◆ Human Brain Mapping 38:2553–2565 (2017) ◆

# Activation in the Angular Gyrus and in the pSTS is Modulated by Face Primes During Voice Recognition

**Cordula Hölig,[1,2]* Julia Föcker,[3] Anna Best,[1] Brigitte Röder,[1] and Christian Büchel[2]**

[1]*Biological Psychology and Neuropsychology, University of Hamburg, Germany*
[2]*Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Germany*
[3]*Department of Psychology, Ludwig Maximilian University, Munich, Germany*

**Abstract:** The aim of the present study was to better understand the interaction of face and voice processing when identifying people. In a S1-S2 crossmodal priming fMRI experiment, the target (S2) was a disyllabic voice stimulus, whereas the modality of the prime (S1) was manipulated blockwise and consisted of the silent video of a speaking face in the crossmodal condition or of a voice stimulus in the unimodal condition. Primes and targets were from the same speaker (person-congruent) or from two different speakers (person-incongruent). Participants had to classify the S2 as either an old or a young person. Response times were shorter after a congruent than after an incongruent face prime. The right posterior superior temporal sulcus (pSTS) and the right angular gyrus showed a significant person identity effect (person-incongruent > person-congruent) in the crossmodal condition but not in the unimodal condition. In the unimodal condition, a person identity effect was observed in the bilateral inferior frontal gyrus. Our data suggest that both the priming with a voice and with a face result in a preactivated voice representation of the respective person, which eventually facilitates (person-congruent trials) or hampers (person-incongruent trials) the processing of the identity of a subsequent voice. This process involves activation in the right pSTS and in the right angular gyrus for voices primed by faces, but not for voices primed by voices. *Hum Brain Mapp 38:2553–2565, 2017.* © 2017 Wiley Periodicals, Inc.

**Key words:** crossmodal processing; voice recognition; person identity; face; multisensory; priming; fMRI

## INTRODUCTION

Human faces and voices are the most important sources of person information such as identity, emotion, age, and gender. Their recognition is essential for us to act adequately in social interactions. In everyday life, we typically encounter crossmodal stimulation as we see a face while simultaneously hearing the corresponding voice.

It has long been known that facial and vocal cues interact during speech perception [reviewed in Campbell, 2008; Navarra et al., 2012]. For instance, it has been demonstrated that seeing a talker's facial movements can facilitate auditory speech perception [Sumby and Pollack, 1954] and that incongruent visual lip movements can create illusionary auditory percepts [e.g. McGurk effect, McGurk and MacDonald, 1976].

Contract grant sponsor: Federal Ministry of Education and Research (to C.B. and B.R.); Contract grant number: G01GW0561; Contract grant sponsor: Deutsche Forschungsgemeinschaft (to C.B. and B.R.); Contract grant number: DFG SFB 936

*Correspondence to: Cordula Hölig, Biological Psychology and Neuropsychology, University of Hamburg, Von-Melle-Park 11, 20146 Hamburg, Germany. E-mail: cordula.hoelig@uni-hamburg.de

Received for publication 12 February 2016; Revised 23 December 2016; Accepted 6 February 2017.

DOI: 10.1002/hbm.23540

Published online 20 February 2017 in Wiley Online Library (wileyonlinelibrary.com).

Much later it has been documented that faces and voices interact extensively during the processing of person information. Voice recognition has been observed to be facilitated when a face with the same identity was simultaneously presented with the voice [Schweinberger et al., 2007, 2011] or prior to the voice [Ellis et al., 1997; Föcker et al., 2011; Stevenage et al., 2012]. Moreover, the learning of face-voice associations in comparison with voice-name associations has been reported to significantly enhance subsequent voice recognition [von Kriegstein and Giraud, 2006; Sheffert and Olson, 2004].

Traditional models of person perception have assumed that faces and voices are processed in separate brain pathways and integration occurs only in postperceptual processing stages [Bruce and Young, 1986]. However, more recent findings have suggested that faces and voice interact at early sensory processing stages during the integration of person information [Chandrasekaran et al., 2013; Föcker et al., 2011; González et al., 2011; Joassin et al., 2004; Schall et al., 2013; Schweinberger et al., 2011;]. Moreover, voice recognition has been shown to elicit activation in classical face-processing areas of the fusiform gyrus [von Kriegstein et al., 2005, 2008; von Kriegstein and Giraud, 2004, 2006; Schall et al., 2013].

Different experimental approaches have been employed to study the interactions between faces and voices with functional magnetic resonance imaging (fMRI). One is to compare brain activation elicited by unimodal stimuli (faces, voices) with brain activation elicited by bimodal stimuli (i.e. faces and voices are presented simultaneously) [Joassin et al., 2011; Watson et al., 2014a]. Using this approach, Joassin et al. [2011] reported person identity related activation in a cerebral network consisting of classical face-sensitive areas of the fusiform gyrus, of voice-sensitive areas along the middle superior temporal sulcus (STS), and of high-level supramodal brain regions such as the angular gyrus and the hippocampus. In contrast, Watson et al. [2014a] did not observe evidence for the integration of person-related information in lower-level visual or auditory areas, but identified the right posterior STS (pSTS) as an audiovisual region which specifically seemed to integrate person-related information. Interestingly, during emotion perception, the pSTS has repeatedly been reported to respond with a higher BOLD signal to the presentation of bimodal emotional stimuli (e.g. happy face paired with laughter) in comparison with purely auditory (e.g. laughter, scream) or face stimuli [Klasen et al., 2011; Kreifelts et al., 2007; Robins et al., 2009].

Experimental designs comparing unimodal stimulation with bimodal stimulation cannot always be unambiguously interpreted: on the one side, it is not clear whether the measured BOLD response is caused by genuine multisensory neurons or by a mixture of unisensory neurons selectively responding to different senses within the voxel [Goebel and van Atteveldt, 2009; Love et al., 2011]. On the other side, common activity arising for each stimulation condition enters the left side of the equation A + V = AV twice resulting in differences in activation which do not reflect genuine multisensory integration [see Gondan and Röder, 2006; James and Stevenson, 2012; Teder-Sälejärvi et al., 2002]. Therefore, different paradigms controlling for these factors need to be employed.

For instance, bimodal emotionally congruent stimuli (e.g., happy face paired with laugh) have been compared with bimodal emotionally incongruent stimuli (e.g. happy face paired with scream, [Dolan et al., 2001; Klasen et al., 2011; Müller et al., 2011]), assuming that only semantically congruent stimuli can be successfully integrated into one percept [Doehrmann and Naumer, 2008]. Bimodal emotionally incongruent stimuli elicited a higher BOLD signal than emotionally congruent stimuli in the angular gyrus, in the cingulate and in prefrontal areas [Klasen et al., 2011; Müller et al., 2011].

Another alternative are priming paradigms: It is a well-established finding that the repeated presentation of the same stimulus (or of the same stimulus attribute) causes the fMRI signal to decline in brain regions that process that stimulus or that stimulus attribute [Grill-Spector et al., 2006; Henson, 2003; Schacter and Buckner, 1998]. Priming effects for crossmodal prime-target combinations have previously been explored with fMRI [Adam and Noppeney, 2010; Blank et al., 2015; Noppeney et al., 2008; Tal and Amedi, 2009; Watson et al., 2014b], but the effects of face primes on voice recognition have not yet been investigated.

In the present study, we employed a S1-S2 priming paradigm to study multisensory interactions during person recognition. The S2 stimulus (the target) was a human voice in all conditions. The S1 stimulus (the prime) preceded the target stimulus and was a dynamic face (video) in the crossmodal condition and a human voice in the unimodal condition. We further manipulated the congruency between the prime and the target, that is, whether the prime and the target belonged to the same speaker (person-congruent) or to different speakers (person-incongruent). Based on the previous literature, we expected in the crossmodal condition a decrease in activation in same-speaker (person-congruent) compared with different-speaker (person-incongruent) trials in face-sensitive areas of the fusiform gyrus [Grill-Spector et al., 2004; Shah et al., 2001] and in voice-sensitive areas along the STS [Andics et al., 2013b; Joassin et al., 2011; Latinus et al., 2011] and the inferior frontal gyrus (IFG, [Andics et al., 2013a,b; Latinus et al., 2011]). We further predicted a similar decline of the BOLD signal in supramodal brain regions which have been previously reported to be activated during the integration of human faces and voices, specifically the pSTS [Blank et al., 2011; Joassin et al., 2011; Klasen et al., 2011; Watson et al., 2013, 2014a,b] and the angular gyrus [Joassin et al., 2011; Klasen et al., 2011; Müller et al., 2011].

## MATERIALS AND METHODS

### Participants

Nineteen university students participated in this study. Due to exceptional slow reactions times (more than three

standard deviations above the mean of the remaining participants), the data of one participant (male, 23 years, right-handed) was excluded from all analyses. All participants in the final sample (eight women, mean age: 24 years, age range: 19–31 years, 17 right handed) reported normal or corrected-to-normal vision and hearing. Written informed consent was given by each participant prior to the beginning of the experiment and all participants received monetary compensation for their participation. This study was approved by the ethic committee of the medical association of Hamburg.

## Experimental Design

### *Stimulus material*

Stimulus material was the same as in Föcker et al. [2011] and consisted of disyllabic German pronounceable pseudowords (baba, dede, fafa, lolo, sasa, wowo, babu, dedu, fafi, lolu, wowe) presented either visually (silent video of a speaking face) or auditorily (voice). We used pseudowords in order to single out voice identity effects by minimizing possible confounds related with real words (e.g. semantic associations, valence, familiarity). Dynamic (videos) instead of static faces were presented because of their higher ecological validity (in naturalistic settings, we see moving faces). Previous research has additionally demonstrated that voice recognition benefits more from simultaneously presented dynamic than static same identity faces and is significantly impaired by dynamic but not by static faces of other people [Schweinberger et al., 2007]. Moreover, in unimodal research, it has been shown that facial motion contributes significantly to the processing of face identity [Knappmeyer et al., 2003].

The pseudowords were spoken by 12 professional actors: three young women (mean age: 25 years, range: 23–27 years), three young men (mean age: 28 years, range: 26–29 years), three old women (mean age: 63 years, range: 61–64 years) and three old men (mean age: 66 years, range: 56–79 years). Each talker spoke all pseudowords. Talker's utterances were recorded in a sound-attenuated recording studio (Faculty of Media Technology at the Hamburg University of Applied Sciences) with a Neumann U87 microphone. Sound material was digitally sampled at 16 bit and offline equated for root mean square at 0.2 for presentation inside and at 0.025 for presentation outside the MR scanner. The mean duration of the auditory stimuli was 1,044 ms (range: 676 ms–1,406 ms). To guarantee a smooth onset of the voice stimulus, a 50 ms period of silence was added before the actor's voicing began. Actors were filmed with frontal view, outer facial features including hair and ears were covered (see example in Fig. 1). The video signal was digitally sampled at 16.66 frames per second with 24-bit resolution at 720 $\times$ 480 pixels. Videos were presented in grey scale. The mean duration of the video stimuli was 2,285 ms (range: 1,740 ms to 2,880 ms). A video file started with a motionless face for 720 ms (12 frames) before the face started moving while uttering the disyllabic pseudoword. It ended again with an image without any facial movements (closed mouth, 300 ms, five frames).

## Procedure

### *Experiment*

Within a S1-S2 paradigm (Fig. 1), we presented successively (1) the silent video of a speaking face and a voice stimulus (crossmodal condition) or (2) two voice stimuli (unimodal condition). Each trial began with a warning sound (550 Hz, duration = 100 ms). After 500 ms, the first stimulus was presented (S1, a silent video of a speaking face in the crossmodal condition, a voice in the unimodal condition) and after an interstimulus interval (ISI) of 700 ms, the second stimulus (S2, a voice in all conditions). The trial ended with the response of the participant, maximal 1,000 ms after the offset of the S2 voice. Each trial was followed by a 4 to 12 s rest period (mean: 8 s, uniform distribution). A white fixation cross was presented throughout scanning except during the presentation of the face stimuli.

As the S1 voice stimuli were of shorter duration than the S1 face stimuli (see Stimulus Material), periods of silence were added before and after the presentation of each S1 voice stimulus. The difference in duration was calculated individually for each face-voice pair. Of that time difference, 450 ms were added at the offset of the S1 voice stimulus; and the remaining time before the onset of the S1 voice stimulus (mean: 729 ms, range: 386–1,389 ms).

In 50% of the trials, S1 and S2 belonged to the same speaker (person-congruent trials); in the other 50% of the trials, S1 and S2 belonged to different speakers (person-incongruent trials) (Fig. 1). Participants indicated whether the S2 voice was from an old or from a young person. An orthogonal task instead of an explicit speaker identity matching task was used in order to dissociate the effect of person identity incongruency from response incongruency. Orthogonal tasks have been successfully employed in previous priming studies [Ellis et al., 1997; Henson, 2003; Noppeney et al., 2008]. Participants responded by pressing one of two buttons on a keypad with the index or the middle finger of the right hand. Response key assignments were counterbalanced across participants. In both modality conditions (unimodal and crossmodal), 48 person-congruent and 48 person-incongruent trials were presented resulting in 96 trials per modality and in a total number of 192 trials (standard trials). To guarantee attention to the S1 stimulus, 12 additional trials with deviant S1 stimuli (deviant trials, 11.1% of trials) were interspersed in each modality condition. Participants had to detect a deviant stimulus by pressing the button which was assigned to the index finger. The experiment was presented in four sessions (two crossmodal and two unimodal sessions); uni- and crossmodal sessions alternated. Half of the participants started with a unimodal session, the other half with a crossmodal session.

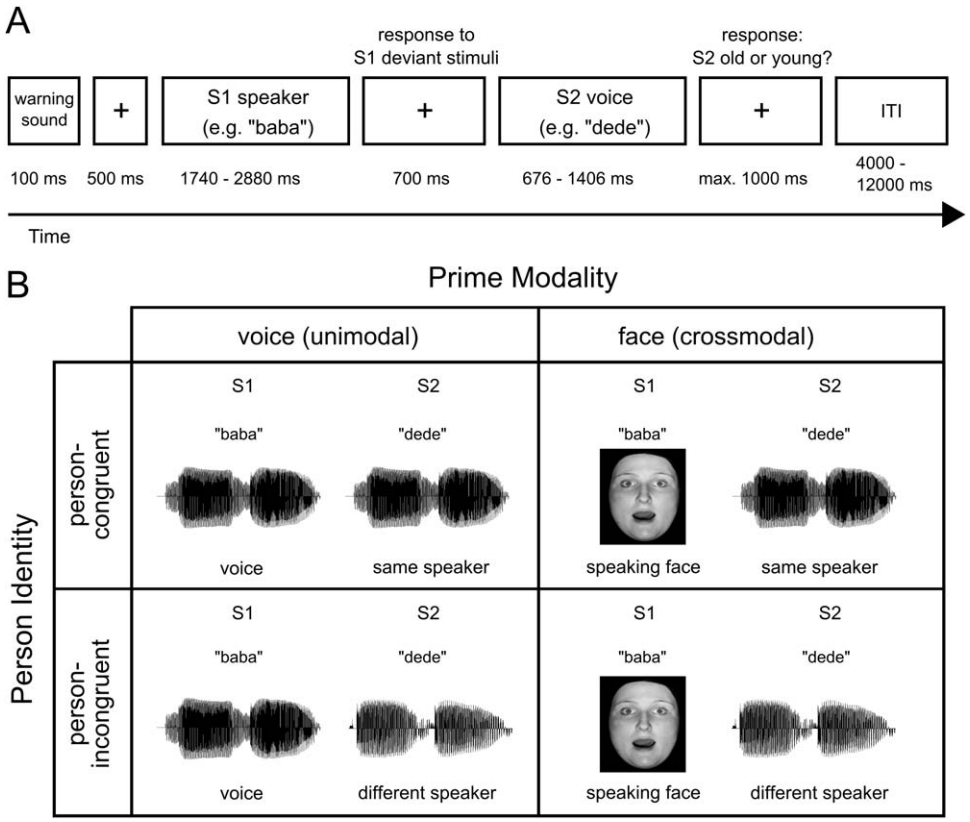


**Figure 1.**

Illustration of the experimental design. (**A**) Timing and tasks of the experiment. Two stimuli (disyllabic pseudowords) were successively presented. The second stimulus (S2) was always a voice. Participants indicated whether the S2 voice was from an old or from a young person. Additionally, participants had to detect deviant S1 stimuli (11.1% of all trials). ITI = intertrial-interval. (**B**) Conditions of the experiment. Two factors (prime modality, person identity) were manipulated within participants. Prime modality: the S1 stimulus was either a silent video of a speaking face (crossmodal condition) or a voice stimulus (unimodal condition). Person identity: in 50% of the trials, S1 and S2 belonged to the same speaker (person-congruent); in the other 50%, S1 and S2 belonged to different speakers (person-incongruent). To avoid physically identical voice pairs in the unimodal person-congruent condition (voice-voice), different pseudowords were used as S1 and S2 in all conditions, e.g. "baba" as S1 and "dede" as S2.

In standard trials, six pseudowords in which the first and second syllable were identical were presented (baba, dede, fafa, lolo, sasa, wowo). Deviant S1 voice stimuli (unimodal condition) consisted of pseudowords with two different syllables (babu, dedu, fafi, lolu, wowe) and deviant S1 face stimuli (crossmodal condition) of non-speaking faces (without any lip movements). To avoid physically identical voice pairs in the unimodal person-congruent condition (voice-voice), different pseudowords were used as S1 and S2 in all conditions, e.g. "baba" as S1 and "dede" as S2. Stimuli were presented in a pseudo-randomized order so that the same actor was never presented in two consecutive trials and deviant stimuli were separated by at least two standard stimuli. Overall, each actor was presented equally often as S1 and as S2 in both conditions. In person-incongruent trials, each speaker was paired once with a different speaker of the same age and gender, once with a different speaker of the same age but a different gender, once with a different speaker of a different age but the same gender and once with a different speaker of a different age and a different gender. Consequently, 50% of person-incongruent trials (i.e. 25% of the total trials) were gender-congruent (S1 and S2 same gender) and 50% (i.e. 25% of the total trials) were gender-incongruent (S1 and S2 different gender). Similarly, 50% of person-incongruent trials were age-congruent (S1 and S2 same age) and 50% were age-incongruent (S1 and S2 different age). Note that age-congruent trials were response-congruent (i.e. S1 primed response to S2) and age-incongruent trials response-incongruent (i.e. S1 did not prime response to S2). This procedure enabled us to

disentangle the effect of person identity from the effects of age, gender, and response.

### Training

Prior to the experiment, participants were familiarized with all face and voice stimuli presented in standard trials in multiple two hour training sessions. The training procedure was the same as in Föcker et al. [2011]. Initially, all stimuli were introduced and associated with a disyllabic proper name for each actor. In each trial, participants listened to an auditorily presented name which was followed by (1) the face or (2) one of six voice stimuli or (3) the bimodal face-voice combination of the corresponding actor. Participants were instructed to memorize all name-voice and name-face associations.

The main training consisted of two phases: a person identity training phase and a person identity matching phase. In the training phase, a face or a voice stimulus was presented and participants were asked to respond with the correct name of the actor. Feedback was provided after each response. Each training sequence consisted of 36 randomly presented stimuli (24 voices and 12 faces), in which each actor was presented three times. This training phase ended as soon as the participant reached the criterion of 85% correct responses (21 out of 24 voices, 10 out of 12 faces) in at least three consecutive training sequences.

In the matching phase, face and voice stimuli were presented within a S1-S2 paradigm. Each matching sequence consisted of 30 face-voice or voice-voice pairs of which 50% were person-congruent and 50% person-incongruent. In contrast to the main experiment, participants explicitly indicated whether the two stimuli belonged to the same or two different persons and received feedback after each response. Participants had to reach a criterion of 85% correct classifications in two successive blocks (26 out of 30 trials) to successfully terminate this training phase.

On the day of scanning, person identity recall and person identity matching performance were assessed again outside the scanner. Furthermore, participants were familiarized with the experiment prior to scanning.

## Data Acquisition

Functional magnetic resonance imaging (fMRI) data were acquired on a 3 T MR scanner (Siemens Magnetom Trio, Siemens, Erlangen, Germany) equipped with a 12-channel standard head coil. Thirty-six transversal slices (3 mm thickness, no gap) were acquired in each volume. A T2*-sensitive gradient echo-planar imaging (EPI) sequence was used (repetition time: 2.35 s, echo time: 30 ms, flip angle: 80°, field of view: 216 × 216, matrix: 72 × 72). A three-dimensional high-resolution ($1 \times 1 \times 1$ mm$^3$ voxel size) T1-weighted structural MRI was acquired for each subject using a magnetization-prepared rapid gradient echo (MP-RAGE) sequence. Face stimuli were projected onto a screen visible to the participant via a mirror mounted on the top of the head coil. Voice stimuli were presented via MR-compatible electrodynamic headphones (MR confon GmbH, Magdeburg, Germany, http://www.mr-confon.de). Sound volume was adjusted to a comfortable level for each participant prior to the experiment. This ensured that stimuli were clearly audible for all participants. Task presentation and recording of behavioral responses were conducted with Presentation software (www.neurobs.com).

## Data Analysis

### Behavioral data

On the day of scanning, we assessed for each participant the recognition rates for voices and faces (in %) and the response accuracy during crossmodal and unimodal person identity matching (in %). Means were statistically compared between modalities by paired *t*-tests.

In the main experiment, reaction times (RTs) were analyzed relative to the onset of the S2 voice stimulus for standard stimuli and relative to the onset of the S1 voice stimulus for deviant stimuli. Trials with incorrect responses or exceptionally fast (before voice onset) or slow responses (more than three standard deviations above a subject's mean in the respective condition) were excluded from all further analyses. For each participant, mean RTs and mean response accuracies were calculated separately for unimodal person-congruent trials, for unimodal person-incongruent trials, for unimodal deviant trials, for crossmodal person-congruent trials, for crossmodal person-incongruent trials and for crossmodal deviant trials. Conditions differences in reactions times and response accuracies were analyzed with two 2 × 2 ANOVA with the repeated measurement factors person identity (person-congruent vs. person-incongruent) and modality (unimodal vs. crossmodal) in standard trials. Mean response accuracies and RTs for deviants trials were statistically compared between modalities by two paired *t*-tests.

Additional analyses within person-incongruent trials were performed for each modality in order to investigate the effects of age and gender priming. Mean RTs and response accuracies were calculated separately for age-congruent and age-incongruent and likewise for gender-congruent and gender-incongruent trials for each participant and then compared by paired *t*-tests within each group and modality.

### fMRI data

Image processing and statistical analyses were performed with statistical parametric mapping (SPM 8 and SPM 12 software, Wellcome Department of Imaging Neuroscience, London, www.fil.ion.ucl.ac.uk/spm). The first five volumes of each session were discarded to allow for T1 saturation effects. Scans from each subject were realigned using the mean scan as a reference. Movement-by-susceptibility artefacts were corrected with the deformation fields implemented in the "realing and unwarp" function in SPM 8 [Andersson et al., 2001]. The structural T1 image was coregistered to the

mean functional image generated during realignment. The coregistered T1 image was then segmented into gray matter, white matter and CSF using the unified segmentation approach provided with SPM8 [Ashburner and Friston, 2005]. Functional images were subsequently spatially normalized to Montreal Neurological Institute space using the normalization parameters obtained from the segmentation procedure, resampled to a voxelsize of $2 \times 2 \times 2$ mm$^3$, and spatially smoothed with a 8 mm full-width at half-maximum isotropic Gaussian kernel.

Statistical analysis was performed within a general linear model (GLM). S1 and S2 stimuli were modeled at the onset of their presentation separately for the four conditions (crossmodal person-incongruent, crossmodal person-congruent, unimodal person-incongruent, unimodal person-congruent; only correct trials were included). The statistical model further included deviant trials and trials with incorrect responses (errors) as regressors of no interest. The duration of conditions was set to the length of the stimulus (in s) for face videos and to 0 s for voice recordings. All regressors were convolved with a hemodynamic response function (HRF). Potential baseline drifts in time series were corrected by applying a high-pass frequency filter (128 s). To analyze age and gender priming effects within person-incongruent trials, we set up two more models which were identical to the model described above except that we split the S2 person-incongruent regressor into a gender-congruent and a gender-incongruent regressor (model gender priming) and into an age-congruent and an age-incongruent regressor (model age priming).

In order to assess the effect of person identity within each modality, we compared brain responses in person-incongruent with brain responses in person-congruent trials (S2 person-incongruent > S2 person-congruent) separately for the crossmodal and the unimodal condition. We then tested for modality differences in the congruency effect with a modality (face prime > voice prime) × person identity (person-incongruent > person-congruent) interaction analysis. Additionally, we analysed the effect of prime modality independent of the factor person identity separately for prime (S1) and target (S2) stimuli, modality unspecific effects of person identity (S2 person-incongruent > S2 person-congruent), and age and gender priming effects within each region of interest.

For all reported analyses, we created appropriate contrasts at the subject level within SPM. Population-level inferences were based on a random-effects model that estimated the second-level t-statistic at each voxel. Activations at the group level were corrected for multiple comparisons using a family-wise error rate approach (FWE, $P < 0.05$). This approach was applied at the peak level on the whole brain volume or, for a priori defined regions of interest, on spherical volumes centered on coordinates reported in previous studies. In particular, spherical volumes consisted of a 15 mm radius sphere centered on [$x = 40$, $y = -46$, $z = -22$] for the fusiform gyrus [Grill-Spector et al., 2004; Shah et al., 2001] and on 10 mm radii spheres centered on [x = 60, y = −22, z = −2] for the middle STS [Andics et al., 2013b; Latinus et al., 2011], on [$x = 56$, $y = -48$, $z = 4$] for the pSTS [Klasen et al., 2011], on [x = 44, y = 22, z = 16] for the IFG [Andics et al., 2013a,b; Latinus et al., 2011], and on [$x = 52$, $y = -50$, $z = 34$] for the angular gyrus [Klasen et al., 2011; Müller et al., 2011]. Spherical volumes for all regions of interests were combined into a single mask for the small volume correction. Spatial references are reported in MNI standard space. Statistical maps are displayed on the MNI template thresholded at $P < 0.05$ (FWE) adjusted for the search volume.

## RESULTS

### Behavioral Data

On the day of scanning, mean recognition rates for both faces and voices were above 90% and higher for faces than for voices (faces: 99.5 ± 0.2% (mean ± SEM); voices: 91.8 ± 1.3%, $t_{(17)} = 5.80$, $P < 0.001$). Response accuracies in the person identity matching task were above 95% and did not differ between crossmodal and unimodal trials (crossmodal trials: 96.5 ± 0.9%; unimodal trials: 97.0 ± 1.1%, $t_{(17)} = 0.77$, $P = 0.454$).

In the main experiment, response accuracies were above 90% in all conditions (Fig. 2B). In both modalities, participants responded more accurately ($F_{(1,17)} = 6.00$, $P = 0.025$) and faster ($F_{(1,17)} = 36.92$, $P < 0.001$, Fig. 2A) in person-congruent than in person-incongruent trials. The RT difference between person-incongruent and person-congruent trials was larger in the unimodal than in the crossmodal condition (Fig. 2A, Modality by Person Identity interaction, $F_{(1,17)} = 8.00$, $P = 0.012$). RTs in person-incongruent trials did not differ between crossmodal and unimodal trials ($t_{(17)} = 0.07$, $P = 1$), but RTs in person-congruent trials were significantly faster in the unimodal condition than in the crossmodal condition ($t_{(17)} = 3.55$, $P = 0.005$). The main effect of Modality showed a trend to significance for RTs ($F_{(1,17)} = 3.06$, $P = 0.098$). No effects of Modality were observed for response accuracy (main effect of Modality: $F_{(1,17)} = 0.84$, $P = 0.371$; Modality by Person Identity interaction: $F_{(1,17)} = 2.67$, $P = 0.121$).

The detection rate of S1 deviants was above 95% in both modalities and significantly higher for S1 face deviants (mean detection rate of 99.6 ± 0.4%) than for S1 voice deviants (mean detection rate of 96.9 ± 1.2; $t_{(17)} = 2.16$, $P = 0.045$). Participants responded faster to S1 voice deviants (mean RT of 1337 ± 41 ms) than to S1 face deviants (1908 ± 105 ms; $t_{(17)} = 6.28$, $P < 0.001$).

To control for gender and age priming effects, we directly compared response accuracies and RTs between age-congruent and age-incongruent and between gender-congruent and gender-incongruent trials within person-incongruent trials. Both comparisons revealed no significant differences, neither in the unimodal condition (age RTs:

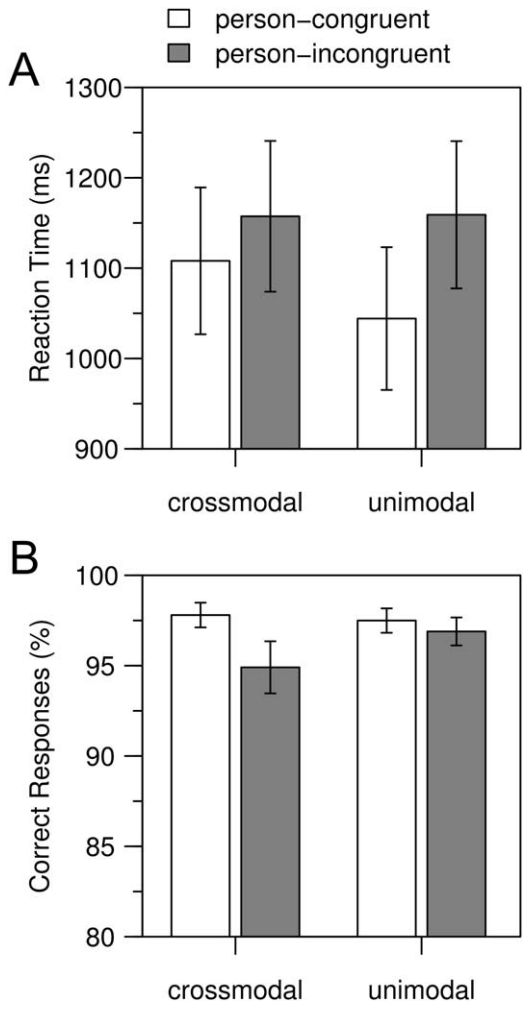


**Figure 2.**

Behavioral data. (**A**) Mean response times and (**B**) mean response accuracies are shown for each modality (crossmodal: face prime, unimodal: voice prime) and separately for person-congruent and person-incongruent trials. Response times were recorded from S2 onset onwards. Error bars indicate the standard error of the mean. In both modalities, participants responded faster and more accurately in person-congruent than in person-incongruent trials.

$t_{(17)} = 2.05$, $P = 0.056$, age response accuracy: $t_{(17)} = 0.30$, $P = 0.767$; gender RTs: $t_{(17)} = 1.44$, $P = 0.167$, gender response accuracy: $t_{(17)} = 0.84$, $P = 0.412$), nor in the crossmodal condition (age RTs: $t_{(17)} = 0.51$, $P = 0.615$, age response accuracy: $t_{(17)} = 0.35$, $P = 0.730$, gender RTs: $t_{(17)} = 0.60$, $P = 0.557$, gender response accuracy: $t_{(17)} = 0.12$, $P = 0.908$).

### fMRI Data

In the crossmodal condition, but not in the unimodal condition, person-incongruent S2 voices elicited a higher BOLD signal than person-congruent S2 voices in the right pSTS (peak coordinates $x\ y\ z$ in mm: 64 −42 4, $z = 3.75$, $P = 0.044$, FWE corrected for small volume) and in the right angular gyrus (58 −50 36, $z = 3.74$, $P = 0.044$, FWE corrected for small volume) (Fig. 3A). These results were confirmed by a significant Modality by Person Identity interaction in the right pSTS (60 −50 2, $z = 3.60$, $P = 0.024$, FWE corrected for small volume) and a marginally significant Modality by Person Identity interaction in the right angular gyrus (58 −50 36, $z = 3.35$, $P = 0.0505$, FWE corrected for small volume). The interaction was driven by person identity priming within the crossmodal condition: there was no significant activation for the contrast person-congruent > person-incongruent in the unimodal condition.

In the unimodal condition, but not in the crossmodal condition, person-incongruent S2 voices elicited a higher BOLD signal than person-congruent S2 voices in the bilateral IFG (pars triangularis, right peak: 54 32 14, $z = 5.52$, $P = 0.002$; left peak: −54 24 18, $z = 4.83$, $P = 0.044$, both FWE corrected for the whole brain; Fig. 3B). However, the Modality by Person Identity interaction did not reveal any significant activation in the IFG or any other brain region.

We further analysed the effect of prime modality on the processing of the target voice, irrespective of whether prime and target matched with regard to person identity. The BOLD signal following crossmodal S2 voices (i.e. voices primed by faces) was significantly higher than after unimodal S2 voices (i.e. voices primed by a voice) in the bilateral auditory cortex, in the right pSTS, in bilateral primary visual areas and in the right angular gyrus (Table I, Fig. 4A). The BOLD response following unimodal S2 voices was significantly higher than after crossmodal S2 voices in the bilateral fusiform gyrus, in primary visual areas (both due to a deactivation of the BOLD signal after crossmodal S2 voices) and in the right IFG (Table I, Fig. 4B).

Irrespective of prime modality, brain responses to person-incongruent S2 stimuli compared to person-congruent S2 stimuli were enhanced in the right IFG (52 28 18, $z = 5.67$, $P = 0.001$ FWE whole brain corrected). There were no significant voxels for the reverse contrast (person-congruent > person-incongruent). The comparison between S1 faces and S1 voices revealed modality-specific effects in sensory areas involved in the perceptual processing of visual and auditory information (Table II, Fig. 5). None of our regions of interest showed effects of gender or age priming.

## DISCUSSION

The aim of the present study was to identify the neural mechanisms of multisensory person identity processing. Voices primed by faces of the same person were faster and more accurately categorized as old or young than voices primed by faces of another person. Person identity processing elicited a higher BOLD response in the pSTS and in the angular gyrus after a crossmodal but not after a unimodal prime.

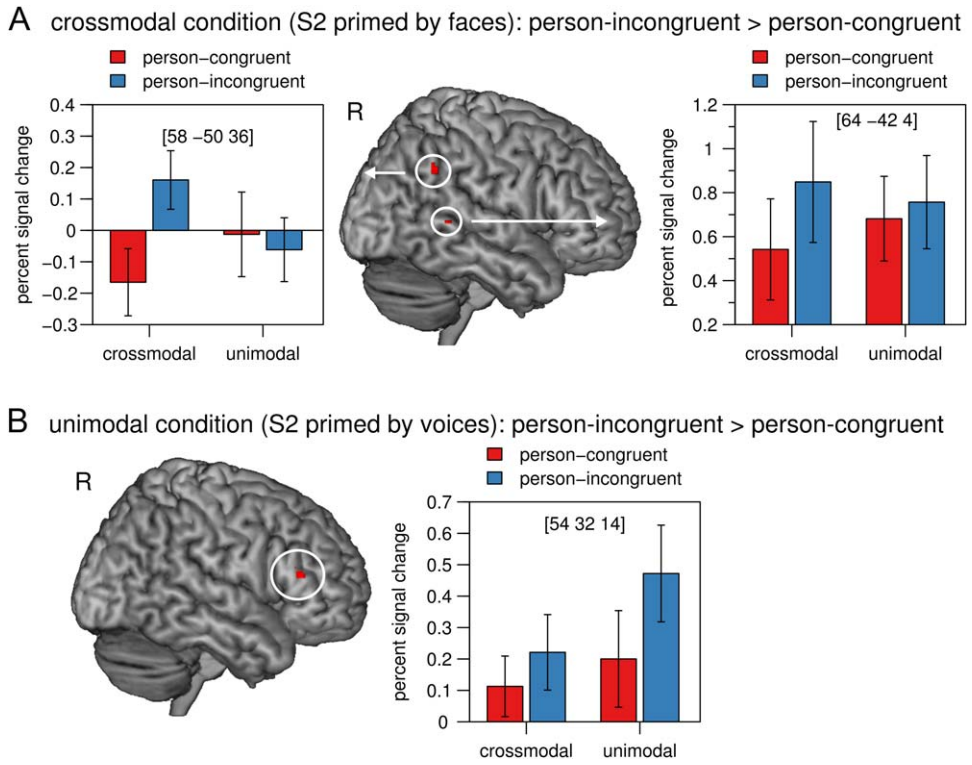


**Figure 3.**

(**A**) In the crossmodal condition (face primes), activation in the right pSTS and in the right angular gyrus was significantly higher in response to person-incongruent S2 stimuli than to person-congruent S2 stimuli. (**B**) In the unimodal condition (voice primes), activation in the bilateral IFG was significantly higher in response to person-incongruent S2 stimuli than to person-congruent S2 stimuli. The mean percent signal change of the peak voxel is plotted for each condition. Error bars indicate the standard error of the mean. Crossmodal = face prime, unimodal = voice prime, person-congruent = S1 and S2 same speaker, person-incongruent = S1 and S2 different speakers, L = left, R = right. [Color figure can be viewed at wileyonlinelibrary.com]

Our behavioral results replicate findings from two earlier studies using similar designs [Ellis et al., 1997; Föcker et al., 2011]. Behavioral priming effects of voice primes on face targets [Blank et al., 2015; Ellis et al., 1997] and face identity aftereffects caused by voice adaptors [Hills et al., 2010] have previously been reported, indicating bidirectional multisensory interactions. Crossmodal behavioral priming effects were further not restricted to the domain of person identity processing as they have been shown during the processing of audiovisual affective person information as well [Skuk and Schweinberger, 2013; Watson et al., 2014b].

Our results show an effect of face primes on voice identity processing in the pSTS and in the angular gyrus, which have both been shown to respond to different modalities in various tasks. The pSTS has been demonstrated to be involved in the integration of audiovisual information for a broad variety of stimuli [Beauchamp et al., 2004; Calvert et al., 2000; van Atteveldt et al., 2004;] including affective [Kreifelts et al., 2013; Watson et al., 2014b] and identity information [Joassin et al., 2011; Watson et al., 2014a] from faces and voices. Animal research has shown that the STS is directly connected with auditory and visual cortices [Seltzer and Pandya, 1994]; and comprises multisensory neurons which respond to both, auditory and visual stimulation [Barraclough et al., 2005; Perrodin et al., 2014]. In humans, activation in the pSTS has been observed during the processing of unimodally presented face stimuli [Baseler et al., 2014; Haxby et al., 2000] or voice stimuli [Andics et al., 2010]. Two studies have provided evidence for direct connections between FFA and pSTS [Baseler et al., 2014; Blank et al., 2011]. Based on these findings, we hypothesize that the pSTS receives information from face processing regions of the fusiform gyrus during the processing of face primes.

In contrast, there is no evidence for the angular gyrus to receive direct input from neither visual nor auditory cortices; and the angular gyrus has been shown to be anatomically connected mainly with other association areas [Binder et al., 2009; Seghier, 2013]. However, this region has been described to be responsive to multiple sensory modalities [Downar et al., 2001, 2002] and to be involved in the audiovisual

**TABLE I. Activation maxima for the comparison of crossmodal and unimodal S2 voices, irrespective of whether prime and target matched with regard to person identity**

| Region | MNI coordinates | | | z-value | $P < 0.05$, FWE corrected |
|---|---|---|---|---|---|
| | $x$ | $y$ | $z$ | | |
| S2 primed by faces > S2 primed by voices | | | | | |
| R Heschl Gyrus | 42 | −20 | 4 | 6.07 | Whole brain |
| L Heschl Gyrus | −46 | −14 | 12 | 6.03 | Whole brain |
| R temporal sulcus | 56 | −20 | 6 | 5.44 | Whole brain |
| L calcarine | −4 | −94 | 8 | 5.01 | Whole brain |
| R calcarine | 4 | −94 | 14 | 4.99 | Whole brain |
| R angular gyrus | 48 | −58 | 30 | 3.97 | Small volume |
| R angular gyrus | 52 | −60 | 34 | 3.90 | Small volume |
| S2 primed by voices > S2 primed by faces | | | | | |
| L fusiform gyrus | −36 | −64 | −14 | 5.78 | Whole brain |
| R fusiform gyrus | 30 | −46 | −20 | 5.66 | Whole brain |
| R fusiform gyrus | 38 | −64 | −14 | 5.31 | Whole brain |
| L calcarine | −10 | −98 | −6 | 4.88 | Whole brain |
| R calcarine | 22 | −88 | −8 | 4.82 | Whole brain |
| R inferior frontal gyrus | 52 | 22 | 22 | 4.47 | Small volume |

Coordinates are denoted by $x$, $y$, $z$ in mm (MNI space) and indicate the peak voxel. Strength of activation is expressed in $z$-scores. Only activations with $P < 0.05$ FWE corrected are listed.
L = Left, R = right.

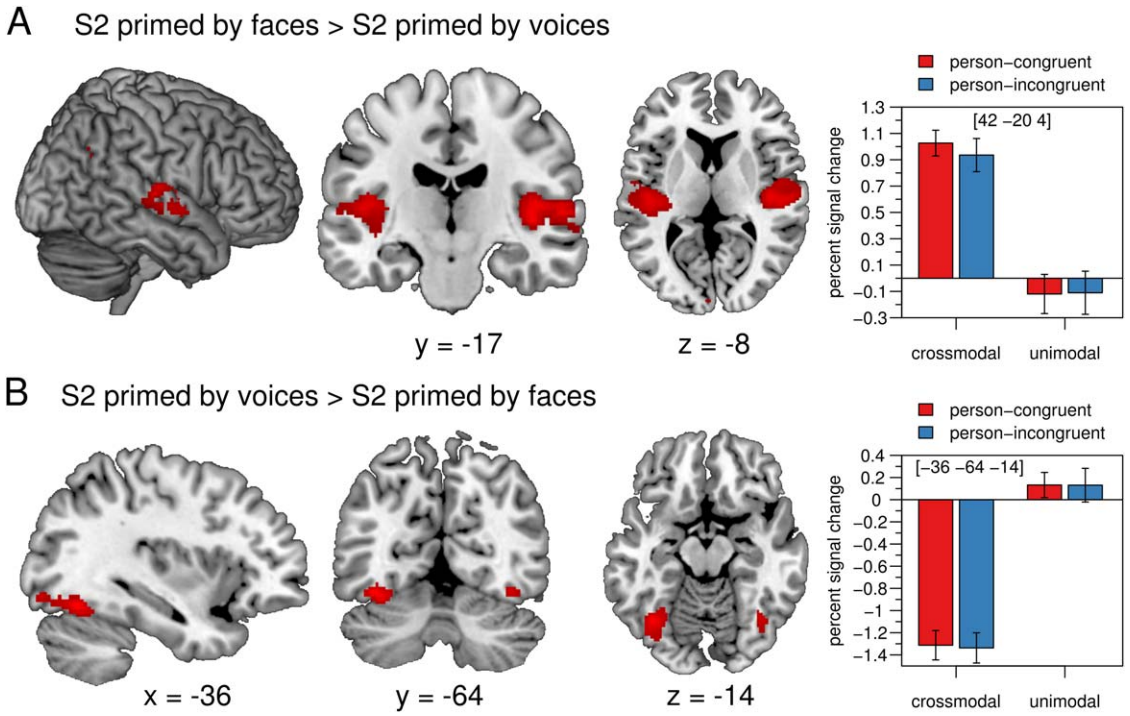


**Figure 4.**

(**A**) The BOLD signal was significantly enhanced in temporal auditory regions, in primary visual areas, and in the right angular gyrus for S2 faces compared with S2 voices. (**B**) The fusiform gyrus showed a significant decrease in activation in response to S2 voices primed by faces compared with S2 voices primed by voices. The mean percent signal change of the peak voxel is plotted for each condition. Error bars indicate the standard error of the mean. Crossmodal = face prime, unimodal = voice prime, person-congruent = S1 and S2 same speaker, person-incongruent = S1 and S2 different speakers, L = left, R = right. [Color figure can be viewed at wileyonlinelibrary.com]

**TABLE II. Activation maxima for the comparison of S1 faces and S1 voices, irrespective of whether prime and target matched with regard to person identity**

| Region | MNI coordinates | | | z-value | $P < 0.05$ FWE corrected |
|---|---|---|---|---|---|
| | x | y | z | | |
| S1: faces > voices | | | | | |
| R fusiform gyrus | 42 | −68 | −16 | 6.26 | Whole brain |
| L fusiform gyrus | −38 | −68 | −14 | 6.08 | Whole brain |
| R inferior occipital gyrus | 26 | −88 | −4 | 5.47 | Whole brain |
| L caudate | −12 | 0 | 24 | 5.12 | Whole brain |
| L fusiform gyrus | −28 | −44 | −16 | 4.99 | Whole brain |
| R middle occipital gyrus | 32 | −88 | 12 | 4.92 | Whole brain |
| R parahippocampal gyrus | 30 | −24 | −16 | 4.91 | Whole brain |
| L caudate | −20 | −4 | 24 | 4.87 | Whole brain |
| R parahippocampal gyrus | 34 | −30 | −14 | 4.87 | Whole brain |
| L calcarine | −4 | −94 | −8 | 4.84 | Whole brain |
| L fusiform gyrus | −32 | −54 | −16 | 4.83 | Whole brain |
| L fusiform gyrus | −22 | −96 | 4 | 4.83 | Whole brain |
| S1: voices > faces | | | | | |
| R Heschl Gyrus | 48 | −16 | 8 | 7.20 | Whole brain |
| L Heschl Gyrus | −38 | −26 | 8 | 6.85 | Whole brain |
| R temporal sulcus | 64 | −14 | −6 | 6.75 | Whole brain |
| L insula | −30 | 24 | 0 | 5.90 | Whole brain |
| R calcarine | 32 | −62 | 4 | 5.40 | Whole brain |
| R thalamus | 8 | −20 | 2 | 5.38 | Whole brain |
| R temporal sulcus | 56 | −38 | 4 | 5.32 | Whole brain |
| L thalamus | −8 | −20 | −2 | 5.23 | Whole brain |
| R cerebellum | 8 | −74 | −24 | 4.98 | Whole brain |
| R cerebellum | 22 | −54 | −28 | 4.84 | Whole brain |
| L inferior frontal operculum | −44 | 18 | 14 | 4.84 | Whole brain |
| R insula | 30 | 28 | −4 | 4.80 | Whole brain |
| R inferior frontal gyrus | 44 | 22 | 6 | 3.82 | Small volume |
| R inferior frontal gyrus | 44 | 18 | 8 | 3.70 | Small volume |
| R inferior frontal gyrus | 48 | 24 | 22 | 3.70 | Small volume |

Coordinates are denoted by *x*, *y*, *z* in mm (MNI space) and indicate the peak voxel. Strength of activation is expressed in *z*-scores. Only activations with $P < 0.05$ FWE corrected are listed.
L = Left, R = Right.

integration of speech [Bernstein et al., 2008] and person information [Joassin et al., 2011]. The angular gyrus is connected with the pSTS via the middle longitudinal fasciculus [Frey et al., 2008; Makris et al., 2009]. This opens the possibility that multisensory processing within the angular gyrus might be mediated by the pSTS. It might be further speculated that the angular gyrus, which is thought to play a role in knowledge retrieval [Binder et al., 2009; Binder and Desai, 2011], mediates the retrieval of the semantic representation of the person.

Our data suggest that both the priming with a voice and with a face results in the preactivation of the voice representation of the respective person. If the S2 voice stimulus matches the preactivated voice representation (person-congruent trial), the processing of the identity of the S2 voice stimulus seems to be facilitated (as suggested by the fastened response times and a reduced BOLD signal in person-congruent trials compared with in person-incongruent trials). This interpretation of the data is in accordance with the predictive coding framework [Egner et al., 2010; Friston, 2010; Rao and Ballard, 1999]. Predictive coding models suggest that the brain does not only passively processes incoming information but actively forms expectations about the nature of an incoming stimulus. These expectations, generated primarily in higher-level cortical areas, are thought to be compared with the actual sensory input. The difference is called "prediction error" and is coded in the brain response: the higher the difference between the expectation and the sensory information the higher is the BOLD signal. In the current study, the expectation may be reflected in the preactivated voice representation. The brain signal may be consequently enhanced for trials in which the S2 voice stimulus differs from the preactivated voice representation. Specifically the angular gyrus has recently been implicated in predictive coding to account for its multiple functions [Geng and Vossel, 2013; Seghier, 2013]. Supporting evidence comes from studies observing angular gyrus activation during predictive motor coding [Jakobs et al., 2009], during

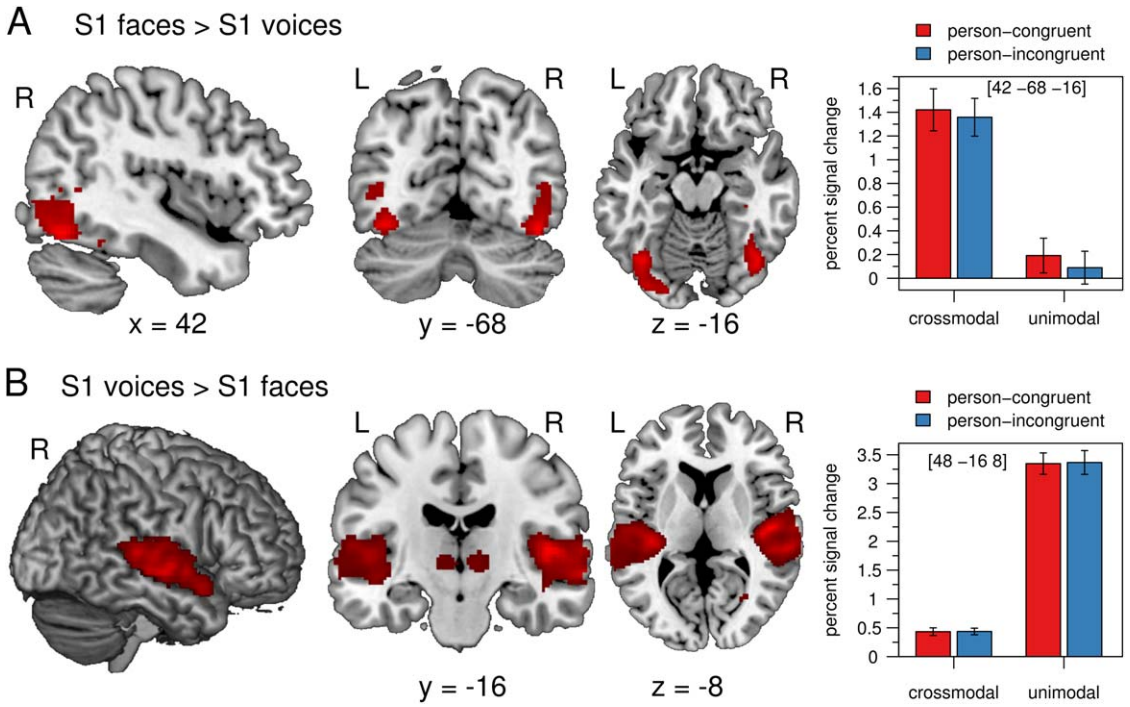


**Figure 5.**

(**A**) The BOLD signal was significantly increased in the fusiform gyrus and other regions of the occipital cortex for S1 faces compared with S1 voices. (**B**) Auditory regions in the temporal cortex showed significantly enhanced activation for S1 voices compared with S1 faces. The mean percent signal change of the peak voxel is plotted for each condition. Error bars indicate the standard error of the mean. Crossmodal = face prime, unimodal = voice prime, person-congruent = S1 and S2 same speaker, person-incongruent = S1 and S2 different speakers, L = left, R = right. [Color figure can be viewed at wileyonlinelibrary.com]

testing whether sensory input is relevant to one's own body [Tsakiris et al., 2008], during the violation of memory expectations [O'Connor et al., 2010], and during audiovisual incongruency processing [Klasen et al., 2011; Müller et al., 2011; Noppeney et al., 2008]. For example, emotional incongruent information (e.g. happy face paired with scream) elicited more activation than emotional congruent information (e.g. happy face paired with laugh) in the right angular gyrus [Klasen et al., 2011; Müller et al., 2011]. There is additional evidence for the STS to be implicated in predictive coding during audiovisual incongruency processing [Arnal et al., 2011; Lee and Noppeney, 2014]. For example, auditory leading and visual leading asynchronous speech compared to synchronous speech elicited activation in the STS [Lee and Noppeney, 2014].

Further evidence for this interpretation is provided by data from the unimodal condition. If voices were primed by voices, priming effects were observable in the IFG. Besides its relevance for voice identity processing [Andics et al., 2013a,b; Latinus et al., 2011; von Kriegstein and Giraud, 2004], the IFG has been reported to be involved in semantic retrieval [Badre et al., 2005; Badre and Wagner, 2007; Wagner et al., 2000] and cognitive control [Miller and Cohen, 2001]. Moreover, it has been shown to be modulated by stimulus expectations about voice identity [Andics et al., 2013a].

Taken together, we hypothesize that the BOLD signal difference between trials with person-congruent and person-incongruent primes might reflect the comparison between the preactivated voice representation and the actual sensory information. Our data suggest that this analysis involves activation in the pSTS and in the angular gyrus for voices primed by faces, but not for voices primed by voices. These findings indicate that the pSTS and the angular gyrus play an important role in mediating interactions between faces and voices during person recognition.

## ACKNOWLEDGMENTS

We thank Katrin Wendt and Kathrin Bergholz for their support in acquiring the fMRI data and Jürgen Finsterbusch for setting up the MR sequence. We are grateful to Boris Schlaack for his support to create the stimulus material and to Ulrike Adam, Kirstin Grewenig and Florence Kroll for helping to record the stimulus material supervised by Prof. Dr. Eva Wilk.

## REFERENCES


Adam R, Noppeney U (2010): Prior auditory information shapes visual category-selectivity in ventral occipito-temporal cortex. Neuroimage 52:1592–1602.

Andersson JLR, Hutton C, Ashburner J, Turner R, Friston K (2001): Modeling geometric deformations in EPI time series. Neuroimage 13:903–919.

Andics A, Gál V, Vicsi K, Rudas G, Vidnyánszky Z (2013a): FMRI repetition suppression for voices is modulated by stimulus expectations. Neuroimage 69:277–283.

Andics A, McQueen JM, Petersson KM (2013b): Mean-based neural coding of voices. Neuroimage 79:351–360.

Andics A, McQueen JM, Petersson KM, Gál V, Rudas G, Vidnyánszky Z (2010): Neural mechanisms for voice recognition. Neuroimage 52:1528–1540.

Arnal LH, Wyart V, Giraud A-L (2011): Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nat Neurosci 14:797–801.

Ashburner J, Friston KJ (2005): Unified segmentation. Neuroimage 26:839–851.

van Atteveldt N, Formisano E, Goebel R, Blomert L (2004): Integration of letters and speech sounds in the human brain. Neuron 43:271–282.

Badre D, Poldrack RA, Paré-Blagoev EJ, Insler RZ, Wagner AD (2005): Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. Neuron 47:907–918.

Badre D, Wagner AD (2007): Left ventrolateral prefrontal cortex and the cognitive control of memory. Neuropsychologia 45: 2883–2901.

Barraclough NE, Xiao D, Baker CI, Oram MW, Perrett DI (2005): Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. J Cogn Neurosci 17:377–391.

Baseler HA, Harris RJ, Young AW, Andrews TJ (2014): Neural responses to expression and gaze in the posterior superior temporal sulcus interact with facial identity. Cereb Cortex 24:737–744.

Beauchamp MS, Lee KE, Argall BD, Martin A (2004): Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41:809–823.

Bernstein LE, Auer ET Jr, Wagner M, Ponton CW (2008): Spatiotemporal dynamics of audiovisual speech processing. Neuroimage 39:423–435.

Binder JR, Desai RH (2011): The neurobiology of semantic memory. Trends Cogn Sci 15:527–536.

Binder JR, Desai RH, Graves WW, Conant LL (2009): Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex 19:2767–2796.

Blank H, Anwander A, von Kriegstein K (2011): Direct structural connections between voice- and face-recognition areas. J Neurosci 31:12906–12915.

Blank H, Kiebel SJ, von Kriegstein K (2015): How the human brain exchanges information across sensory modalities to recognize other people. Hum Brain Mapp 36:324–339.

Bruce V, Young A (1986): Understanding face recognition. Br J Psychol 77:305–327.

Calvert GA, Campbell R, Brammer MJ (2000): Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr Biol 10:649–657.

Campbell R (2008): The processing of audio-visual speech: empirical and neural bases. Philos Trans R Soc Lond B Biol Sci 363: 1001–1010.

Chandrasekaran C, Lemus L, Ghazanfar AA (2013): Dynamic faces speed up the onset of auditory cortical spiking responses during vocal detection. Proc Natl Acad Sci USA 110:E4668–E4677.

Doehrmann O, Naumer MJ (2008): Semantics and the multisensory brain: How meaning modulates processes of audio-visual integration. Brain Res 1242:136–150.

Dolan RJ, Morris JS, de Gelder B (2001): Crossmodal binding of fear in voice and face. Proc Natl Acad Sci USA 98:10006–10010.

Downar J, Crawley AP, Mikulis DJ, Davis KD (2001): The effect of task relevance on the cortical response to changes in visual and auditory stimuli: An event-related fMRI study. Neuroimage 14:1256–1267.

Downar J, Crawley AP, Mikulis DJ, Davis KD (2002): A cortical network sensitive to stimulus salience in a neutral behavioral context across multiple sensory modalities. J Neurophysiol 87: 615–620.

Egner T, Monti JM, Summerfield C (2010): Expectation and surprise determine neural population responses in the ventral visual stream. J Neurosci 30:16601–16608.

Ellis HD, Jones DM, Mosdell N (1997): Intra- and inter-modal repetition priming of familiar faces and voices. Br J Psychol 88: 143–156.

Föcker J, Hölig C, Best A, Röder B (2011): Crossmodal interaction of facial and vocal person identity information: An event-related potential study. Brain Res 1385:229–245.

Frey S, Campbell JSW, Pike GB, Petrides M (2008): Dissociating the human language pathways with high angular resolution diffusion fiber tractography. J Neurosci 28:11435–11444.

Friston K (2010): The free-energy principle: A unified brain theory? Nat Rev Neurosci 11:127–138.

Geng JJ, Vossel S (2013): Re-evaluating the role of TPJ in attentional control: Contextual updating? Neurosci Biobehav Rev 37:2608–2620.

Goebel R, van Atteveldt N (2009): Multisensory functional magnetic resonance imaging: A future perspective. Exp Brain Res 198:153–164.

Gondan M, Röder B (2006): A new method for detecting interactions between the senses in event-related potentials. Brain Res 1073–1074:389–397.

González IQ, León MAB, Belin P, Martínez-Quintana Y, García LG, Castillo MS (2011): Person identification through faces and voices: An ERP study. Brain Res 1407:13–26.

Grill-Spector K, Henson R, Martin A (2006): Repetition and the brain: Neural models of stimulus-specific effects. Trends Cogn Sci 10:14–23.

Grill-Spector K, Knouf N, Kanwisher N (2004): The fusiform face area subserves face perception, not generic within-category identification. Nat Neurosci 7:555–562.

Haxby JV, Hoffman EA, Gobbini MI (2000): The distributed human neural system for face perception. Trends Cogn Sci 4: 223–233.

Henson RNA (2003): Neuroimaging studies of priming. Prog Neurobiol 70:53–81.

Hills PJ, Elward RL, Lewis MB (2010): Cross-modal face identity aftereffects and their relation to priming. J Exp Psychol Hum Percept Perform 36:876–891.

Jakobs O, Wang LE, Dafotakis M, Grefkes C, Zilles K, Eickhoff SB (2009): Effects of timing and movement uncertainty implicate the temporo-parietal junction in the prediction of forthcoming motor actions. Neuroimage 47:667–677.

James TW, Stevenson RA (2012): The Use of fMRI to Assess Multisensory Integration. In: Murray, MM, Wallace, MT, editors. The Neural Bases of Multisensory Processes. Boca Raton (FL): CRC Press/Taylor & Francis, 131–146.

Joassin F, Maurage P, Bruyer R, Crommelinck M, Campanella S (2004): When audition alters vision: An event-related potential study of the cross-modal interactions between faces and voices. Neurosci Lett 369:132–137.

Joassin F, Pesenti M, Maurage P, Verreckt E, Bruyer R, Campanella S (2011): Cross-modal interactions between human faces and voices involved in person recognition. Cortex 47: 367–376.

Klasen M, Kenworthy CA, Mathiak KA, Kircher TTJ, Mathiak K (2011): Supramodal Representation of Emotions. J Neurosci 31: 13635–13643.

Knappmeyer B, Thornton IM, Bülthoff HH (2003): The use of facial motion and facial form during the processing of identity. Vision Res 43:1921–1936.

Kreifelts B, Ethofer T, Grodd W, Erb M, Wildgruber D (2007): Audiovisual integration of emotional signals in voice and face: An event-related fMRI study. Neuroimage 37:1445–1456.

Kreifelts B, Wildgruber D, Ethofer T (2013): Audiovisual integration of emotional information from voice and face. In: Belin P, Campanella S, Ethofer T, editors. Integrating Face and Voice in Person Perception. New York: Springer. pp 225–251.

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–e326.

von Kriegstein K, Dogan O, Grüter M, Giraud A-L, 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 USA 105:6747–6752.

Latinus M, Crabbe F, Belin P (2011): Learning-induced changes in the cerebral processing of voice identity. Cereb Cortex 21: 2820–2828.

Lee H, Noppeney U (2014): Temporal prediction errors in visual and auditory cortices. Curr Biol 24:R309–R310.

Love SA, Pollick FE, Latinus M (2011): Cerebral correlates and statistical criteria of cross-modal face and voice integration. Seeing Perceiving 24:351–367.

Makris N, Papadimitriou GM, Kaiser JR, Sorg S, Kennedy DN, Pandya DN (2009): Delineation of the middle longitudinal fascicle in humans: a quantitative, in vivo, DT-MRI study. Cereb Cortex 19:777–785.

McGurk H, MacDonald J (1976): Hearing lips and seeing voices. Nature 264:746–748.

Miller EK, Cohen JD (2001): An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202.

Müller VI, Habel U, Derntl B, Schneider F, Zilles K, Turetsky BI, Eickhoff SB (2011): Incongruence effects in crossmodal emotional integration. Neuroimage 54:2257–2266.

Navarra J, Yeung HH, Werker JF, Soto-Faraco S (2012): Multisensory interactions in speech perception. In: Stein, BE, editor. The New Handbook of Multisensory Processes. Cambridge, MA: MIT Press. pp 435–452.

Noppeney U, Josephs O, Hocking J, Price CJ, Friston KJ (2008): The effect of prior visual information on recognition of speech and sounds. Cereb Cortex 18:598–609.

O'Connor AR, Han S, Dobbins IG (2010): The inferior parietal lobule and recognition memory: Expectancy violation or successful retrieval? J Neurosci 30:2924–2934.

Perrodin C, Kayser C, Logothetis NK, Petkov CI (2014): Auditory and visual modulation of temporal lobe neurons in voice-sensitive and association cortices. J Neurosci 34:2524–2537.

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.

Robins DL, Hunyadi E, Schultz RT (2009): Superior temporal activation in response to dynamic audio-visual emotional cues. Brain Cogn 69:269–278.

Schacter DL, Buckner RL (1998): Priming and the brain. Neuron 20:185–195.

Schall S, Kiebel SJ, Maess B, von Kriegstein K (2013): Early auditory sensory processing of voices is facilitated by visual mechanisms. Neuroimage 77:237–245.

Schweinberger SR, Kloth N, Robertson DMC (2011): Hearing facial identities: Brain correlates of face–voice integration in person identification. Cortex 47:1026–1037.

Schweinberger SR, Robertson D, Kaufmann JM (2007): Hearing facial identities. Q J Exp Psychol 60:1446–1456.

Seghier ML (2013): The angular gyrus multiple functions and multiple subdivisions. Neuroscientist 19:43–61.

Seltzer B, Pandya DN (1994): Parietal, temporal, and occipita projections to cortex of the superior temporal sulcus in the rhesus monkey: A retrograde tracer study. J Comp Neurol 343: 445–463.

Shah NJ, Marshall JC, Zafiris O, Schwab A, Zilles K, Markowitsch HJ, Fink GR (2001): The neural correlates of person familiarity. A functional magnetic resonance imaging study with clinical implications. Brain 124:804–815.

Sheffert SM, Olson E (2004): Audiovisual speech facilitates voice learning. Percept Psychophysiol 66:352–362.

Skuk VG, Schweinberger SR (2013): Adaptation aftereffects in vocal emotion perception elicited by expressive faces and voices. PLoS One 8:e81691.

Stevenage SV, Hugill AR, Lewis HG (2012): Integrating voice recognition into models of person perception. J Cogn Psychol 24: 409–419.

Sumby WH, Pollack I (1954): Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215.

Tal N, Amedi A (2009): Multisensory visual-tactile object related network in humans: insights gained using a novel crossmodal adaptation approach. Exp Brain Res 198:165–182.

Teder-Sälejärvi WA, McDonald JJ, Di Russo F, Hillyard SA (2002): An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. Cogn Brain Res 14: 106–114.

Tsakiris M, Costantini M, Haggard P (2008): The role of the right temporo-parietal junction in maintaining a coherent sense of one's body. Neuropsychologia 46:3014–3018.

von Kriegstein K, Kleinschmidt A, Sterzer P, Giraud A-L (2005): Interaction of face and voice areas during speaker recognition. J Cogn Neurosci 17:367–376.

Wagner AD, Koutstaal W, Maril A, Schacter DL, Buckner RL (2000): Task-specific repetition priming in left inferior prefrontal cortex. Cereb Cortex 10:1176–1184.

Watson R, Latinus M, Charest I, Crabbe F, Belin P (2014a): People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. Cortex 50:125–136.

Watson R, Latinus M, Noguchi T, Garrod O, Crabbe F, Belin P (2013): Dissociating task difficulty from incongruence in face-voice emotion integration. Front Hum Neurosci 7:744.

Watson R, Latinus M, Noguchi T, Garrod O, Crabbe F, Belin P (2014b): Crossmodal adaptation in right posterior superior temporal sulcus during face-voice emotional integration. J Neurosci 34:6813–6821.