes typically results in performance costs and longer processing times (e.g., Magnuson and Nusbaum, 2007). The present behavioral results extend these findings in two ways. First, they showed that performance costs do not only occur during recognition of the speech message (e.g., words) but also when recognizing linguistic prosody (e.g., whether a sentence is intonated as a question or as a statement). Furthermore, our results showed greatest performance costs when recognizing linguistic prosody from speakers who differ in glottal fold parameters which implies that processing speech from different speakers is particularly difficult when the two types of information are carried by the same acoustic parameter.

Our findings highlight interactions between processes necessary for both the analyses of speech and speaker information. This is in accordance with recent neuroimaging work suggesting an integrated processing of speech and speaker information (Chandrasekaran et al., 2011; Kaganovich et al., 2006; Schweinberger et al., 2011). However, it remained so far unclear what kind of mechanisms explains this integration. Based on the present and previous fMRI results (von Kriegstein et al., 2010), we propose a potential mechanism of how the human brain processes auditory speech information when the same acoustic parameters also signal changes in speaker identity (Fig. 5). According to this view, right-hemispheric areas are specifically involved in extracting speaker-related acoustic features and communicate this information to homologous areas in the left hemisphere. Knowledge about the vocal characteristics of a speaker might place constraints on the linguistic analysis of the signal which might be especially helpful in situations where information necessary for both speech and speaker recognition is carried by the same acoustic parameter. For instance, knowledge about the relatively constant GPR specific to a speaker's voice might help in distinguishing between questions and statements from that speaker. This view is in accordance with theoretical considerations suggesting that the human brain uses a mechanism in which typically slow-varying information predicts more rapidly changing information to optimize recognition (Balaguer-Ballester et al., 2009; Kiebel et al., 2008). In our case, slow-varying information refers to speaker changes which occur between consecutive syllables in the present study and on a much slower time-scale in real-life conversations. Prosody information – determined by pitch variations within a syllable – changes more rapidly. A predictive-coding account in which slow-varying speaker parameters are informative for more rapidly changing speech information could explain behavioral findings which demonstrate benefits in speech processing from knowledge about the speaker's voice (reviewed in Nygaard, 2005). The fMRI findings suggest that interacting areas in the left and right hemispheres are specific to the acoustic parameter. For VTL, right posterior STG/STS regions deal with VTL-induced speaker changes when recognizing speech and functionally interact with left STG/STS (von Kriegstein et al., 2010). For GPR, right Heschl's gyrus deals with GPR-induced speaker changes when recognizing linguistic prosody and functionally interacts with left Heschl's gyrus during that process. PPI analyses rely on correlations of activity (Friston et al., 1997) and, thus, do not allow drawing conclusions on directionality. However, the results are in accordance with the speculation that an advantage of knowing the relatively constant GPR specific to a speaker's voice is represented by right-hemispheric areas providing this information to homologous areas in the left hemisphere.

Interactions between right and left Heschl's gyri might not be re- stricted to the analysis of speech- and speaker-specific GPR information but they could as well represent a general feature in the neural processing of dynamic pitch variations while dealing with concomitant variations in average pitch. It has been shown that areas in right auditory cortex – in close proximity to Heschl's gyrus – are involved in detecting pitch direction from frequency modulated tones (Brechmann and Scheich, 2005). Furthermore, we speculate that interactions between right and left Heschl's gyri might also support functions in other auditory domains. In music perception, for example, such interactions could be used to disentangle local melody-related pitch changes from more global pitch variations indicating a change in key. Recent behavioral findings showed that speaker-related GPR information also affects the perception of vowel quality (i.e., speech-related vocal tract information) (Barreda and Nearey, 2012). We speculate that at the neural level, such interactions might be associated with an increased functional coupling of right Heschl's gyrus sensitive to speaker-specific GPR and left posterior STG/STS processing vocal tract dynamics. However, as the present study was designed to investigate how speaker-specific GPR information interacts with speech-specific GPR information (i.e., prosody) we cannot test this hypothesis with the current dataset.

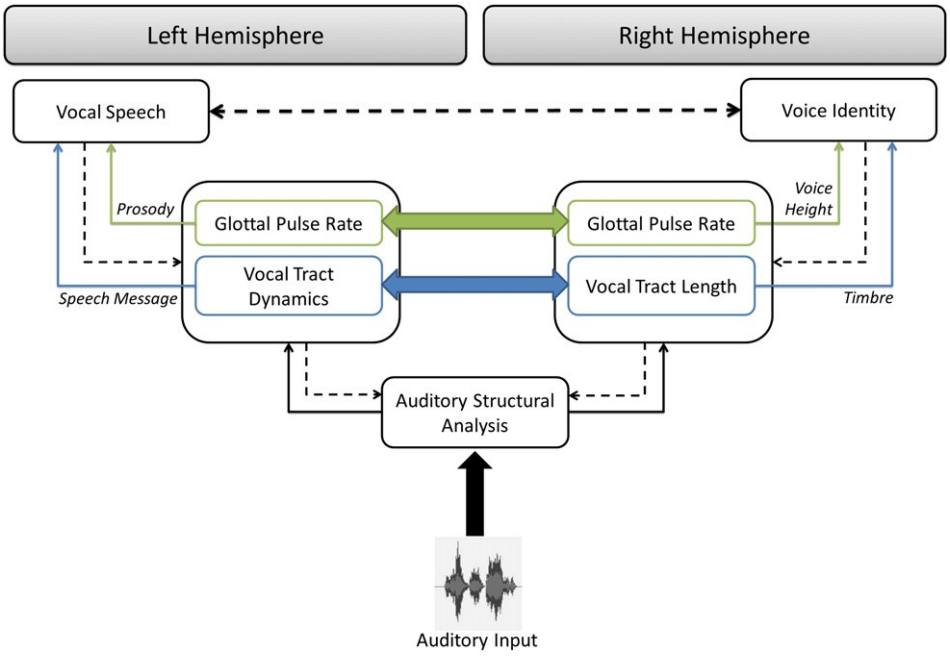


Fig. 5. Schematic illustration of a predictive-coding mechanism that the human brain might use to process auditory speech despite acoustic differences between speakers. This mechanism involves interactions between neural processes that are involved in speech (left side of the figure) and speaker recognition (right side of the figure). These interactions serve to inform speech processing areas in the left hemisphere about speaker parameters processed in the right hemisphere. Speaker parameters (i.e., glottal fold and vocal tract parameters) are processed in distinct cortical areas (i.e., Heschl's gyrus for glottal fold parameters and STG/STS for vocal tract parameters) and interact with homologous regions of the speech-processing pathway in the left hemisphere. This mechanism could explain behavioral findings showing interactions between speech- and speaker-related processes, which are difficult to explain under a view that considers analyses of speech and voices to be independent processes. Solid lines, forward connections conveying prediction error; dashed lines, backward and lateral connections conveying predictions.

The proposed mechanism (Fig. 5) implicitly assumes that speech recognition, including recognition of the message as well as the linguistic prosody, is processed predominantly in the left hemisphere. It is commonplace that processing of the speech message is predominantly leftlateralized (e.g., Leff et al., 2008; McGettigan and Scott, 2012; Scott, 2005; Vigneau et al., 2006). The functional lateralization in prosody processing, however, is more controversial (reviewed in Friederici and Alter, 2004). Processing of emotional prosody seems to be rightlateralized, whereas linguistic prosody processing occurs in a bilaterally distributed network of brain regions (Doherty et al., 2004; Gandour et al., 2003; Tong et al., 2005). Human fMRI (Wildgruber et al., 2004) and clinical studies (Pell and Baum, 1997) directly comparing emotional and linguistic prosody have revealed a left-hemispheric bias in the processing of linguistic prosody. Support for the assumption that the left Heschl's gyrus is involved in the analysis of linguistic prosody comes from a study examining the relationship between brain volume and tone language learning (Wong et al., 2008). In that study, participants who successfully learned to relate pitch patterns to word meaning had greater brain volume in the left Heschl's gyrus than less successful learners. The stimuli of that study comprised monosyllabic pseudowords with rising and falling pitch contours similar to the syllables used in the present study. Interestingly, it seems that a posteromedial region of left Heschl's gyrus was associated with successful learning of linguistic pitch patterns (cf. Fig. 3 in Wong et al., 2008) — the same region we found to be functionally connected to the right Heschl's gyrus when recognizing linguistic prosody from speakers who differ in glottal fold parameters.

Conclusions

In summary, the present results suggest that the human brain uses interactions between right- and left-hemispheric areas as a mechanism to deal with speaker changes when processing linguistic aspects of speech; this includes both the recognition of the speech message and linguistic prosody. We proposed a potential neural mechanism (Fig. 5) in which acoustic parameters of voices are extracted in right-hemispheric areas (i.e., Heschl's gyrus for glottal fold

parameters and STG/STS for vocal tract parameters) and communicated to homologous areas in the left hemisphere processing linguistic aspects of the input. Interactions between the left and right hemispheres might not only represent a powerful neural mechanism for robust speech recognition in the context of acoustic speaker variability but might also apply to other auditory domains. Furthermore, the proposed mechanism might inspire novel types of artificial speech recognition systems which currently still have difficulties adapting to speaker-related variations (O'Shaughnessy, 2008) and do not efficiently exploit speaker-related glottal fold parameters (Meyer et al., 2007).

Supplementary data to this article can be found at the end of this manuscript.

Funding

This work was supported by a Max Planck Research Group grant to KVK. EG was supported by a VIDI grant from the Netherlands Organization for Scientific Research, NWO, and Netherlands Organization for Health Research and Development, ZonMw (grant no. 016.096.397), and by the UK Medical Research Council (G9900369 and G0500221).

Acknowledgments

We thank Samuel R. Mathias and Stefan J. Kiebel for comments on an earlier version of this manuscript and Katja Mayer for her help with acquiring the data.

References

Backes, W.H., Van Dijk, P., 2002. Simultaneous sampling of event-related BOLD responses in auditory cortex and brainstem. Magn. Reson. Med. 47 (1), 90–96.

Balaguer-Ballester, E., Clark, N.R., Coath, M., Krumbholz, K., Denham, S.L., 2009. standing pitch perception as a hierarchical process with top-down modulation. PLoS Comput. Biol. 5.

Barreda, S., Nearey, T.M., 2012. The direct and indirect roles of fundamental frequency in vowel perception. J. Acoust. Soc. Am. 131, 466–477.

Belin, P., Fecteau, S., Bedard, C., 2004. Thinking the voice: neural correlates of voice perception. Trends Cogn. Sci. 8, 129–135.

Best, V., Ozmeral, E.J., Kopco, N., Shinn-Cunningham, B.G., 2008. hances selective auditory attention. Proc. Natl. Acad. Sci. U. S. A. 105, 13174–13178. Bradlow, A.R., Pisoni, D.B., 1999. Recognition of spoken words by native and non-native listeners: talker-, listener-, and item-related factors. J. Acoust. Soc. Am. 106, 2074–2085.

Bradlow, A.R., Nygaard, L.C., Pisoni, D.B., 1999. ation on recognition memory for spoken words. Percept. Psychophys. 61, 206–219.

Brechmann, A., Scheich, H., 2005. Hemispheric shifts of sound representation in auditory cortex with conceptual listening. Cereb. Cortex 15 (5), 578–587.

Brechmann, A., Gaschler-Markefski, B., Sohr, M., Yoneda, K., Kaulisch, T., Scheich, H., 2007. Working Memory–Specific Activity in Auditory Cortex: Potential Correlates of Sequential Processing and Maintenance. Cereb. Cortex 17 (11), 2544–2552.

Chandrasekaran, B., Chan, A.H.D., Wong, P.C.M., 2011. Neural processing of what and who information in speech. J. Cogn. Neurosci. 23, 2690–2700.

Cutler, A., Eisner, F., McQueen, J.M., Norris, D., 2010. How abstract phonemic categories are necessary for coping with speaker-related variation. Laboratory Phonology, 10.

Doherty, C.P., West, W.C., Dilley, L.C., Shattuck-Hufnagel, S., Caplan, D., 2004. Question/ statement judgments: an fMRI study of intonation processing. Hum. Brain Mapp. 23, 85–98.

Fitch, W.T., Giedd, J., 1999. Morphology and development of the human vocal tract: a study using magnetic resonance imaging. J. Acoust. Soc. Am. 106, 1511–1522.

Friederici, A.D., Alter, K., 2004. Lateralization of auditory language functions: a dynamic dual pathway model. Brain Lang. 89, 267–276.

Friston, K.J., Buechel, C., Fink, G.R., Morris, J., Rolls, E., Dolan, R.J., 1997. Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6, 218–229.

Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., Penny, W.D., 2007. metric Mapping: The Analysis of Funtional Brain Images. Elsevier/Academic Press.

Gaab, N., Gabrieli, J.D.E., Glover, G.H., 2007. Assessing the influence of scanner background noise on auditory processing. I. An fMRI study comparing three experimental designs with varying degrees of scanner noise. Hum. Brain Mapp. 28, 703–720.

Gandour, J., Wong, D., Dzemidzic, M., Lowe, M., Tong, Y.X., Li, X.J., 2003. A cross-linguistic fMRI study of perception of intonation and emotion in Chinese. Hum. Brain Mapp. 18, 149–157.

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: 10th Annual Conference of the International Speech Communication Association 2009, vol. 1–5, pp. 152–155.

Griffiths, T.D., Hall, D.A., 2012. Mapping pitch representation in neural ensembles with FMRI. J. Neurosci. 32, 13343–13347.

Griffiths, T.D., Uppenkamp, S., Johnsrude, I., Josephs, O., Patterson, R.D., 2001. Encoding of the temporal regularity of sound in the human brainstem. Nat. Neurosci. 4, 633–637. Griffiths, T.D., Kumar, S., Sedley, W., Nourski, K.V., Kawasaki, H., Oya, H., Patterson, R.D.,

Brugge, J.F., Howard, M.A., 2010. Direct recordings of pitch responses from human auditory cortex. Curr. Biol. 20, 1128–1132.

Gutschalk, A., Patterson, R.D., Scherg, M., Uppenkamp, S., Rupp, A., 2004. Temporal dynamics of pitch in human auditory cortex. Neuroimage 22, 755–766.

Hall, D.A., Haggard, M.P., Akeroyd, M.A., Palmer, A.R., Summerfield, A.Q., Elliott, M.R., Gurney, E.M., Bowtell, R.W., 1999. "Sparse" temporal sampling in auditory fMRI. Hum. Brain Mapp. 7, 213–223.

Hickok, G., Poeppel, D., 2007. Opinion — the cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402.

Kaganovich, N., Francis, A.L., Melara, R.D., 2006. teraction between talker and linguistic information during speech perception. Brain Res. 1114, 161–172.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., Banno, H., 2008. straight: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1–12, pp. 3933–3936.

Kiebel, S.J., Daunizeau, J., Friston, K.J., 2008. A hierarchy of time-scales and the brain. PLoS Comput. Biol. 4.

Kitterick, P.T., Bailey, P.J., Summerfield, A.Q., 2010. Benefits of knowing who, where, and when in multi-talker listening. J. Acoust. Soc. Am. 127, 2498–2508.

Kumar, S., Sedley, W., Nourski, K.V., Kawasaki, H., Oya, H., Patterson, R.D., Howard, M.A., Friston, K.J., Griffiths, T.D., 2011. Predictive coding and pitch processing in the auditory cortex. J. Cogn. Neurosci. 23, 3084–3094.

Laing, E.J., Liu, R., Lotto, A.J., Holt, L.L., 2012. Tuned with a tune: talker normalization via general auditory processes. Front. Psychol. 3.

Lavner, Y., Gath, I., Rosenhouse, J., 2000. tification of familiar voices speaking isolated vowels. Speech Comm. 30, 9–26.

Leff, A.P., Schofield, T.M., Stephan, K.E., Crinion, J.T., Friston, K.J., Price, C.J., 2008. The cortical dynamics of intelligible speech. J. Neurosci. 28, 13209–13215.

Levi, S.V., Winters, S.J., Pisoni, D.B., 2011. Effects of cross-language voice training on speech perception: whose familiar voices are more intelligible? J. Acoust. Soc. Am. 130, 4053–4062.

Luks, T.L., Nusbaum, H.C., Levy, J., 1998. tactic prosody is dynamically dependent on task demands. Brain Lang. 65 (2), 313–332. Magnuson, J.S., Nusbaum, H.C., 2007. Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. J. Exp. Psychol. Hum. Percept. Perform. 33, 391–409.

McGettigan, C., Scott, S.K., 2012. Cortical asymmetries in speech perception: what's wrong, what's right and what's left? Trends Cogn. Sci. 16, 269–276.

Meyer, B.T., Wachter, M., Brand, T., Kollmeier, B., 2007. Phoneme confusions in human and automatic speech recognition. Interspeech 2007: 8th Annual Conference of the International Speech Communication Association, vol. 1–4, pp. 2740–2743.

Morey, R.D., 2008. Confidence intervals from normalized data: a correction to Cousineau (2005). Tutor. Quant. Methods Psychol. 4, 61–64.

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., Zilles, K., 2001. Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. Neuroimage 13, 684–701.

Nusbaum, H., Magnuson, J., 1997. Talker normalization: phonetic constancy as a cognitive process. Talker Variability in Speech Processing 109–132.

Nygaard, L.C., 2005. Perceptual integration of linguistic and nonlinguistic properties of speech. The Handbook of Speech Perception, pp. 390–413.

Nygaard, L.C., Pisoni, D.B., 1998. Talker-specific learning in speech perception. Percept. Psychophys. 60, 355–376.

Nygaard, L.C., Sommers, M.S., Pisoni, D.B., 1994. Speech-perception as a talker-contingent process. Psychol. Sci. 5, 42–46.

Obleser, J., Eisner, F., 2009. Pre-lexical abstraction of speech in the auditory cortex. Trends Cogn. Sci. 13, 14–19.

Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh Inventory. Neuropsychologia 9, 97–113.

O'Shaughnessy, D., 2008. Invited paper: automatic speech recognition: history, methods and challenges. Pattern Recogn. 41, 2965–2979.

Palmeri, T.J., Goldinger, S.D., Pisoni, D.B., 1993. Episodic encoding of voice attributes and recognition memory for spoken words. J. Exp. Psychol. Learn. Mem. Cogn. 19, 309–328. Patterson, R.D., Uppenkamp, S., Johnsrude, I.S., Griffiths, T.D., 2002. The processing of temporal pitch and melody information in auditory cortex. Neuron 36, 767–776.

Pell, M.D., Baum, S.R., 1997. Unilateral brain damage, prosodic comprehension deficits, and the acoustic cues to prosody. Brain Lang. 57, 195–214.

Penagos, H., Melcher, J.R., Oxenham, A.J., 2004. A neural representation of pitch salience in aging. J. Neurosci. 24, 6810–6815.

Peterson, G.E., Barney, H.L., 1952. Control methods used in a study of the vowels. J. Acoust. Soc. Am. 24, 175–184.

Pisoni, D.B., 1993. Long-Term-memory in speech-perception — some new findings on talker variability, speaking rate and perceptual-learning. Speech Comm. 13, 109–125. Pisoni, D.B., 1997. Some thoughts on "normalization" in speech perception. Talker Variability in Speech Processing 9–32.

Plante, E., Creusere, M., Sabin, C., 2002. cessing: activation interacts with task demands. Neuroimage 17 (1), 401–410.

Poldrack, R.A., Mumford, J.A., 2010. On the proper role of nonindependent ROI analysis: a commentary on Vul and Kanwisher. Foundational Issues in Human Brain Mapping 93–96.

Puschmann, S., Uppenkamp, S., Kollmeier, B., Thiel, C.M., 2010. Dichotic pitch activates pitch processing centre in Heschl's gyrus. Neuroimage 49, 1641–1649.

Reiterer, S.M., Erb, M., Droll, C.D., Anders, S., Ethofer, T., Grodd, W., Wildgruber, D., 2005. Impact of task difficulty on lateralization of pitch and duration discrimination. NeuroReport 16 (3), 239–242.

Remez, R.E., Dubowski, K.R., Broder, R.S., Davids, M.L., Grossman, Y.S., Moskalenko, M., Pardo, J.S., Hasbun, S.M., 2009. Auditory-phonetic projection and lexical structure in the recognition of sine-wave words. J. Exp. Psychol. Hum. Percept. Perform. 125, 2656.

Rinne, T., Koistinen, S., Salonen, O., Alho, K., 2009. Task-dependent activations of human auditory cortex during pitch discrimination and pitch memory tasks. J. Neurosci. 29 (42), 13338–13343.

Schönwiesner, M., Krumbholz, K., Rübsamen, R., Fink, G.R., von Cramon, D.Y., 2007. Hemispheric asymmetry for auditory processing in the human auditory brain stem, thalamus, and cortex. Cereb. Cortex 17 (2), 492–499.

Schweinberger, S.R., Walther, C., Zaske, R., Kovacs, G., 2011. tion to voice identity. Br. J. Psychol. 102, 748–764.

Scott, S.K., 2005. Auditory processing — speech, space and auditory objects. Curr. Opin. Neurobiol. 15, 197–201.

Scott, S.K., Johnsrude, I.S., 2003. The neuroanatomical and functional organization of speech perception. Trends Neurosci. 26, 100–107.

Smith, D.R.R., Patterson, R.D., 2005. The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. J. Acoust. Soc. Am. 118, 3177–3186.

Tong, Y.X., Gandour, J., Talavage, T., Wong, D., Dzemidzic, M., Xu, Y.S., Li, X.J., Lowe, M., 2005. Neural circuitry underlying sentence-level linguistic prosody. Neuroimage 28, 417–428.

Vigneau, M., Beaucousin, V., Herve, P.Y., Duffau, H., Crivello, F., Houde, O., Mazoyer, B., Tzourio-Mazoyer, N., 2006. ogy, semantics, and sentence processing. Neuroimage 30, 1414–1432.

von Kriegstein, K., Smith, D.R.R., Patterson, R.D., Ives, D.T., Griffiths, T.D., 2007. resentation of auditory size in the human voice and in sounds from other resonant sources. Curr. Biol. 17, 1123–1128.

von Kriegstein, K., Smith, D.R.R., Patterson, R.D., Kiebel, S.J., Griffiths, T.D., 2010. How the human brain recognizes speech in the context of changing speakers. J. Neurosci. 30, 629–638.

Wildgruber, D., Hertrich, I., Riecker, A., Erb, M., Anders, S., Grodd, W., Ackermann, H., 2004. Distinct frontal regions subserve evaluation of linguistic and emotional aspects of speech intonation. Cereb. Cortex 14, 1384–1389.

Wong, P.C.M., Nusbaum, H.C., Small, S.L., 2004. Neural bases of talker normalization. J. Cogn. Neurosci. 16, 1173–1184.

Wong, P.C.M., Warrier, C.M., Penhune, V.B., Roy, A.K., Sadehh, A., Parrish, T.B., Zatorre, R.J., 2008. Volume of left Heschl's gyrus and linguistic pitch learning. Cereb. Cortex 18, 828–836.

Yonan, C.A., Sommers, M.S., 2000. The effects of talker familiarity on spoken word identification in younger and older listeners. Psychol. Aging 15, 88–99.

Zhang, W.T., Mainero, C., Kumar, A., Wiggins, C.J., Benner, T., Purdon, P.L., Bolar, D.S., Kwong, K.K., Sorensen, A.G., 2006. tivation in trigeminal pathways with cardiac gating. Neuroimage 31, 1506–1512.

Inline Supplementary Table 1.
Activity results associated with prosody processing

This table presents the activity results as revealed by the main effect of prosody task: prosody task - speaker task. Behavioral performance was added as a covariate in the analysis (see the section "Activity analysis"). The table shows activity in areas that was present at $p < 0.001$ uncorrected together with sub-clusters (indented). Areas that showed activity at $p < 0.05$ FWE corrected are shown in bold font. Anatomical labels of cortical areas were based on the Harvard-Oxford cortical structural atlas. Anatomical labels of subcortical and cerebellar areas were based on the Harvard-Oxford subcortical structural atlas and the MNI structural atlas, respectively. All atlases were used as implemented in FSLview.

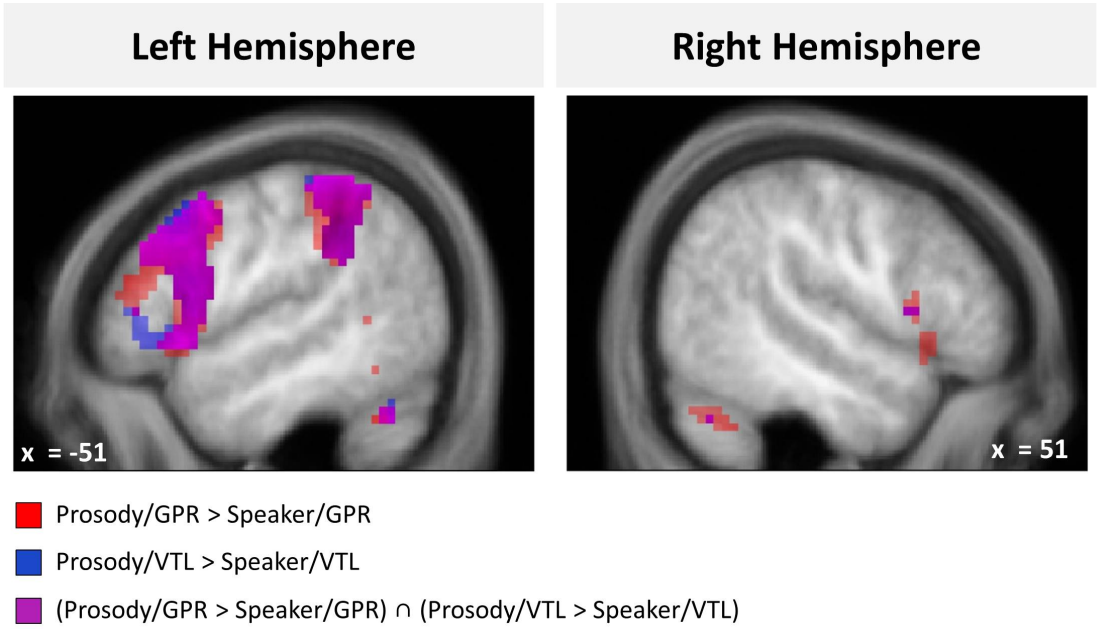
Region	k	t-value	p-value		MNI coordinates		
			FWE-corrected	uncorrected	x	y	z
Left Thalamus	1445	10.69	0.001	1.04×10^{-8}	-9	-10	10
Right Caudate					12	2	10
Left Putamen					-24	14	4
Right Cerebellum	1073	10.59	0.001	1.04×10^{-8}	18	-76	-41
Left Supramarginal Gyrus	631	9.79	0.002	3.32×10^{-8}	-54	-43	43
Left Superior Parietal Lobule					-39	-49	49
Right Frontal Pole	193	9.59	0.002	4.32×10^{-8}	33	47	19
Right Middle Frontal Gyrus					39	35	37
Left Inferior Frontal Gyrus	2178	8.54	0.011	1.91×10^{-7}	-54	14	13
Right Superior Frontal Gyrus					24	2	64
Left Precuneus	117	6.70	0.161	3.56×10^{-6}	-12	-67	55
Right Supramarginal Gyrus	16	4.59	0.948	1.78×10^{-4}	57	-34	43
Right Inferior Frontal Gyrus	10	4.46	0.970	2.28×10^{-4}	51	14	7
Left Middle Temporal Gyrus	5	4.08	0.997	4.96×10^{-4}	-51	-52	7
Left Frontal Pole	1	3.76	0.999	9.45×10^{-4}	-21	44	25

Inline Supplementary Table 2.
Activity results associated with voice processing

This table presents the activity results as revealed by the main effect of speaker task: speaker task - prosody task. Behavioral performance was added as a covariate (see the section "Activity analysis"). The table shows activity in areas that was present at $p < 0.001$ uncorrected together with sub-clusters (indented). Areas that showed activity at $p < 0.05$ FWE corrected are shown in bold font. Anatomical labels of cortical areas were based on the Harvard-Oxford cortical structural atlas. Anatomical labels of subcortical areas were based on the Harvard-Oxford subcortical structural atlas. Both atlases were used as implemented in FSLview.

Region	k	t-value	p-value		MNI coordinates		
			FWE-corrected	uncorrected	x	y	z
Left Hippocampus	727	11.58	0.0002	3.53×10^{-9}	-21	-10	-17
Left Insular Cortex					-42	-7	-2
Right Precuneus	1077	9.83	0.002	3.14×10^{-8}	24	-76	40
Right Lateral Occipital Cortex					33	-82	13
Right Central Operculum	919	8.58	0.010	1.79×10^{-7}	39	-1	16
Right Frontal Pole	91	8.36	0.014	2.49×10^{-7}	27	32	-14
Frontal Pole	494	7.99	0.025	4.34×10^{-7}	0	62	-5
Paracingulate Gyrus					-9	50	22
Right Hippocampus	350	7.53	0.051	9.02×10^{-7}	24	-10	-17
Right Parahippocampal Gyrus					27	-7	-29
Right Insular Cortex					36	2	-11
Left Lateral Occipital Cortex	618	6.71	0.159	3.49×10^{-6}	-39	-73	4
Left Occipital Pole					-24	-97	16
Left Middle Temporal Gyrus	105	6.62	0.176	4.05×10^{-6}	-57	-4	-17
Left Superior Temporal Gyrus					-51	-10	-11
Left Cingulate Gyrus	68	5.20	0.721	5.38×10^{-5}	-12	-43	28
Right Cingulate Gyrus					9	-7	43
Right Inferior Temporal Gyrus	9	4.53	0.959	2.00×10^{-4}	48	-1	-32
Left Precentral Gyrus	28	4.40	0.978	2.56×10^{-4}	-3	-28	61
Left Postcentral Gyrus	11	4.37	0.981	2.37×10^{-4}	-21	-31	70
Left Lingual Gyrus	5	4.33	0.997	4.73×10^{-4}	-30	-43	-5

Supplementary Figure 1



Supplementary Figure 1. BOLD response associated with recognition of linguistic prosody. The figure shows BOLD responses to the prosody as compared to the speaker task separately for GPR change (i.e., prosody task/GPR change > speaker task/GPR change) (red) and VTL change (i.e., prosody task/VTL change > speaker task/VTL change) (blue) as well as for conjunction of both contrasts [i.e., (prosody task/GPR change > speaker task/GPR change) ∩ (prosody task/GPR change > speaker task/GPR change)] (magenta). For display purposes only, the activation patterns are shown at $p < 0.001$ uncorrected.