Behavioral/Systems/Cognitive

How the Human Brain Recognizes Speech in the Context of Changing Speakers

Katharina von Kriegstein,[1,2,3] David R. R. Smith,[4,5] Roy D. Patterson,[5] Stefan J. Kiebel,[1,3] and Timothy D. Griffiths[1,2]

[1]Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, United Kingdom, [2]Auditory Group, Medical School, University of Newcastle-upon-Tyne, Newcastle-upon-Tyne NE2 4HH, United Kingdom, [3]Max Planck Institute for Cognitive and Brain Science, 04103 Leipzig, Germany, [4]Department of Psychology, University of Hull, Hull HU6 7RX, United Kingdom, and [5]Centre for the Neural Basis of Hearing, University of Cambridge, Cambridge CB2 3EG, United Kingdom

We understand speech from different speakers with ease, whereas artificial speech recognition systems struggle with this task. It is unclear how the human brain solves this problem. The conventional view is that speech message recognition and speaker identification are two separate functions and that message processing takes place predominantly in the left hemisphere, whereas processing of speaker-specific information is located in the right hemisphere. Here, we distinguish the contribution of specific cortical regions, to speech recognition and speaker information processing, by controlled manipulation of task and resynthesized speaker parameters. Two functional magnetic resonance imaging studies provide evidence for a dynamic speech-processing network that questions the conventional view. We found that speech recognition regions in left posterior superior temporal gyrus/superior temporal sulcus (STG/STS) also encode speaker-related vocal tract parameters, which are reflected in the amplitude peaks of the speech spectrum, along with the speech message. Right posterior STG/STS activated specifically more to a speaker-related vocal tract parameter change during a speech recognition task compared with a voice recognition task. Left and right posterior STG/STS were functionally connected. Additionally, we found that speaker-related glottal fold parameters (e.g., pitch), which are not reflected in the amplitude peaks of the speech spectrum, are processed in areas immediately adjacent to primary auditory cortex, i.e., in areas in the auditory hierarchy earlier than STG/STS. Our results point to a network account of speech recognition, in which information about the speech message and the speaker's vocal tract are combined to solve the difficult task of understanding speech from different speakers.

Introduction

The same sentence, spoken by different speakers, can sound very different, and the acoustic differences between speakers enable us to relate speech to a specific person and recognize each other by voice. Also, changing voice properties can be useful in adapting to a specific context, such as whispering in quiet surroundings. However, to extract meaning and understand what has been said, the variability within and between speakers must be resolved. This is a nontrivial problem, and the sophisticated algorithms used by speech recognition machines still perform well below humans (Deng et al., 2006; O'Shaughnessy, 2008). Currently, it is unclear why human speech recognition is so robust to variation in speaker characteristics (Friederici, 2002; Wong et al., 2004; Hickok and Poeppel, 2007; Obleser and Eisner, 2009).

One of the main obstacles to understanding speech from different speakers is that the formant frequencies (i.e., the amplitude peaks in the frequency spectrum of speech sounds) contain information about both the type of speech sound (/a/, /i/, /n/, etc.) and speaker-related vocal tract parameters. This information is fundamentally intermingled and difficult to separate (Joos, 1948; Ladefoged and Broadbent, 1957; Sussman, 1986; Nearey, 1989; Welling et al., 2002; Adank et al., 2004; Johnson, 2005; Ames and Grossberg, 2008; Turner et al., 2009) (see Fig. 1). In contrast, glottal fold parameters do not affect the formant position or timbre. Rather, they determine voice pitch or whether speech is voiced or whispered.

With a recently described approach (Kawahara et al., 2004), natural speech sounds can be modified in a major speaker-related vocal tract parameter, i.e., vocal tract length (VTL). Previous functional magnetic resonance imaging (fMRI) studies using these resynthesized sounds revealed that regions in posterior superior temporal gyrus/superior temporal sulcus (STG/STS) respond specifically to VTL information in human speech in contrast to similar information in, for example, animal calls (von Kriegstein et al., 2007). The reason for this specificity is unclear. VTL information is an important cue for speaker recognition (Lavner et al., 2000), and regions responding to this information might be involved in recognizing other humans by voice (Belin et al., 2004; von Kriegstein and Giraud, 2004). Another reason might be that posterior STG/STS is responsive to VTL changes, because this area contributes to speech recognition by processing information about speaker-specific vocal tract dynamics. Using two fMRI studies, we focus on testing this latter hypothesis. We show that VTL-sensitive regions in posterior STG/STS are in-

Received June 11, 2009; revised Oct. 1, 2009; accepted Nov. 5, 2009.

This work was supported by Volkswagen Stiftung Grant I/79 783, the Wellcome Trust, and the United Kingdom Medical Research Council Grants G9900362 and G0500221. We thank Tom C. Walters for helping with preparation of Figure 1.

Correspondence should be addressed to Katharina von Kriegstein, Max Planck Institute for Cognitive and Brain Science, Stephanstrasse 1A, 04103 Leipzig, Germany. E-mail: kriegstein@cbs.mpg.de.

DOI:10.1523/JNEUROSCI.2742-09.2010

 0270-6474/10/300629-10$15.00/0

volved in speech recognition and that left and right posterior STG/STS are functionally connected when recognizing speech in the context of changing speakers. In addition, we harnessed the distinction between vocal tract and glottal fold parameters to show that (1) posterior STG/STS is involved in processing both speaker-specific formant and speech information and that (2) speaker-related glottal fold and vocal tract parameters are processed in separate brain regions. We present a hypothesis of how speaker-related acoustic variability is dealt with by the human brain. In addition, we discuss the implications of our findings for two influential but opposing theoretical accounts (abstractionist vs exemplar models) of speech processing (Goldinger, 1996; Pisoni, 1997).

Materials and Methods

Stimuli

The stimuli were based on syllables recorded from a single speaker (16 bit resolution, 48 kHz sample rate) that were preprocessed with level balancing to minimize loudness differences, and perceptual centering to reduce rhythmic distractions as described previously (von Kriegstein et al., 2006). Experiment 1 contained 96 syllables (48 consonant–vowel, 48 vowel–consonant). Experiment 2 contained 150 vowel–consonant–vowel syllables. Versions of the original syllables were synthesized to simulate speakers with different glottal pulse rate (GPR) and VTL using a sophisticated vocoder referred to as STRAIGHT (Kawahara et al., 1999, 2004). In addition, whispered syllables were produced by resynthesizing the recorded speech sounds with a broadband noise and lifting the spectrum 6 dB per octave to match the spectral slope of whispered speech (Fujimura and Lindqvist, 1971). For both experiments, spoken syllables were concatenated to form syllable sequences. Example syllable sequences for both experiments are available online as supplemental data (available at www.jneurosci.org as supplemental material). In experiment 1 (supplemental Fig. S1, available at www.jneurosci.org as supplemental material), sequences lasted 9.44 s and contained eight syllabic events (680 ms stimulus, 500 ms pause). In experiment 2 (supplemental Fig. S1, available at www.jneurosci.org as supplemental material), all syllable sequences lasted 8.4 s and contained six syllabic events (1100 ms stimulus, 300 ms pause). Before each sequence, participants received a visual instruction to perform either a speech recognition task ("speech task") or a control task (which was a "loudness task" in experiment 1 and a "speaker task" in experiment 2) (see below).

Experimental design

Experiment 1

Experiment 1 was a $2 \times 2 \times 2$ factorial design with the factors VTL (VTL varies/VTL fixed), task (speech task/loudness task), and glottal fold parameters (voiced/whispered) (supplemental Fig. S1, available at www.jneurosci.org as supplemental material). It was used to address three questions that we will detail in the following.

Do VTL-sensitive regions in posterior STG/STS also participate in speech recognition tasks? To locate VTL-sensitive regions, half of the syllable sequences contained syllable events that differed in vocal tract length (VTL varies); during the other half, the VTL of the speaker was fixed (VTL fixed). VTL values were resynthesized to range from 10.6 to 21.7 cm in eight equal logarithmic steps. To investigate responses to speech recognition, we included a speech task (speech task) and a control task (loudness task) in the design. In the speech task, subjects indicated via button press whether the current syllable was different from the previous one. In the loudness task, subjects indicated via button press whether the level of the current syllable was different from the previous one. Within each syllable sequence, there were three different syllables (e.g., /ga/, /ke/, /la/; /mu/, /mi/, /ka/; etc.) and three different values of sound level [values differed by 9–12 dB sound pressure level (SPL)]. The changes in syllable and sound level were independent. Each sequence (with a specific stimulus combination) always occurred twice, once in the speech task and once in the loudness task. To address the question whether posterior STG/STS responds to VTL as well as to the speech task, we tested regions responsive to VTL ("main effect of VTL"), to the speech task ("main effect of task"), as well as the interaction between the two ("VTL × task"). In this interaction, we were specifically interested in regions responding more to a speech task when VTL varied while controlling for stimulus as well as for task effects: (VTL varies/speech task > VTL varies/loudness task) > (VTL fixed/speech task > VTL fixed/loudness task).

Do VTL-sensitive regions in posterior STG/STS respond differently to different glottal fold parameters, i.e., voiced and whispered speech? Glottal fold parameters do not influence the formant position in the speech spectrum (Fig. 1). Therefore, if posterior STG/STS contains a mechanism for formant processing, responses to voiced and whispered speech should be similar in this region. Half of the syllable sequences in experiment 1 were voiced (fundamental frequency set at 160Hz) and half of them were whispered. To check whether VTL-sensitive regions in posterior STG/STS respond similarly to voiced and whispered speech, we used two approaches. First, we performed contrasts for the main effect of VTL, the main effect of task, and the VTL × task interaction separately for voiced and whispered speech and entered these contrasts in a second-level t statistic for a conjunction analysis (e.g., conjunction of "main effect of VTL voiced" and "main effect of VTL whispered"). Second, we tested the interaction between the contrasts of interest with the factor "glottal fold parameter" at a relatively low statistical threshold ($p = 0.01$ uncorrected).

Where are glottal fold parameters processed in the human brain? The inclusion of voiced and whispered speech permits a test of where these two glottal fold parameters are processed differentially in the human brain. We probed the "main effect of glottal fold parameter" in both directions, i.e., voiced > whispered and whispered > voiced.

In summary, experiment 1 had a $2 \times 2 \times 2$ factorial design with eight experimental conditions: (1) speech task, VTL varies, whispered; (2) speech task, VTL varies, voiced; (3) speech task, VTL fixed, whispered; (4) speech task, VTL fixed, voiced; (5) loudness task, VTL varies, whispered; (6) loudness task, VTL varies, voiced; (7) loudness task, VTL fixed, whispered; (8) loudness task, VTL fixed, voiced. The experiment also included a silence condition. The order of conditions was randomized.

Experiment 2

Experiment 2 was a 2×2 factorial design with the factors VTL (VTL varies/GPR varies) and task (speech task/speaker task) (supplemental Fig. S1, available at www.jneurosci.org as supplemental material). It was designed to complement experiment 1 by addressing the following two questions.

Is VTL-sensitive posterior STG/STS specifically processing formant information? When listening to sequences with varying VTL, subjects usually have the impression that the speech sounds are produced by different synthetic speakers (E. Gaudrain, S. Li, V. S. Ban, R. D. Patterson, unpublished observations). In contrast, the sequences with fixed VTL are perceived as spoken by the same speaker. The high-level percept of different speakers is a confound when investigating the acoustic effect of vocal tract length. By acoustic effect, we mean the speaker-related shift in formant positions. In experiment 2, half of the syllable sequences were spoken by speakers that differed in VTL, and the other half was spoken by speakers that differed in the vibration rate of the glottal folds (GPR). The GPR and VTL values (GPR: 95, 147, 220 Hz; VTL: 9.1, 13.6, 20.3 cm) were chosen because preliminary behavioral studies indicated that subjects perceive these values as a change of speaker rather than a change of the voice characteristics within the speaker. GPR changes affect the pitch of the syllable but do not alter the formant positions. In contrast, VTL changes shift the formant frequencies but not the pitch. Thus, only VTL information is intermingled with the formant information determining the speech message (e.g., /a/), whereas GPR information is independent (Fig. 1). To test whether posterior STG/STS is processing speaker-related formant information, we probed the main effect of VTL (i.e., VTL varies > GPR varies).

Is VTL-sensitive posterior STG/STS specifically modulated by the speech task? During a speech task, subjects might automatically process the speaker characteristics of the stimulus significantly more than they do in a comparable loudness task. This could potentially explain differential responses in experiment 1 for the main effect of task (speech task >

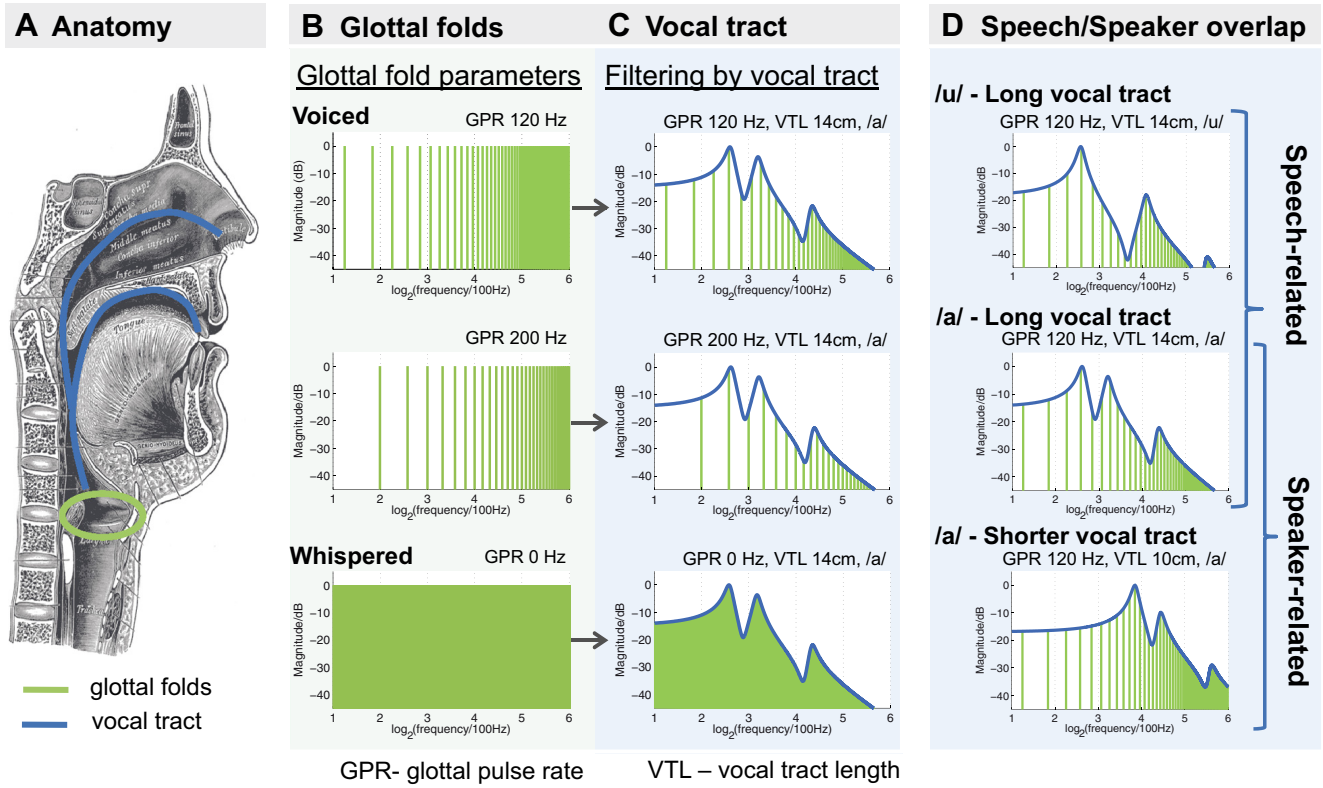

Figure 1. The contribution of glottal fold and vocal tract parameters to the speech output. ***A***, Shown is a sagittal section through a human head and neck. Green circle, Glottal folds; blue lines, extension of the vocal tract from glottal folds to tip of the nose and lips. ***B***, The three plots show three different sounds determined by glottal fold parameters. In voiced speech, the vibration of the glottal folds results in lower voices (120 Hz GPR; top) or higher voices (200 Hz GPR; middle). If glottal folds are constricted, they produce a noise-like sound that is heard as whispered speech (0 Hz GPR; bottom). ***C***, The vocal tract filters the sound wave coming from the glottal folds, which introduces amplitude peaks at certain frequencies ("formants"; blue lines). Note that the different glottal fold parameters do not influence the formant position. ***D***, Both speech- and speaker-related vocal tract parameters influence the position of the formants. Here we show as an example the formant shifts associated with the speech sounds /u/ and /a/ (first and second plot) and an /a/ with a shorter and longer vocal tract length (second and third plot).

loudness task) and the VTL × task interaction [(VTL varies/speech task > VTL varies/loudness task) > (VTL fixed/speech task > VTL fixed/loudness task)]. Such an explanation would counter our hypothesis that posterior STG/STS is responding to speaker-specific formant information to use this information for speech recognition. Accordingly, in experiment 2, we included not only a speech task but also a speaker task. In the speech task, subjects indicated via button press whether the current syllable was different from the previous one. In the speaker task, subjects indicated via button press whether the current speaker was different from the previous one. Subjects were asked to only score two consecutive syllable events as different if they clearly perceived a change of speaker rather than a change of the voice of one speaker. Within each sequence, there were three different syllables (e.g., /aga/, /ake/, /ala/; or /esi/, /elu/, /ero/; etc.) and three different speakers (i.e., different VTLs or different GPRs). Changes in syllable and speaker were independent. Each sequence (with a specific stimulus combination) always occurred twice, once in the speech task and once in the speaker task.

To test whether VTL-sensitive posterior STG/STS is specifically involved in speech recognition, we tested the blood oxygen level-dependent (BOLD) signal changes for the contrast main effect of task (i.e., in the direction speech > speaker task) and for the task × VTL interaction, i.e., (VTL varies/speech task > VTL varies/speaker task) > (GPR varies/speech task > GPR varies/speaker task).

In summary, experiment 2 was a 2 × 2 factorial design with four conditions: (1) speech task, VTL varies; (2) speech task, GPR varies; (3) speaker task, VTL varies; (4) speaker task, GPR varies. The experiment additionally included a silence condition. The order of conditions was randomized.

Participants

Eighteen subjects participated in experiment 1 (all right handed; 10 female, 8 male; aged 19–40 years; mean age of 26 years; native language: 15 English, 2 German, 1 Spanish). In experiment 2, 14 subjects were included in the analysis (all right handed; 5 female, 9 male; aged 20–37 years; mean age of 26 years; native language: 11 English, 2 German, 1 Spanish). All subjects were proficient in English and had been living in the United Kingdom for at least 3 years at time of testing. None of the subjects was trained in a tone language. Five additional subjects were excluded from experiment 2 to match behavioral performance for the different conditions across the group (supplemental Table S1, available at www.jneurosci.org as supplemental material). All subjects gave informed consent, and the experiment was performed with the approval of the Institute of Neurology Ethics Committee (London, UK). None of the subjects had any history of neurological or psychiatric disorder. All subjects reported having normal hearing, and they all had normal structural MRI brain scans.

Scanner setup

The stimuli were delivered using a custom electrostatic system at 70 dB SPL. After each syllable sequence, functional gradient-echo planar images (EPIs) were acquired [sparse imaging (Hall et al., 1999)] on a 3 T scanner (42 slices; −5° tilt; slice thickness of 2 mm, interslice distance of 1 mm; cardiac triggering; Siemens). Because of the cardiac triggering, there was a variable scan repetition time (time to repeat, 2.73 s + length of stimulus presentation + time to next pulse; time to echo, 65 ms). The 42 transverse slices of each brain volume covered the entire brain. The task instruction was presented during the last 10 slice acquisitions of each volume. It was followed by a fixation cross displayed during the subsequent stimulus sequence. Experiment 1 included 222 brain volumes for each subject (3 runs of 74 volumes each). Experiment 2 included 210 brain volumes for each subject (5 runs of 42 volumes each). Subjects were allowed to rest for several minutes between runs. The first two volumes were discarded from each run. Thus, there were 24 volumes for each of the eight experimental conditions plus 24 volumes for the silence condi-

tion in experiment 1. In experiment 2, there were 40 volumes for each of the four experimental condition plus 40 volumes for the silence condition.

Data analysis

The behavioral data were analyzed using SPSS 12.02.

Imaging data were analyzed using the statistical parametric mapping package (SPM5; http://www.fil.ion.ucl.ac.uk/spm). Scans were realigned, unwarped, and spatially normalized (Friston et al., 1995a) to Montreal Neurological Institute (MNI) standard stereotactic space (Evans et al., 1993) and spatially smoothed with an isotropic Gaussian kernel of 8 mm full-width at half-maximum.

Activity analyses

Statistical parametric maps were generated by modeling the evoked hemodynamic response for the different stimuli as boxcars convolved with a synthetic hemodynamic response function in the context of the general linear model (Friston et al., 1995b). Population-level inferences concerning BOLD signal changes between conditions of interest were based on a random-effects model that estimated the second-level *t* statistic at each voxel. To display common activations for similar contrasts in both experiments (see Figs. 2, 4), we entered the contrasts of interest (e.g., "speech task > loudness task" and "speech task > speaker task") for each subject in a second-level *t* statistic and performed a conjunction across the two contrasts.

Connectivity analyses (psychophysiological interactions)

Based on our activity results (see Results), the hypothesis was generated that left and right STG/STS are functionally connected when recognizing speech in the context of changing speakers. To test this hypothesis, we performed psychophysiological interaction analyses (PPI) (Friston et al., 1997). We selected the left posterior STG/STS as seed region of interest, identified at the individual subject level using the main effect of VTL (supplemental Table S2, available at www.jneurosci.org as supplemental material). Subjects for which a cluster of the main effect could be localized in left posterior STG/STS were included in the PPI analyses (experiment 1, $n = 17$; experiment 2, $n = 11$; *Z*-score >1.5). We extracted the first eigenvariate from these clusters (PPI-seed regions). PPI regressors were created using routines implemented in SPM5. The psychological variables were the interaction contrasts: (syllable task/VTL varies > syllable task/VTL fixed) > (loudness task/VTL varies > loudness task/VTL fixed) in experiment 1 and (syllable task/VTL varies > syllable task/GPR varies) > (speaker task/VTL varies > speaker task/GPR varies) in experiment 2. 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.

Significance thresholds for fMRI data and anatomical hypotheses

For each contrast, responses were considered significant at $p < 0.001$, uncorrected, if the localization of activity was in accordance with a priori anatomical hypotheses. Anatomical hypotheses for the main effect of VTL, main effect of task, and the interaction between the two were restricted to STG/STS based on previous studies on VTL processing (von Kriegstein et al., 2006, 2007). Hypotheses for the PPI interactions were restricted to right STG/STS. Based on studies investigating pitch processing with complex artificial nonspeech sounds, the hypothesis for BOLD responses related to pitch in the present experiment (GPR) were restricted to anterolateral Heschl's gyrus and planum polare (Griffiths et al., 2001; Patterson et al., 2002; Penagos et al., 2004; Bendor and Wang, 2005). Otherwise, responses were considered significant at $p < 0.05$ familywise error (fwe) corrected.

Regions are claimed to overlap if they adhere to both of the following criteria: (1) overlap on visual inspection given the above significance criteria and (2) activation by one contrast (i.e., speech task > control task) is significant in a region of interest defined by another contrast (i.e., VTL varies > VTL fixed) at $p < 0.05$ fwe corrected.

In the text, we only refer to activations that conform to these significance criteria. All other regions at $p < 0.001$ uncorrected are listed in the tables for the respective contrast.

To plot percentage signal changes for significant activations, we extracted the parameter estimates from the region of interest at the voxel in which we found the maximum value of the statistic. These values were then plotted using SPSS 12.02.

Results

We begin with the contrasts relevant to the hypothesis that posterior STG/STS is responsive to speaker-specific formant information (VTL) and at the same time contributes to speech recognition. These contrasts are partly also relevant to the second hypothesis, which is that VTL is processed in different regions from glottal fold parameters. This is tested by the contrast VTL varies > GPR varies and GPR varies > VTL varies in experiment 2 and by testing for differences in BOLD activation between conditions with the two glottal fold parameters voiced and whispered speech in experiment 1.

Speaker-related vocal tract changes are processed in posterior STG/STS: VTL varies > VTL fixed (experiment 1) and VTL varies > GPR varies (experiment 2)

In experiment 1, the contrast VTL varies > VTL fixed revealed BOLD responses along bilateral STG and STS (supplemental Table S3, Fig. S2, red, available at www.jneurosci.org as supplemental material). In experiment 2, the contrast VTL varies > GPR varies shows responses in left posterior STG/STS (supplemental Table S3, Fig. S3, red, available at www.jneurosci.org as supplemental material). Activation in the right posterior STG/STS only shows a trend to significance ($Z = 2.9$) (supplemental Table S3, available at www.jneurosci.org as supplemental material). A conjunction analysis involving these two contrasts is displayed in Figure 2 (red) showing the common activation in left posterior STG/STS [MNI coordinates: (−58, −50, 10) and (−60, −36, 10)].

In experiment 1, behavioral performance was better in conditions with fixed VTL compared with varying VTL (main effect of VTL, $F_{(1,17)} = 41$, $p < 0.0001$). In experiment 2, the behavioral performance was matched for the two conditions $F_{(1,13)} = 2.408$, $p = 0.15$ (Table 1).

Left posterior VTL-sensitive STG/STS is modulated by speech recognition: Speech task > loudness task (experiment 1) and speech task > speaker task (experiment 2)

There is more activation in STG/STS during the speech task than in the loudness task (experiment 1) (supplemental Fig. S2, green, available at www.jneurosci.org as supplemental material) and also in contrast to the speaker task (experiment 2) (supplemental Fig. S3, green, available at www.jneurosci.org as supplemental material). All activated regions are listed in supplemental Table S4 (available at www.jneurosci.org as supplemental material). A conjunction analysis of the two contrasts is displayed in Figure 2 (green) showing the common activation in left STG/STS. In both experiments, activation for the speech task, in contrast to the control task, overlaps with the main effect for VTL in left posterior STG/STS (Fig. 2) (supplemental Figs. S2, S3, green and red overlap, available at www.jneurosci.org as supplemental material).

At the behavioral level, in experiment 1, the speech task was easier than the control task, i.e., the loudness task ($F_{(1,17)} = 157$, $p < 0.001$) (Table 1). Behavioral performance in experiment 2 was the same for the syllable and the control task, i.e., the speaker task ($F_{(1,13)} = 0.246$, $p = 0.63$) (Table 1).

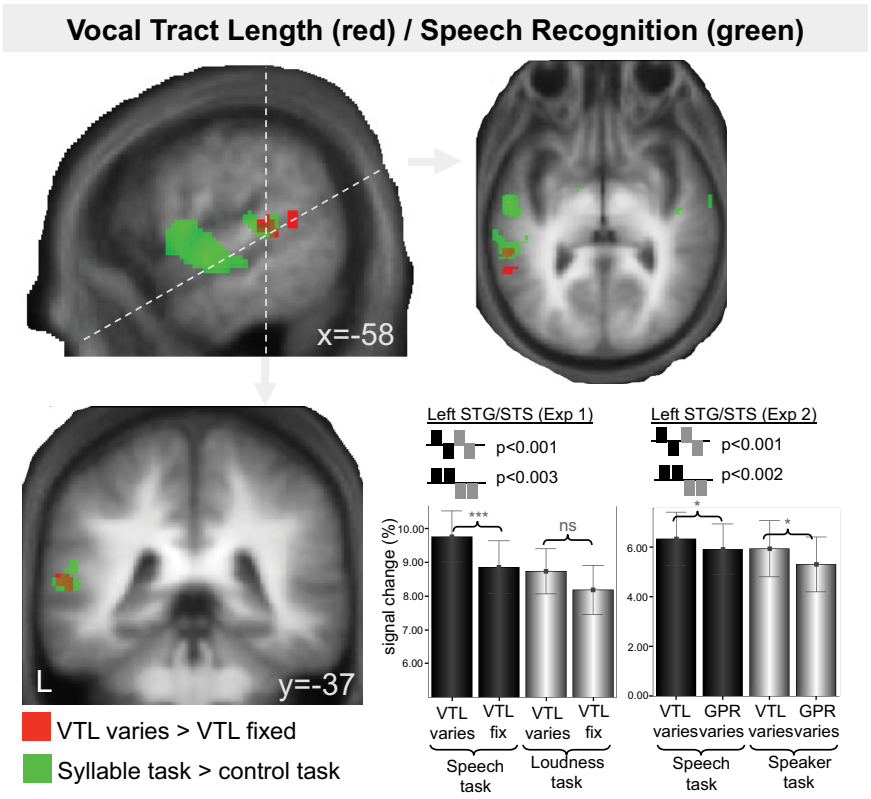

Figure 2. BOLD responses associated with the main effect of VTL (red) and main effect of task (green) as revealed by the conjunction analysis of experiment 1 and experiment 2. The group mean structural image is overlaid with the statistical parametric maps for the respective contrasts. "Control task" refers to loudness task in experiment 1 and to speaker task in experiment 2. L, Left hemisphere; VTL, acoustic effect of vocal tract length. The dotted lines on the sagittal section indicate the slices displayed as horizontal and coronal sections. The plots show the parameter estimates for experiments 1 and 2 separately. The small bar graphs on top of the plots display the main effects and their significance threshold in a repeated-measures ANOVA. Results of *post hoc t* tests are indicated by the brackets within the plot. $*p < 0.05$, $***p < 0.001$. ns, Nonsignificant. Error bars represent ±1 SEM.

Table 1. Behavioral results

	% correct	SEM
Experiment 1 ($n = 18$)		
Speech task		
VTL varies	86.62	1.15
Voiced	86.10	1.27
Whispered	87.14	1.29
VTL same	89.91	1.11
Voiced	89.23	1.31
Whispered	90.59	0.98
Loudness task		
VTL varies	74.27	1.09
Voiced	77.23	1.25
Whispered	71.30	1.23
VTL same	77.31	1.21
Voiced	81.22	1.23
Whispered	73.39	1.49
Experiment 2 ($n = 14$)		
Speech task		
VTL varies	91.77	1.4
GPR varies	91.31	1.5
Speaker task		
VTL varies	90.85	2.5
GPR varies	89.45	2.7

The table presents the percentage correct responses over the group for each condition of the experiments. Although there were performance differences in experiment 1 between the different conditions (see description in Results), experiment 2 was matched for behavioral performance across all conditions.

Right posterior STG/STS is modulated by speech recognition when speaker-related vocal tract parameters vary

Task × VTL interaction (experiment 1)

We tested the following interaction: (speech task/VTL varies > speech task/VTL fixed) > (loudness task/VTL varies > loudness task/VTL fixed). For this contrast, we found enhanced BOLD responses in right posterior STG/STS (Fig. 3, magenta) (supplemental Table S5, available at www.jneurosci.org as supplemental material).

At the behavioral level, there was no interaction between task and VTL ($F_{(1,17)} = 17$, $p = 0.73$).

In contrast to results of a previous fMRI study on speaker normalization (Wong et al., 2004), the right hemispheric activation in our study can be explained by neither increased task difficulty nor activity attributable to processing voice information per se (the stimulus input was exactly the same during the speech and loudness tasks). One potential reason, however, for differential responses in right STG/STS is an implicit processing of voice characteristics, i.e., subjects might automatically process the speaker characteristics of the stimulus significantly more during a speech task compared with a loudness task. In experiment 2, we control additionally for this potential confound by using a speaker task as control task.

Task × VTL interaction (experiment 2)

The speaker task in experiment 2 focuses attention explicitly on voice characteristics, which controls for implicit processing of voice characteristics during the speech task. We examined the interaction (speech task/VTL varies > speech task/GPR varies) > (speaker task/VTL varies > speaker task/GPR varies). For this contrast, there are enhanced BOLD responses in right posterior STG/STS in a similar location to the responses in experiment 1 (Fig. 3, cyan) (supplemental Table S5, available at www.jneurosci.org as supplemental material). At the behavioral level, there was no VTL × task interaction ($F_{(1,13)} = 1.4$, $p < 0.3$).

The right posterior STG/STS is commonly activated for the interactions in experiments 1 and 2 (supplemental Table S5, available at www.jneurosci.org as supplemental material). The conjunction analysis for the two experiments is displayed in Figure 4 (right panel, red). In addition, Figure 4 relates the current findings to findings of a previous study (von Kriegstein et al., 2007) (Fig. 4, blue). The contrast shows activations that are specific to VTL information in speech in contrast to similar acoustic changes in another communication sound, e.g., frog calls. VTL responsive regions in that study were located in right posterior STG/STS [(60, −42, −2), $Z = 3.12$; (60, −34, 4), $Z = 2.99$] (Fig. 4, right panel, blue) and in left posterior STG/STS [(−60, −32, 6), $Z = 3.74$; (−60, −48, 14), $Z = 3.23$] (Fig. 4, left panel, blue).

Responses in posterior STG/STS are similar for voiced and whispered speech

The percentage signal change in left posterior STG/STS is similar for the main effect VTL (VTL varies > VTL fixed) in voiced speech and in whispered speech (supplemental Fig. S4, red bars, available at www.jneurosci.org as supplemental material), and also for the main effect of task (speech task > loudness task) (supplemental Fig. S4, green bars, available at www.jneurosci.org as supplemental material). The percentage signal change for the interaction in right STG/STS is also similar for voiced and whispered speech (supplemental Fig. S4, magenta bars, available at www.jneurosci.org as supplemental material).

Functional connectivity between left and right posterior STG/STS is increased during speech recognition when speaker-related vocal tract parameters vary: PPI analyses: task × VTL interaction (experiments 1 and 2)

We tested the functional connectivity of left posterior STG/STS for the following psychological variables: (speech task/VTL varies > speech task/VTL fixed) > (loudness task/VTL varies > loudness task/VTL fixed) in experiment 1 and (speech task/VTL varies > speech task/GPR varies) > (speaker task/VTL varies > speaker task/GPR varies) in experiment 2. The analyses reveal that activity in VTL-sensitive left posterior STG/STS (seed region) (Fig. 5 red) has a stronger correlation to activity in right posterior STG/STS (target region) (Fig. 5, green) when recognizing speech from varying speakers than when recognizing speech from the same speaker. Importantly, this connectivity increase is specific to speech recognition in the context of changing speakers, because we use the task × VTL interaction as psychological variable in the PPI. Experiment 2 additionally shows that enhanced connectivity between left and right posterior STG/STS during speech recognition is attributable to speaker VTL changes rather than speaker GPR changes. In both experiments, the PPI target region is located in consistently close proximity posterior to regions showing enhanced activity in the task × VTL interactions [Fig. 5, magenta; the same contrast is also displayed in Fig. 3 (magenta, experiment 1; cyan, experiment 2)].

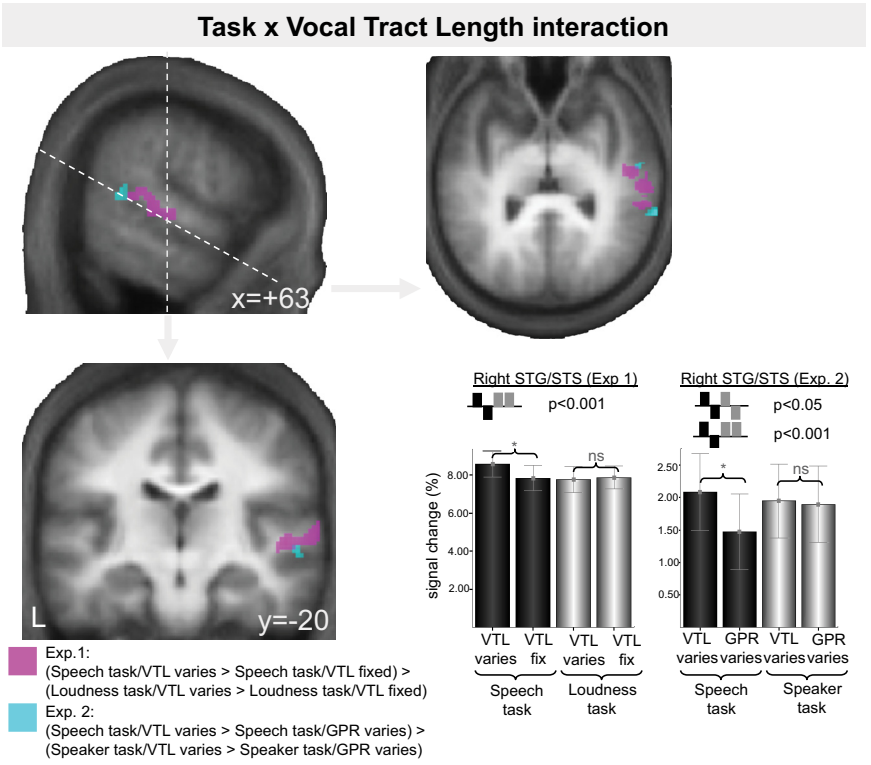

Figure 3. BOLD responses associated with the interaction between task and VTL. The contrast for experiment 1 is rendered in magenta and for experiment 2 in cyan. The plots show the parameter estimates for experiments 1 and 2 separately [MNI coordinates: experiment 1, (52, −22, 0); experiment 2, (68, −42, 16)]. The small bar graphs on top of the plots show the significant interaction and main effects and their significance threshold in a repeated-measures ANOVA. Results of *post hoc t* test are indicated by the brackets within the plot. $*p < 0.05$. ns, Nonsignificant. Error bars represent ±1 SEM.

Glottal fold parameters are processed along Heschl's gyrus

Voiced > whispered and whispered > voiced (experiment 1)

Contrasting all conditions containing voiced sounds with all conditions containing whispered sounds reveals an enhanced BOLD response adjacent to primary auditory cortex in anterolateral Heschl's gyrus (auditory cortex area Te1.2) (Morosan et al., 2001) of both hemispheres (Fig. 6, red) (supplemental Table S6, available at www.jneurosci.org as supplemental material). There was a significant location (Te1.1, Te1.2) × glottal fold parameter (voiced/whispered) interaction ($F_{(1,17)} = 123$, $p < 0.001$) (Fig. 6, plot).

Behavioral performance was better, on average, for voiced than whispered speech ($F_{(1,17)} = 38$, $p < 0.001$) (Table 1). This difference was attributable to performance differences in the loudness task [task × glottal fold parameter interaction, $F_{(1,17)} = 53$, $p < 0.0001$; *post hoc* paired *t* tests: whispered > voiced in the loudness task (size fixed, $t = 5.6$, $p < 0.0001$; size variable, $t = 4.8$, $p < 0.001$); whispered > voiced in the speech task (size fixed, $t = -1.9$, $p < 0.08$; size variable, $t = -0.7$, $p < 0.5$)]. We probed a task × glottal fold parameter interaction to check whether the pattern of BOLD responses in the main effect of glottal fold parameter is attributable to these differences in behavioral performance. There was no such interaction in Heschl's gyrus even at a low statistical threshold ($p = 0.05$ uncorrected). Furthermore, we found that contrasts, for which the behavioral performance is similar (speech task whispered > speech task voiced; speech task voiced > speech task whispered), reveal the same pattern of responses as the main effect in Te1.1 and Te1.2.

GPR varies > VTL varies (experiment 2)

The rate of glottal fold vibration (i.e., GPR) results in speech with different fundamental frequencies. This is heard as voices with different pitch. BOLD responses for contrasting all conditions in which GPR varies with all conditions in which VTL varies (GPR varies > VTL varies) partly overlap with those for the contrast voiced > whispered but extend farther along the superior temporal plane (Fig. 6, cyan). The behavioral performance is matched for this contrast (Table 1).

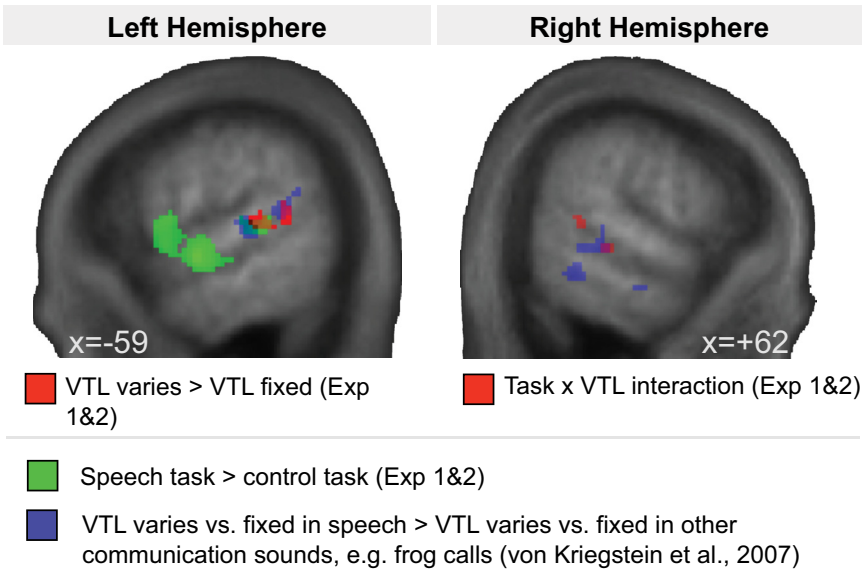

Figure 4. Overview of BOLD responses in right and left hemisphere. This figure also includes the BOLD responses reported in a previous study (von Kriegstein et al., 2007). The right-sided activation for the previous study is shown at a threshold of $p < 0.003$ for display purposes. The voxel with the maximum statistic for this study is at (60, −42, −2), $Z = 3.12$.

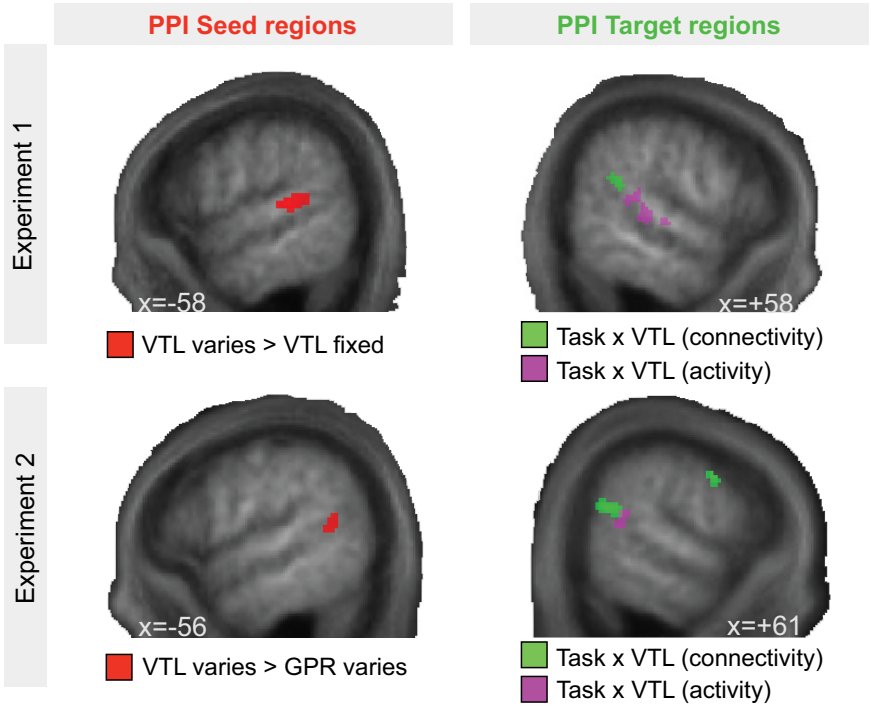

Figure 5. Functional connectivity (PPI) between left and right posterior STG/STS. Seed regions were taken from individual subject clusters; here the group mean is shown (red). Target regions identified by the PPI analysis (VTL × task, connectivity) are shown in green [MNI coordinates: experiment 1, (58, −46, 20), $Z = 3.03$; experiment 2, (60, −52, 20), $Z = 3.26$)]. BOLD responses associated with the interaction between task and VTL (VTL × task, activity) are displayed to demonstrate their consistently close proximity to PPI target regions in right posterior STG/STS.

Discussion

Our results show that speaker-related vocal tract parameters, which influence the formant position of the speech signal, are processed in posterior STG/STS. In contrast, speaker-related glottal fold parameters, which do not influence the formant position, are processed in areas immediately adjacent to primary auditory cortex, i.e., Te1.0 (Kaas and Hackett, 2000; Morosan et al., 2001). Vocal tract parameter-sensitive areas in posterior STG/STS are also involved in speech recognition. Left posterior STG/STS is (1) responsive to changes in vocal tract parameters (main effect of VTL) and (2) modulated by a speech recognition task (main effect of task). Right posterior STG/STS is modulated by the speech task only if vocal tract length varies but not if glottal fold parameters vary (VTL × task interaction). Functional connectivity between left and right posterior STG/STS is increased when recognizing speech from different speakers.

Representation of speaker-related acoustic variability

Vocal tract parameters

The experiments reported here are in accordance with previous studies investigating speaker-related vocal tract parameters (i.e., VTL) (von Kriegstein et al., 2006, 2007). For all studies, the maximum of BOLD responses to VTL changes occurs in left posterior STG/STS. In all studies, there is also activation in similar STG/STS areas of the right hemisphere at a relatively lower, sometimes nonsignificant, statistical threshold (for experiments 1 and 2, see Table S3, available at www.jneurosci.org as supplemental material) (for a previous experiment, see Fig. 4).

Glottal fold parameters

In voiced speech, the glottal pulse rate is perceived as voice pitch. Studies that contrast pitch-producing nonspeech sounds with spectrally matched noises reveal differential activation in anterolateral Heschl's gyrus (Te1.2) adjacent to primary auditory cortex (Griffiths et al., 2001; Patterson et al., 2002; Penagos et al., 2004). The differential activation to voiced over whispered speech overlaps with this putative pitch processing area (Fig. 6, red). Furthermore, differential activation in a region adjacent to lateral Heschl's gyrus (Te1.2) for GPR (pitch)-varying versus VTL-varying syllable sequences (Fig. 6, cyan) complements similar findings for artificial sounds (Patterson et al., 2002). These findings imply that voice pitch is processed in similar areas as the pitch of nonspeech sounds. The results are in accordance with the assumption that there are increasingly independent representations of pitch (here elicited by GPR) and timbre (here elicited by VTL) beyond primary auditory cortex (Nelken and Bar-Yosef, 2008; Bizley et al., 2009).

The inclusion of whispered conditions reveals a surprising result: whispered speech produces differential activation, not in primary auditory cortex but in regions immediately adjacent to it:

posteromedial Heschl's gyrus (Te1.1) (Fig. 6, yellow). In whispered speech, the constriction of the glottal folds produces noise that is, after passing through the vocal tract, perceived as whispered speech (Abercrombie, 1967). Previous studies involving noise bursts never found comparable effects (Griffiths et al., 2001; Patterson et al., 2002; von Kriegstein et al., 2006). Whispered speech in the current study is technically noise, but its spectrum contains formants. Because the voiced and whispered stimuli are resynthesized from the same recordings, the conditions are precisely matched with regard to spectral characteristics. We speculate that the activity in Te1.1 for whispered speech is not noise processing per se but noise that is specifically being processed as a communication signal.

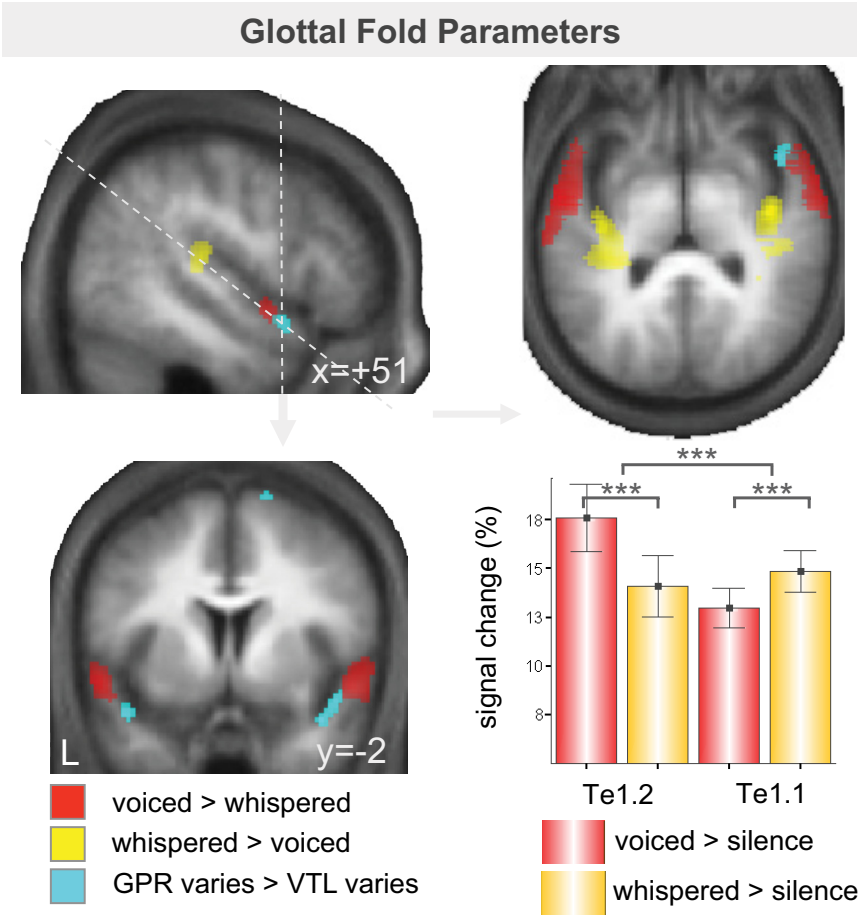

Figure 6. BOLD responses for voiced and whispered speech. The group mean structural image is overlaid with the statistical parametric maps for the contrasts between (1) voiced > whispered speech (red), (2) whispered > voiced speech (yellow), and (3) pitch varies > VTL varies (cyan). The plot shows parameter estimates for voiced and whispered speech in Te1.2 and Te1.1 (volume of interest). Error bars represent ±1 SEM. A repeated-measures ANOVA with the factors location (Te1.1, Te1.2) and sound quality (voiced, whispered) revealed a significant interaction of sound quality × location ($F_{(1,17)} = 28, p < 0.0001$), indicating differential responsiveness to whispered sounds in Te1.1 and to voiced sounds in Te1.2. ***$p < 0.001$.

Vocal tract information in speech recognition

One of the present key findings is that regions responding to changes in vocal tract length in posterior STG/STS are also involved in speech recognition.

There is neurophysiological evidence that relatively fast changing aspects of auditory input, which are relevant for speech perception, are processed in left-hemispheric temporal lobe areas, whereas slower changing information, e.g., identity, is predominantly processed in the right temporal lobe (Poeppel, 2003; von Kriegstein et al., 2003; Belin et al., 2004; Boemio et al., 2005; Giraud et al., 2007; Overath et al., 2007; Abrams et al., 2008; Lewis et al., 2009). In view of this dichotomy, an involvement of left-hemispheric areas in speech recognition, as found in the current study, is expected, but left-hemispheric processing of speaker-related vocal tract parameters is unexpected. Conversely, processing of speaker-related parameters in right STG/STS is expected, but involvement of right-hemispheric areas in speech recognition (compared with a high-level control condition) is surprising (Scott, 2005; Vigneau et al., 2006; Leff et al., 2008).

Why are regions in bilateral STG/STS responsive to changes in speaker-related vocal tract parameters and to speech recognition? It has been suggested that speech recognition involves both hemispheres but with computational differences between the hemispheres (Boatman et al., 1998; Hickok and Poeppel, 2000; Boatman, 2004; Boatman et al., 2006). The exact nature of these differences is unclear. In the following, we provide a speculative theoretical account that explains our results in terms of distinct but coupled mechanisms in the left and right hemisphere.

A potential mechanism for dealing with speaker-related vocal tract variability in speech recognition

In speech recognition, the brain needs to decode a fast-varying, information-rich, auditory input stream online. Theoretical accounts of brain function suggest that online recognition can be accomplished using dynamic models of the environment that predict sensory input (Knill et al., 1998; Friston, 2005; Kiebel et al., 2008). There is increasing evidence that the brain uses such a mechanism (Wolpert et al., 1995; Rao and Ballard, 1999; Bonte et al., 2006; Summerfield et al., 2006; von Kriegstein and Giraud, 2006; Overath et al., 2007). This scheme might be especially powerful if information changing at slower time scales predicts information at faster time scales (Kiebel et al., 2008; Balaguer-Ballester et al., 2009). For example, knowledge of the relatively constant vocal tract length of a speaker helps, among other constraints, to identify possible formant positions determining the phonemes of that speaker. Prediction of speech trajectories also implies that the dynamic uncertainty about speaker- and speech-related parameters is encoded. Dynamic uncertainty measures are valuable for online recognition because they preclude premature interpretations of speech input.

A prediction mechanism, which is based on knowledge about speaker characteristics, would prove useful in everyday conversational situations in which the speaker does not change rapidly. Such a scheme, which exploits the temporal stability of speaker parameters, would explain findings that speech from the same speaker is more intelligible than speech from changing speakers (Creelman, 1957; Mullennix et al., 1989; Pisoni, 1997). In this view, brain regions that encode speaker-specific parameters and dynamic uncertainty about these are critical in a speech recogni-

tion network (von Kriegstein et al., 2008). This would explain our findings (1) that bilateral posterior STG/STS is involved in both processing changes in vocal tract parameters and speech recognition and (2) that functional connectivity between left and right posterior STG/STS increases during speech recognition in the context of changing speakers.

We hypothesize that the right posterior STG/STS activation reflects the extraction of speaker parameters (here VTL), which is used by an internal model to effectively recognize the speech message. A change in VTL, during speech recognition, would prompt an adjustment of the vocal tract parameters of the internal model. This additional processing would not be necessary when VTL is fixed. We assume that VTL is just one of many speaker-related parameters that can be used to adjust an internal model. Other relevant parameters may include speaking rate (Adank and Devlin, 2010), visual information (e.g., face), and social information about the speaker (e.g., accent). Furthermore, in tone languages GPR changes are used to mark both message and speaker changes (Wong and Diehl, 2003). We speculate that, especially in these languages, GPR-sensitive regions in the right hemisphere provide information about the speaker-related variation of pitch to the left hemisphere.

In line with the assumption that left- and right-hemispheric function is specialized for distinct time scales in speech (Poeppel, 2003; Boemio et al., 2005; Giraud et al., 2007; Overath et al., 2008), we speculate that the left posterior STG/STS deals with vocal tract dynamics at a short time scale, e.g., at the length of one syllable or shorter. In this view, the main function of this area is not to determine the vocal tract parameters. Rather, left posterior STG/STS uses speaker-related vocal tract parameters, probably in part provided by right STG/STS, to represent fast vocal tract dynamics for speech recognition. Because certainty about VTL will lend more certainty to the representation of fast vocal tract dynamics, a sudden speaker change will lead to increased uncertainty about the fast speech dynamics. We assume that this burst in uncertainty triggers adjustment processes in the representation of fast speech dynamics, which explains why the speech-processing left STG/STS is also sensitive to speaker changes.

From the viewpoint of theoretical accounts of human and artificial speech processing, the proposed mechanism is a hybrid between abstract and exemplar models and captures the advantages of both (1) the ability to extract abstract features from the input and (2) the representation and use of speaker-related details during speech recognition. Such a scheme, which integrates abstract and exemplar approaches, may be implemented computationally using techniques as described previously (Kiebel et al., 2009). The location of such mechanisms in posterior STG/STS is in line with the implication of this region in phonological representation of speech sounds (Hickok and Poeppel, 2007; Desai et al., 2008; Leff et al., 2008), as well as the processing of visual vocal tract movements (Puce et al., 1998; O'Toole et al., 2002). We hypothesize that right and left posterior STG/STS encode speech features at various time scales and serve recognition by using a comprehensive, internal speech model that is updated during a speaker change.

References

Abercrombie D (1967) Elements of general phonetics. Edinburgh: Edinburgh UP.

Abrams DA, Nicol T, Zecker S, Kraus N (2008) Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. J Neurosci 28:3958–3965.

Adank P, Devlin JT (2010) On-line plasticity in spoken sentence comprehension: adapting to time-compressed speech. Neuroimage 49:1124–1132.

Adank P, van Hout R, Smits R (2004) An acoustic description of the vowels of Northern and Southern Standard Dutch. J Acoust Soc Am 116:1729–1738.

Ames H, Grossberg S (2008) Speaker normalization using cortical strip maps: a neural model for steady-state vowel categorization. J Acoust Soc Am 124:3918–3936.

Balaguer-Ballester E, Clark NR, Coath M, Krumbholz K, Denham SL (2009) Understanding pitch perception as a hierarchical process with top-down modulation. PLoS Comput Biol 5:e1000301.

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

Bendor D, Wang X (2005) The neuronal representation of pitch in primate auditory cortex. Nature 436:1161–1165.

Bizley JK, Walker KM, Silverman BW, King AJ, Schnupp JW (2009) Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. J Neurosci 29:2064–2075.

Boatman D (2004) Cortical bases of speech perception: evidence from functional lesion studies. Cognition 92:47–65.

Boatman D, Hart J Jr, Lesser RP, Honeycutt N, Anderson NB, Miglioretti D, Gordon B (1998) Right hemisphere speech perception revealed by amobarbital injection and electrical interference. Neurology 51:458–464.

Boatman DF, Lesser RP, Crone NE, Krauss G, Lenz FA, Miglioretti DL (2006) Speech recognition impairments in patients with intractable right temporal lobe epilepsy. Epilepsia 47:1397–1401.

Boemio A, Fromm S, Braun A, Poeppel D (2005) Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nat Neurosci 8:389–395.

Bonte M, Parviainen T, Hytönen K, Salmelin R (2006) Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. Cereb Cortex 16:115–123.

Creelman CD (1957) Case of the unknown talker. J Acoust Soc Am 29:655–655.

Deng L, Dong Y, Acero A (2006) Structured speech modeling. IEEE Trans Audio Speech Lang Processing 14:1492–1504.

Desai R, Liebenthal E, Waldron E, Binder JR (2008) Left posterior temporal regions are sensitive to auditory categorization. J Cogn Neurosci 20:1174–1188.

Evans AC, Collins DL, Mills SR, Brown ED, Kelly RL, Phinney RE (1993) 3D statistical neuroanatomical models from 305 MRI volumes. Proc IEEE Nucl Sci Symp Med Imag Conf 3:1813–1817.

Friederici AD (2002) Towards a neural basis of auditory sentence processing. Trends Cogn Sci 6:78–84.

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815–836.

Friston KJ, Ashburner J, Frith CD, Poline JB, Heather JD, Frackowiak RSJ (1995a) Spatial registration and normalisation of images. Hum Brain Mapp 2:165–189.

Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ (1995b) Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 2:189–210.

Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6:218–229.

Fujimura O, Lindqvist J (1971) Sweep-tone measurements of vocal-tract characteristics. J Acoust Soc Am 49:Suppl 2:541+.

Giraud AL, Kleinschmidt A, Poeppel D, Lund TE, Frackowiak RS, Laufs H (2007) Endogenous cortical rhythms determine cerebral specialization for speech perception and production. Neuron 56:1127–1134.

Goldinger SD (1996) Words and voices: episodic traces in spoken word identification and recognition memory. J Exp Psychol Learn Mem Cogn 22:1166–1183.

Griffiths TD, Uppenkamp S, Johnsrude I, Josephs O, Patterson RD (2001) Encoding of the temporal regularity of sound in the human brainstem. Nat Neurosci 4:633–637.

Hall DA, Haggard MP, Akeroyd MA, Palmer AR, Summerfield AQ, Elliott MR, Gurney EM, Bowtell RW (1999) "Sparse" temporal sampling in auditory fMRI. Hum Brain Mapp 7:213–223.

Hickok G, Poeppel D (2000) Towards a functional neuroanatomy of speech perception. Trends Cogn Sci 4:131–138.

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402.
Johnson K (2005) Speaker normalization in speech perception. In: The handbook of speech perception (Pisoni DB, Remez RE, eds), pp 363–389. Oxford: Blackwell Publishing.
Joos M (1948) Acoustic phonetics. Language 24:1–136.
Kaas JH, Hackett TA (2000) Subdivisions of auditory cortex and processing streams in primates. Proc Natl Acad Sci U S A 97:11793–11799.
Kawahara H, Masuda-Kasuse I, de Cheveigne A (1999) Restructuring speech representations using pitch-adaptive time-frequency smoothing and instantaneous-frequency-based F0 estraction: possible role of repetitive structure in sounds. Speech Commun 27:187–207.
Kawahara H, Irino T, Divenyi P (2004) Underlying prinicples of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. In: Speech separation by humans and machines, pp 167–180. Norwell, MA: Kluwer Academic.
Kiebel SJ, Daunizeau J, Friston KJ (2008) A hierarchy of time-scales and the brain. PLoS Comput Biol 4:e1000209.
Kiebel SJ, von Kriegstein K, Daunizeau J, Friston KJ (2009) Recognizing sequences of sequences. PLoS Comput Biol 5:e1000464.
Knill D, Kersten D, Yuille A, Richards W (1998) Introduction: a Bayesian formulation of visual perception. In: Perception as Bayesian inference, pp 1–21. Cambridge, UK: Cambridge UP.
Ladefoged P, Broadbent DE (1957) Information conveyed by vowels. J Acoust Soc Am 29:98–104.
Lavner Y, Gath I, Rosenhouse J (2000) The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. Speech Commun 30:9–26.
Leff AP, Schofield TM, Stephan KE, Crinion JT, Friston KJ, Price CJ (2008) The cortical dynamics of intelligible speech. J Neurosci 28:13209–13215.
Lewis JW, Talkington WJ, Walker NA, Spirou GA, Jajosky A, Frum C, Brefczynski-Lewis JA (2009) Human cortical organization for processing vocalizations indicates representation of harmonic structure as a signal attribute. J Neurosci 29:2283–2296.
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.
Mullennix JW, Pisoni DB, Martin CS (1989) Some effects of talker variability on spoken word recognition. J Acoust Soc Am 85:365–378.
Nearey TM (1989) Static, dynamic, and relational properties in vowel perception. J Acoust Soc Am 85:2088–2113.
Nelken I, Bar-Yosef O (2008) Neurons and objects: the case of auditory cortex. Front Neurosci 2:107–113.
Obleser J, Eisner F (2009) Pre-lexical abstraction of speech in the auditory cortex. Trends Cogn Sci 13:14–19.
O'Shaughnessy D (2008) Invited paper: automatic speech recognition: history, methods and challenges. Pattern Recognit 41:2965–2979.
O'Toole AJ, Roark DA, Abdi H (2002) Recognizing moving faces: a psychological and neural synthesis. Trends Cogn Sci 6:261–266.
Overath T, Cusack R, Kumar S, von Kriegstein K, Warren JD, Grube M, Carlyon RP, Griffiths TD (2007) An information theoretic characterisation of auditory encoding. PLoS Biol 5:e288.
Overath T, Kumar S, von Kriegstein K, Griffiths TD (2008) Encoding of spectral correlation over time in auditory cortex. J Neurosci 28: 13268–13273.
Patterson RD, Uppenkamp S, Johnsrude IS, Griffiths TD (2002) The processing of temporal pitch and melody information in auditory cortex. Neuron 36:767–776.
Penagos H, Melcher JR, Oxenham AJ (2004) A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. J Neurosci 24:6810–6815.
Pisoni DB (1997) Some thoughts on "normalization" in speech perception. In: Talker variability in speech processing (Johnson K, Mullenix JW, eds), pp 9–32. San Diego: Academic.
Poeppel D (2003) The analysis of speech in different temporal integration windows: cerebral lateralization as "asymmetric sampling in time." Speech Commun 41:245–255.
Puce A, Allison T, Bentin S, Gore JC, McCarthy G (1998) Temporal cortex activation in humans viewing eye and mouth movements. J Neurosci 18:2188–2199.
Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87.
Scott SK (2005) Auditory processing: speech, space and auditory objects. Curr Opin Neurobiol 15:197–201.
Summerfield C, Egner T, Greene M, Koechlin E, Mangels J, Hirsch J (2006) Predictive codes for forthcoming perception in the frontal cortex. Science 314:1311–1314.
Sussman HM (1986) A neuronal model of vowel normalization and representation. Brain Lang 28:12–23.
Turner RE, Walters TC, Monaghan JJ, Patterson RD (2009) A statistical formant-pattern model for estimating vocal-tract length from formant frequency data. J Acoust Soc Am 125:2374–2386.
Vigneau M, Beaucousin V, Hervé PY, Duffau H, Crivello F, Houdé O, Mazoyer B, Tzourio-Mazoyer N (2006) Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. Neuroimage 30:1414–1432.
von Kriegstein K, Giraud AL (2004) Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage 22:948–955.
von Kriegstein K, Giraud AL (2006) Implicit multisensory associations influence voice recognition. PLoS Biol 4:e326.
von Kriegstein K, Eger E, Kleinschmidt A, Giraud AL (2003) Modulation of neural responses to speech by directing attention to voices or verbal content. Brain Res Cogn Brain Res 17:48–55.
von Kriegstein K, Warren JD, Ives DT, Patterson RD, Griffiths TD (2006) Processing the acoustic effect of size in speech sounds. Neuroimage 32:368–375.
von Kriegstein K, Smith DR, Patterson RD, Ives DT, Griffiths TD (2007) Neural representation of auditory size in the human voice and in sounds from other resonant sources. Curr Biol 17:1123–1128.
von Kriegstein K, Dogan O, Grüter M, Giraud AL, Kell CA, Grüter T, Kleinschmidt A, Kiebel SJ (2008) Simulation of talking faces in the human brain improves auditory speech recognition. Proc Natl Acad Sci U S A 105:6747–6752.
Welling L, Ney H, Kanthak S (2002) Speaker adaptive modeling by vocal tract normalization. IEEE Trans Speech Audio Process 10:415–426.
Wolpert DM, Ghahramani Z, Jordan MI (1995) An internal model for sensorimotor integration. Science 269:1880–1882.
Wong PC, Diehl RL (2003) Perceptual normalization for inter- and intratalker variation in Cantonese level tones. J Speech Lang Hear Res 46:413–421.
Wong PC, Nusbaum HC, Small SL (2004) Neural bases of talker normalization. J Cogn Neurosci 16:1173–1184.