Interaction of Face and Voice Areas during Speaker Recognition

Katharina Von Kriegstein, Andreas Kleinschmidt, Philipp Sterzer, Anne-Lise Giraud

Interaction of Face and Voice Areas during Speaker Recognition

Katharina von Kriegstein, Andreas Kleinschmidt, Philipp Sterzer, and Anne-Lise Giraud

Abstract

■ Face and voice processing contribute to person recognition, but it remains unclear how the segregated specialized cortical modules interact. Using functional neuroimaging, we observed cross-modal responses to voices of familiar persons in the fusiform face area, as localized separately using visual stimuli. Voices of familiar persons only activated the face area during a task that emphasized speaker recognition over recognition of verbal content. Analyses of functional connectivity between cortical territories show that the fusiform face region is coupled with the superior temporal sulcus voice region during familiar speaker recognition, but not with any of the other cortical regions normally active in person recognition or in other tasks involving voices. These findings are relevant for models of the cognitive processes and neural circuitry involved in speaker recognition. They reveal that in the context of speaker recognition, the assessment of person familiarity does not necessarily engage supramodal cortical substrates but can result from the direct sharing of information between auditory voice and visual face regions. ■

INTRODUCTION

We can recognize people we know by seeing, hearing, touching, or even smelling them. Under normal circumstances, several of these person-specific attributes are simultaneously available to our senses, thus processed in parallel, and presumably associated to form unified supramodal stored representations of the individuals we know (Ellis, Jones, & Mosdell, 1997; Burton, Bruce, & Johnston, 1990; Bruce & Young, 1986). Our proficiency, in particular, for face and voice processing, is high enough that each of these attributes in isolation is normally sufficient to identify an individual familiar person, for example, faces on a photograph or voices on the phone. There is solid neurophysiological evidence that faces and voices are preferentially processed in distinct temporal lobe regions (von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003; Belin, Zatorre, & Ahad, 2002; Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Kanwisher, McDermott, & Chun, 1997; Puce, Allison, Gore, & McCarthy, 1995; Sergent, Ohta, & MacDonald, 1992). The fusiform face area (FFA) (Kanwisher et al., 1997) and the superior temporal sulcus (STS) voice regions (Belin, Zatorre, & Ahad, 2002; Belin, Zatorre, Lafaille, et al., 2000) have been established by comparing the neural responses associated with faces and voices with those when looking at objects or listening to environmental sounds, respectively. The counterpart of these anatomically segregated neurophysiological findings are neuropsychological syndromes where patients with lesions at different sites in the temporal lobe either fail to identify persons from their faces (prosopagnosia) (De Renzi, Perani, Carlesimo, Silveri, & Fazio, 1994; Damasio, Tranel, & Damasio, 1990) or their voices (phonagnosia) (Neuner & Schweinberger, 2000; Van Lancker, Cummings, Kreiman, & Dobkin, 1988).

Although anatomically segregated, voice- and face-processing modules are usually engaged in parallel and assumed to interact during person recognition (Ellis et al., 1997). Accordingly, models of person recognition include the possibility of reciprocal connections between attribute-specific perceptual modules. An influential model by Burton et al. (1990) assumes that the processing modules for names and faces converge onto a supramodal person identity node that performs a familiarity check, independent from person-related semantic processing. More recent models add a convergent path from voice-recognition units onto the person identity node (Ellis et al., 1997). These models hence posit supramodal nodes as the link between attribute-specific modules.

For speech perception, we have previously shown specific auditory-to-visual cross-modal effects in response to semantically meaningful stimuli (von Kriegstein et al., 2003; Giraud & Truy, 2002; Giraud, Price, Graham, Truy, & Frackowiak, 2001). Here, we tested for a corresponding cross-modal effect in the context of recognition of persons through voices and further investigated

Johann Wolfgang Goethe University, Germany

the underlying functional connectivity, that is, how areas are coupled to mediate cross-modal effects. In addition to the person recognition models discussed above, which include a supramodal node as an obligatory interface between face- and voice-recognition units, we considered an alternative model where attribute-specific modules can be directly and reciprocally functionally connected (Figure 1). This latter model accommodates the existing evidence that reciprocal interactions between the senses can be relayed through association cortices and do not necessarily involve supramodal feedback (Bavelier & Neville, 2002).

RESULTS

Using functional magnetic resonance imaging (fMRI), we first examined whether fusiform cortex is activated by voices of familiar persons as opposed to voices of nonfamiliar persons and assessed whether the fusiform cortex requires explicit person identification to be activated. Familiar speakers were personal acquaintances of the subjects (colleagues), whereas nonfamiliar speakers had never been encountered by the participants before the experiment. We compared responses during a speaker recognition task to those during a verbal task performed on the same voices of familiar persons. The effects of task (voice vs. verbal recognition) and familiarity (familiar vs. nonfamiliar speakers) were dissociated by virtue of a 2 × 2 factorial design (Figure 2, Table 1).

Voice recognition compared with verbal recognition performed on the same spoken material activated bilateral STS, preorbital and orbitofrontal cortices, superior parietal regions, the right temporal pole, and the cerebellum ($p < .05$, corrected). Recognition of voices of familiar compared with nonfamiliar persons further activated bilateral temporo-occipito-parietal (TOP), medial parietal/retrosplenial and anterior inferior temporal regions, and the fusiform cortex bilaterally, the latter with a right predominance (Figure 3A). The significance of fusiform activation by recognition of familiar persons' voices (compared with nonfamiliar persons' voices) was confirmed in a random effects model ($p < .001$, 42, −45, −21) as well as in single subjects. Two subjects activated the fusiform cortex only on the right side. The other seven showed bilateral responses.

Individual comparisons with the results from a separate localizer study involving visual stimuli (see Methods) revealed that the voice-induced effect overlapped or was located in very close proximity to responses to faces versus objects in all subjects. Figure 3B shows the responses in the fusiform region to both recognition of familiar persons' voices (individual thresholds, $p < .000$, two subjects; $p < .001$, two subjects; $p < .002$, one subject; and $p < .01$, one subject) and passive viewing of faces in the six subjects showing a significant response in the face localizer experiment. The overlap between the fusiform response to voices of familiar persons and the FFA was also reflected at the group-analysis level (Figure 3A) by contrasts of nonfamiliar faces with scrambled faces (group maximum at 42, −46, −30) or with objects (group maximum at 46, −44, −20), in accord with previous studies (Kanwisher et al., 1997). Interestingly, the group effect of the voice-induced response was best colocalized with the activation found when comparing faces of familiar versus nonfamiliar persons (group maximum at 40, −48, −24).

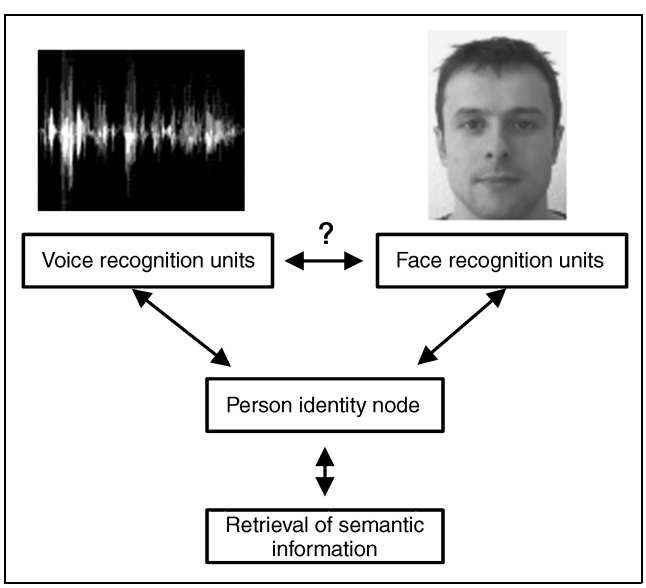

Figure 1. Excerpt of a person recognition model (Ellis et al., 1997; Burton et al., 1990). Arrows indicate reciprocal connections between levels of processing. The question mark indicates an alternative route of connection between voice- and face-specific areas.

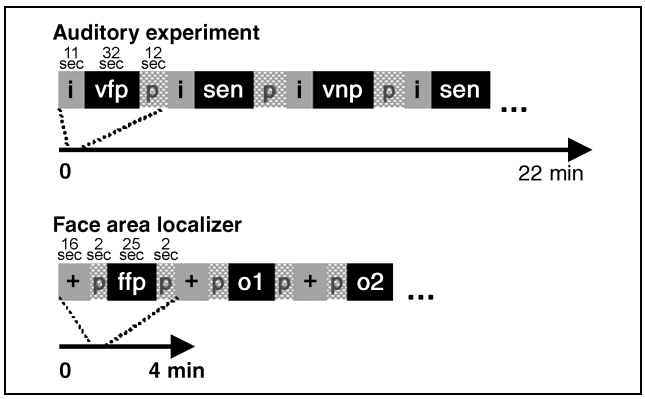

Figure 2. Experimental design. The auditory experiment comprised blocks of experimental conditions (vfp, vnp, cfp, cnp; see Methods) or blocks of speech envelope noises (sen) preceded by an instruction/target presentation (i). Each of the blocks was followed by a pause of 12 sec. The face area localizer comprised blocks of fixation (+) and blocks of experimental conditions (faces of familiar people [ffp], objects [o1/o2], faces of nonfamiliar people, scrambled versions of the faces). Each block was followed by a pause of 2 sec. Arrows display the length of one session.

Table 1. Local Response Maxima in SPM of Main Effects and Interaction

A. Contrast of Voice versus Verbal Content Recognition Independent of Familiarity

	Voice > Verbal Recognition: vfp + vnp > cfp + cnp			
Region	*x*	*y*	*z*	*Z*
Temporal				
Right pole/anterior STS	48	21	−18	7.3
Right STS middle/posterior	63	−42	−6	7.5
	63	*−27*	*0*	*6.5*
	66	*−18*	*3*	*6.4*
Left STS posterior	−63	−45	−3	5.5
Left STS middle	−60	−9	−3	5
	−51	*−15*	*−15*	*4.7*
Parietal				
Medial/superior parietal	3	−69	51	5
	0	*−60*	*42*	*4.8**
	9	*−63*	*39*	*4.7*
Superior right	36	−57	57	7.3
Superior left	−24	−54	51	5.3
Frontal				
Right inferior prefrontal lateral	45	30	18	8.8
	36	*18*	*18*	*7*
Right orbito-frontal	*36*	*33*	*−21*	*6.1*
	30	*21*	*−27*	*5.5*
Left inferior prefrontal lateral	−51	15	27	6.5
	−45	30	24	6.4
Left orbito-frontal	−33	27	−18	7.3
	−45	*21*	*−12*	*6.5*
	−33	*15*	*−12*	*6.5*
Cerebellum				
Left	−24	−51	−30	5.7
	−21	*−63*	*−21*	*5.6*
Right cerebellum/fusiform	48	−51	−30	4.2

Table 1. *continued*

B. Contrast of Familiar Speaker versus Nonfamiliar Speaker Independent of Task

	Familiar > Nonfamiliar Speaker: vfp + cfp > vnp + cnp			
Region	*x*	*y*	*z*	*Z*
Temporal				
Right anterior middle/inferior	60	−3	−30	7.5
Right pole	*51*	*15*	*−30*	*5.2*
Right temporo-parietal	54	−60	18	5.8*
Right fusiform	42	−45	−24	3.7*
Right amygdala/para-/hippocampus	27	9	−24	4
Right hippocampus	24	−15	−15	3.4
Left anterior middle/inferior	−60	−3	−24	5.6*
	−63	*−18*	*−21*	*4.9*
	−54	*3*	*−33*	*4.6*
Left temporo-parietal	−39	−60	18	6.4*
	−54	*−69*	*18*	*5.8**
Left fusiform	−36	−45	−30	3.5*
Left amygdala/para-/hippocampus	−30	0	−21	3.8
Medial parietal				
Precuneus/retrosplenial	6	−57	21	9.7

All response maxima are shown at $p < .001$, uncorrected, masked by vfp > sen. Asterisks (*) indicate a significant interaction of speaker familiarity and voice task. Abbreviations: vfp-recognition of voices (familiar person); vnp-recognition of voices (non-familiar person); cfp-recognition of verbal content (familiar person); cnp-recognition of verbal content (non-familiar person); *x*, *y*, *z* are the Talairach coordinates of the local maxima (in mm); *x*-medial-lateral axis; *y*-anterior-posterior axis; *z*-dorsal-ventral axis; *Z*-level of significance; STS-superior temporal sulcus; TOP-temporo-occipital-parietal junction; subclusters are displayed in *italics*.

The fMRI signal time course from the fusiform activation peak in the auditory study (group analysis) (Figure 3C) shows a strong response to voices of familiar persons during speaker's recognition, whereas during the verbal task, the response to the same stimuli did not significantly exceed the activity levels in the other tasks with recognition of noises or of voices of nonfamiliar persons. A positive interaction between task and person familiarity confirmed that the fusiform recruitment required both the voice of a familiar person and engagement in the voice-recognition task. Table 1 shows the results of tests for interaction in all regions related to the main effects.

The fusiform response profile to voices could not be accounted for by differences in behavioral responses to the conditions. Reaction times were shorter during the verbal task than during the voice-recognition task but were not affected by familiarity with the speakers for either type of task (Figure 4A). Although the fusiform region only showed a significant response during recog-

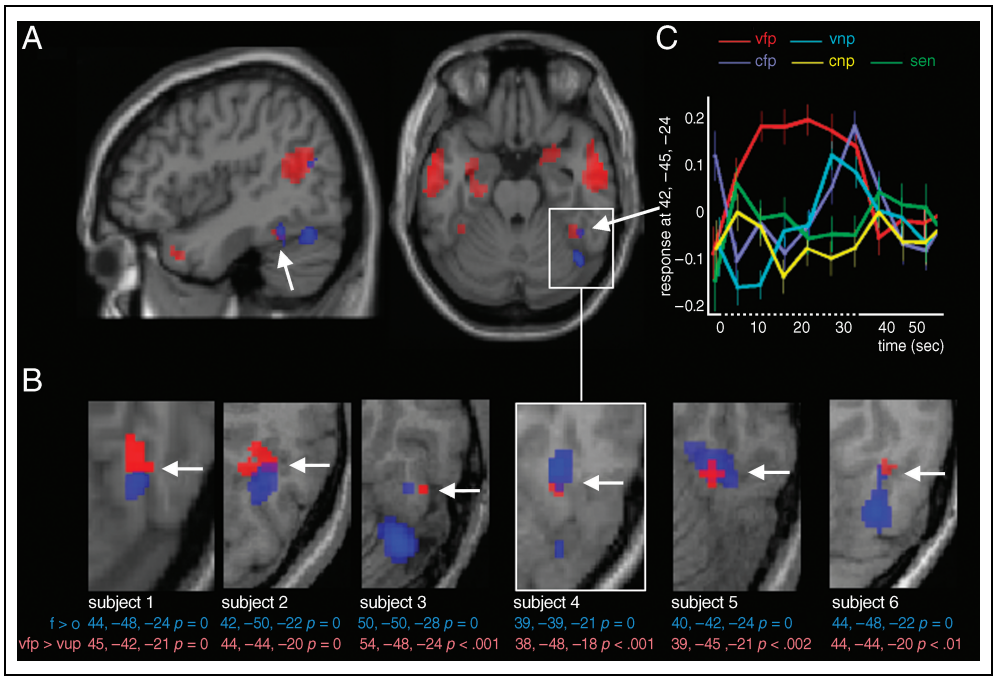

Figure 3. Activation of fusiform regions in the auditory experiment and the face area localizer study. (A) Group analysis. Contrast of familiar speaker versus nonfamiliar speaker (red), contrast of nonfamiliar faces versus objects (blue), $p < .001$, uncorrected, on sagittal and transversal sections. (B) Single-subject analyses. Contrast of recognition of familiar speakers' voices versus nonfamiliar speakers' voices (red) and faces versus objects (blue), results from six of the subjects displayed on details of appropriate transversal slices (as indicated in white frame of A) with coordinates and p values. (C) Time course of fMRI signal in the right fusiform region in response to the experimental conditions. Red = vfp, voice task (familiar); purple = cfp, verbal content task (familiar); cyan = vnp, voice task (nonfamiliar); yellow = cnp, verbal content task (nonfamiliar); green = sen, noise task (speech envelope noises); block length is displayed as dotted line on the x-axis.

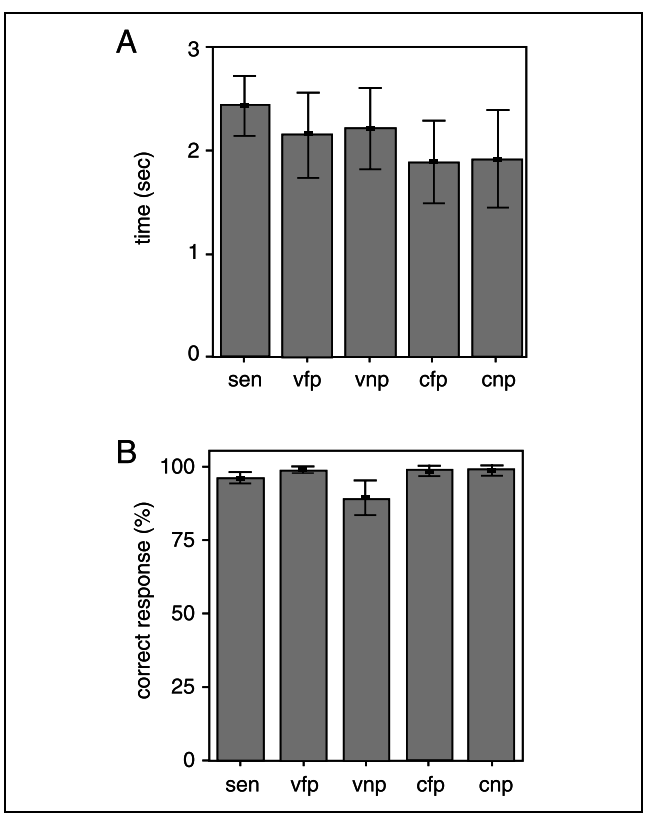

Figure 4. Behavioral data. (A) Response time in seconds. A repeated measure ANOVA indicated a significant task effect ($p = .001$). (B) Correct responses in different conditions. An ANOVA on repeated measure revealed an interaction of task and familiarity with a lower accuracy during recognition of voices of nonfamiliar persons ($p = .02$). For abbreviations see Figure 3.

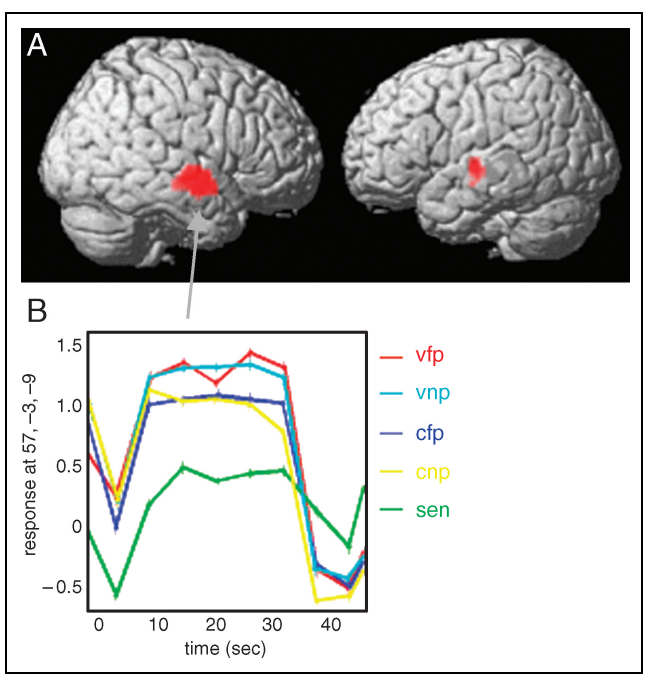

Figure 5. Connectivity analysis. (A) Cortical rendering of the brain regions functionally connected with the fusiform voice-responsive region during voice conditions with familiar speakers in contrast to conditions with voices of nonfamiliar speakers. The analysis reveals that the right STS interacts with the fusiform voice-responsive area ($p < .05$, corrected). (B) Time course of fMRI signal in the STS voice area in response to the experimental conditions. Note an effect of task (voice vs. verbal recognition) but not of familiarity of speakers.

nition of voices of familiar persons, the accuracy level was equally high for all tasks except for the recognition of nonfamiliar speakers (Figure 4B).

The cross-modal activation of the face region by familiar speakers' voices was the basis to delineate the potential neural circuits underlying voice-induced cross-modal effects in fusiform regions. As models of person recognition (Figure 1) propose a supramodal interface between sensory modules, we examined the functional connectivity pattern of the cross-modally recruited region in the fusiform cortex and tested whether this pattern was modulated by the cognitive task performed on voices (Friston et al., 1997) (see Methods). Although we consider the person recognition network as a whole, the way we performed functional connectivity analyses was constrained by the unimodal nature of our design and hence cross-modal nature of our observations. In particular, our reasoning was to take the fusiform voice-responsive region as a sample, instead of more liberally take the voice area as a starting point. Larger scale and less hypothesis-driven connectivity analyses could have been performed. However, when performing such analyses, we loose statistical power, because in an auditory study, activity in an auditory region shares its variance with the many other brain regions that get activated (bottom-up and top-down), whereas modulation of activity in the FFA should mostly reflect the modulations of the area(s) that provides its input.

The condition-specific connectivity analysis for familiarity and task showed increased correlation of the fusiform region only with bilateral middle/anterior STS (with a right predominance, Figure 5A, Table 2). The STS region targeted by the correlation analysis overlapped with the voice regions defined by the main effect of task (voice > verbal recognition, Table 1A). The time course of the response in the STS voice area that was identified by the correlation analysis revealed a task but no person familiarity effect (Figure 5B). Conversely, even at lower statistical thresholds, none of the regions that did respond to familiarity of speakers showed enhanced functional connectivity with the right fusiform region during recognition of familiar speakers. We confirmed this in a series of complementary tests on these regions (medial parietal/retrosplenial, bilateral anterior inferior/middle temporal, and bilateral TOP cortices). None of them showed a significant change of coupling with the right fusiform region (Table 2, Figure 6). Yet, all regions showed familiarity dependent correlation with medial parietal cortex and the voice-responsive STS. This suggests that, in our setting, the voice-responsive STS regions were functionally involved in two distinct interactions, one with the right FFA and the other with a person identity-retrieval network.

DISCUSSION

The main purpose of the present study was to detect and functionally characterize cross-modal neural activations of face-specific regions and their functional connections in the context of speaker identification. We replicated earlier findings that human voices are specifically processed along both STS (von Kriegstein et al., 2003; Belin, Zatorre, & Ahad, 2002; Belin, Zatorre, Lafaille, et al., 2000) and that the recognition of familiar

Table 2. Analyses of Condition Specific Functional Connectivity (Familiar Speakers) in Regions Showing a Significant Effect of Familiarity

		Target region						
Sampled region		*Right STS*	*Right fusiform*	*Medial parietal*	*Right temporo-parietal*	*Left temporo-parietal*	*Right anterior temporal*	*Left anterior temporal*
Right STS	63, −3, −9	–	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
Right fusiform	42, −45, −24	63, −3, −9	–	*ns*	*ns*	*ns*	*ns*	*ns*
Medial parietal	6, −57, 21	60, −6, −12	*ns*	–	60, −51, 9	−48, −51, 12	60, −3, −30	−60, −3, −24
Right temporo-parietal	54, −60, 18	60, −9, −12	*ns*	*ns*	–	*ns*	*ns*	*ns*
Left temporo-parietal	−39, −60, 18	60, −9, −12	*ns*	*ns*	*ns*	–	*ns*	*ns*
Right anterior temporal	60, −3, −30	57, −9, −12	*ns*	*ns*	*ns*	*ns*	–	*ns*
Left anterior temporal	−60, −3, −24	66, −21, −12	*ns*	*ns*	*ns*	*ns*	*ns*	–

Statistical threshold is $p < .05$, corrected. Numbers present the Talairach coordinates (x, y, z). *ns* = not significant.

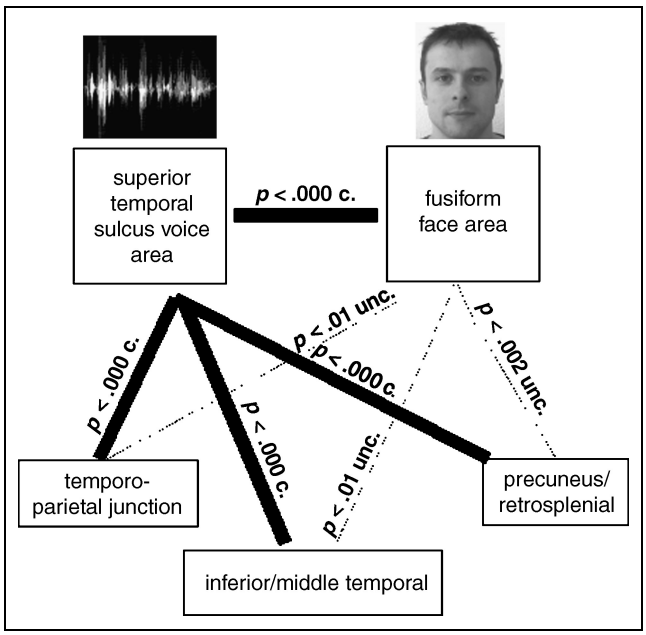

Figure 6. Summary of functional connectivity of the face and voice areas with regions showing a significant effect of familiarity. During speaker recognition, the voice area shows two types of interaction, one with the FFA and the other with supramodal familiarity responsive regions ($p < .000$ corrected). In contrast to that, the FFA interacts with the voice area ($p < .000$ corrected), but the interaction with supramodal regions is not significant. (c. = corrected; unc. = uncorrected).

speakers involves a distributed brain system, including bilateral medial parietal, TOP, and anterior temporal regions (Gorno-Tempini & Price, 2001; Shah et al., 2001; Leveroni et al., 2000; Gorno-Tempini, Price, Josephs, et al., 1998). This brain system is known to participate in episodic memory retrieval and is, therefore, probably related to person recognition in a nonspecific way (Cabeza & Nyberg, 2000).

Speech-driven cross-modal effects in face-specific visual areas have been previously observed in cochlear implant patients who, by experience, highly rely on visual phonological information (lip reading) to understand speech, but not in normal hearing controls (Giraud & Truy, 2002; Giraud, Price, et al., 2001). These data suggested that auditory-to-visual effects are not simply driven by stimulus material, for example, speech, but depend on task demands and expertise in linking visual and auditory input from speech. We confirmed this general notion in the present study by showing responses of the FFA to the voices of familiar persons only in a task that emphasized speaker recognition.

The fusiform response to voices was observed both at the group level and individually in each of the nine normal-hearing participants. This result confirms that speech-induced effects in visual areas are ''physiological'' and do not only occur after previous injury to a sensory modality (Giraud & Truy, 2002). Our findings indicate that cortical modules specialized for voices and faces, respectively, are coupled when one recognizes familiar persons from listening to their voices. The face area localizer experiment confirmed that the response to familiar persons' voices was located in regions specialized for face processing. It was colocalized with the anterior region of the two face responsive areas assessed by the contrast of faces and objects (Kanwisher et al., 1997). It has been proposed that the posterior face regions that fall within the region of the so-called lateral occipital complex encode the structure of faces and possibly achieve invariance (normalization for variable aspects such as expression, glasses, haircut, etc.) whereas more anterior regions (FFA) perform face recognition (Courtney, Ungerleider, Keil, & Haxby, 1997; Haxby, Horwitz, et al., 1994; Sergent et al., 1992; Haxby, Grady, et al., 1991) and familiar face processing (George et al., 1999). The bilateral activation in our study is in agreement with the assumption that the fusiform cortices from both hemispheres cooperate in face-selective processing (de Gelder & Rouw, 2001) and more specifically during the production of familiarity judgments on faces (Vuilleumier, Mohr, Valenza, Wetzel, & Landis, 2003).

One way to interpret the concurrent activation of the face and voice perceptual modules could be that there is a common input to the fusiform and the STS voice responsive areas from earlier processing stages. Such polymodal mechanisms have been proposed to explain some of the postlesional cross-modal effects (Bavelier & Neville, 2002). However, contrary to visually driven responses obtained in the same region for famous faces (George et al., 1999) that show typical stimulus dependency, the voice-induced activations in the fusiform cortex found here were not primarily related to a given sensory input (voices of familiar persons) but highly modulated by the task that emphasized voice over verbal content recognition. This observation suggests that the face area receives input from a region that in itself already shows a task effect, and thus speaks against a mere branching of bottom-up processing into both STS and fusiform cortex.

As effects of familiarity and expertise have been observed in the FFA (Gauthier, Skudlarski, Gore, & Anderson, 2000; Tarr & Gauthier, 2000), its response to voices of familiar persons could alternatively be seen as a supramodal effect resulting from convergence of inputs from segregated perceptual modules for faces and voices. Yet, neuropsychological reports do not reflect an association of prosopagnosia with phonagnosia that one would expect if the fusiform response to voices corresponded to a supramodal process. To the contrary, case descriptions of prosopagnosia emphasized that the patients affected relied on voices for compensation and thus maintained their capacity to identify friends and relatives (Pallis, 1955). Therefore, FFA activation by voices does not label the FFA as a

supramodal person region but rather indicates a cross-modal effect, that is, the recruitment of one specific sensory module through another modality.

The question remains to determine by which route the face-specific module is cross-modally activated by voices. One potential mechanism would involve a top-down influence from a supramodal relay as postulated by some models of person recognition (Figure 1). They assume a person identity node onto which unimodal information converges and where a familiarity check is performed (Schweinberger & Burton, 2003; Neuner & Schweinberger, 2000; Ellis et al., 1997; Burton et al., 1990). This type of connection would imply that during the processing of familiar persons' voices, activity in the voice-responsive fusiform region should share most of its variance with the supramodal relay node, and thus also, but more indirectly and hence to a lesser degree, with the voice-responsive STS regions. Interestingly, our functional connectivity analyses of the voice-responsive fusiform region and of the individually mapped FFA did not show this pattern. Instead, fusiform activity during familiar speaker's recognition correlated selectively with activity in the STS voice region.

The strength of functional connectivity between the fusiform and the STS region was modulated by both the familiarity of the speaker and the task. Conversely, the factor-related response profile of the STS region showed only an effect of task demand but not of speaker familiarity. Hence, the FFA response profile cannot simply be accounted for by a propagation of information from the STS to the fusiform region. In light of these findings, we propose that the familiarity effect in the fusiform region emerged from the coupling between the voice and the face area. This interpretation was further corroborated by extensive control analyses in which we found no evidence of an additional input to the fusiform region that could have generated its familiarity effect. Conceivably, such an input could have been expected from the candidate areas for a supramodal person identity node as postulated in previous models (Burton et al., 1990), that is, areas with larger responses to voices of familiar than nonfamiliar persons and common responses to faces and voices of familiar persons. Our results do not contradict the notion of a supramodal familiarity check but question whether in the context of speaker recognition this process has a dedicated (and segregated) cortical substrate. Instead, our data suggest that different aspects of person identity may be bound together through a supramodal "process." By *process*, we mean an interaction between unimodal perceptual modules that would not necessarily relay through a separate person identity node. The observation of familiarity responses from direct cross-modal coupling indeed suggests that familiarity assessment can already be performed at the perceptual stage. This, however, is not incompatible with the existence of supramodal regions involved in person-identification processing as proposed by the existing models, but it suggests that there might be an additional and earlier mode of coupling between modalities. A first level of familiarity check may take place at the sensory module level and hence work with some independence from the retrieval of semantic knowledge about the person identity. In this scheme, person identity nodes could correspond to convergent person-related semantic information. Equivalence between person identity nodes and semantic knowledge was in fact proposed in the original model of person recognition by Bruce and Young (1986).

What are the perceptual correlates of a cross-modal FFA activation by voices of familiar people? It has been established that the FFA responds to faces as a sensory stimulus (Kanwisher et al., 1997; Puce et al., 1995), to attending to faces (Wojciulik, Kanwisher, & Driver, 2000), to perceiving faces during binocular rivalry (Tong, Nakayama, Vaughan, & Kanwisher, 2000), and to imagery of faces (Ishai, Haxby, & Ungerleider, 2002; Ishai, Ungerleider, & Haxby, 2000; O'Craven & Kanwisher, 2000). As the first three mechanisms require the presence of a face in the sensory input, imagery of faces is the most likely perceptual correlate of FFA activation by voices (Ishai, Haxby, et al., 2002; Ishai, Ungerleider, et al., 2000; O'Craven & Kanwisher, 2000). However, different from the aforementioned studies where imagery was explicitly instructed by the task, we take our result to reflect the "implicit imagery" of a face when hearing the voice of a familiar person. Our results are not compatible with explicit visual imagery as a result of a top-down influence on the FFA from a supramodal region subsequent to person recognition through voices because there was no evidence for a primary input to the FFA from other cortical regions (Ishai, Haxby, et al., 2002; Ishai, Ungerleider, et al., 2000) than auditory. We, therefore, propose that if voices are processed to recognize speakers, this engenders an FFA activation that is routed via the voice area and induces face imagery as a perceptual consequence. Whether this mechanism of implicit imagery is corollary or contributes to speed and precision with which familiar persons can be recognized from their voices remains to be tested in future experiments.

In conclusion, our study demonstrates systematic recruitment of the fusiform cortex during recognition of familiar persons' voices. We show that (1) this effect occurs in areas responding to faces as sensory stimuli, (2) it reflects direct cross-modal cooperativity rather than supramodal convergence or relay, and (3) it is driven by active voice recognition rather than by mere familiarity of the speaker. We propose that (1) the right fusiform response to the voice of familiar persons is primarily driven by input from the right STS voice area, (2) the familiarity effect observed in the fusiform response results from a process that consists in the sharing of information between auditory and visual association cortices, and (3) the coupling between voice and face regions during speaker recognition does not require a

dedicated supramodal cortical locus where person related information converges to then be relayed back onto specific sensory modules.

METHODS

Subjects

Nine volunteers participated in the study (four women and five men; aged 27–36 years, with written informed consent in accordance with local Ethics Committee requirements). They were all right handed as determined by a modified version of the Edinburgh Inventory of Handedness (Oldfield, 1971), had normal hearing, and no history of neurological disease.

Auditory Experiment

We used 47 German sentences spoken by 14 unknown and 14 familiar speakers. Familiar speakers (and participants) were members of the clinical staff from the local neurology department. Nonfamiliar speakers were not known to the participants before the experiment neither by voice, face, or name. The standardization of stimuli was accomplished by using the same recording environment, software, and processing steps. Furthermore, the stimuli had the same linguistic content and the amount of sentences said by speakers with the same sex and age was matched.

Vocal stimuli were recorded (32-kHz sampling rate, 16-bit resolution), adjusted to the same overall sound pressure level, and processed using CoolEdit 2000 (Syntrillium Software, Scottsdale, Arizona, USA) and Soundprobe (Hisoft, Bedford, UK). Stimuli in the control conditions were speech envelope noises derived from the sentences (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Stimuli were delivered in the MRI scanner with a commercially available high-quality sound system (mr-confon, Magdeburg, Germany, stimuli 80 dB SPL, scanner noise 100 dB, passive attenuation by sound system 40 dB).

The auditory fMRI experiment (Figure 2) comprised two 22-min scanning sessions. Before each session, subjects were familiarized with the stimulation setting and listened passively to voices of familiar persons and of persons they had never seen before (each voice was presented three times saying different sentences) as well as with all sentences and the speech envelope noises. Subjects were additionally trained on the target voices and target speech envelope noises before the sessions. Subjects were not informed about the hypothesis of the experiment. Each session comprised four experimental conditions consisting in recognizing (1) the target voice of a familiar person (vfp), (2) the target voice of an nonfamiliar person (vnp), (3) the verbal content of a target sentence spoken by familiar person (cfp), and (4) the verbal content of a target sentence spoken by an nonfamiliar person (cnp). Matched control conditions involved recognizing the speech envelope noises by virtue of their temporal structure. Conditions were split into three blocks presented in random order within and across conditions. Blocks with the control conditions alternated with the experimental sentence/voice-recognition conditions.

Each block lasted 32 sec and contained 8 items (sentences or noises), of which 3 were targets. Prior to each block, subjects were verbally instructed to pay attention to the voices, the verbal content, or the temporal structure of speech envelope noises, depending on the condition, and presented with the target for the ensuing block. The acoustic material was divided into familiar (condition 1, 3) and nonfamiliar (condition 2, 4) speakers but was the same across the tasks, voice versus verbal content recognition. As a target voice would speak different sentences, and vice versa, a target sentence would be spoken by different voices, verbal content, and voice, respectively, could not serve as a cue to identify the targets. Subjects were requested to respond to each item with the right hand by pressing one button if it was a target and another button if it was not.

Face Area Localizer

Eight of the nine participants were studied in a visual face localizer study that comprised three 4.7-min fMRI sessions (Figure 2). There were four stimulus conditions (objects, faces of familiar and nonfamiliar people, and scrambled pictures) presented in blocks of 25.2 sec. The stimuli employed were frontal view pictures of 35 familiar (colleagues) and 35 nonfamiliar faces, scrambled versions of the faces, and pictures of 70 objects in canonical view. All stimuli were digital color photos with a size of 300 × 300 pixels. Single stimuli were presented every 720 msec (with the stimulus on for 525 msec and off for 195 msec). A fixation cross was introduced between the blocks for 16.8 sec.

Imaging and Data Analysis

Functional imaging was performed on a 1.5-T magnetic resonance scanner (Siemens Vision, Erlangen, Germany) with a standard head coil and gradient booster. We used echo-planar imaging to obtain image volumes with 24 contiguous oblique transverse slices every 2.7 sec (voxel size 3.44 × 3.44 × 4 mm, 1 mm gap, TE 60 msec) covering the whole brain. We acquired 494 volumes per session (988 in total per subject) in the auditory experiment and 105 volumes per session (315 in total per subject) in the visual face area localizer.

The fMRI data were preprocessed (realignment, slice-time correction, spatial normalization into stereotactic space, smoothing with a 10-mm Gaussian kernel) and analyzed using the Statistical Parametric Mapping software (SPM99; Wellcome Department of Cognitive Neu-

rology, London, UK, http://www.fil.ion.ucl.ac.uk/spm) in a MATLAB 6.1 environment (Mathworks, Sherborn, MA). A fixed-effects analysis of the auditory experiment was used to analyze the data as a 2 × 2 factorial design. We modeled subjects and conditions and applied high-pass (cutoff 512 sec) and low-pass (Gaussian 4 sec) filtering. Responses at the group level were considered significant at $p < .05$, corrected, or at $p < .001$, uncorrected, if motivated by a prior hypothesis (fusiform cortex). For the face area localizer, the same thresholds were applied.

A cross-modal response to voices of familiar persons was also checked using a random-effects analysis. Individual probabilistic maps derived from the contrasts recognition of voices of familiar > nonfamiliar persons (vfp > vnp) and recognition of voice/verbal content of familiar > nonfamiliar persons (vfp + cfp > vnp + cnp) were tested across subjects using a one-sample t test.

To investigate the functional connectivity, we performed psychophysiological interaction analyses (Gitelman, Penny, Ashburner, & Friston, 2003; Friston et al., 1997). Functional MRI signal changes over time were extracted from a volume of interest (VOI) with a radius of 5 mm centered on the response maximum for each single subject (for the right fusiform voice responsive area) or with the group maximum (for all familiarity responsive regions including the right fusiform) as representative time courses in terms of the first eigenvariate of the data. We multiplied these mean-corrected data (y) with a mean-corrected condition specific regressor (r) probing a familiarity effect (r = vfp + cfp + (vnp * −1) + (cnp * −1)) and a task effect (r = vfp + vnp + (cfp * −1) + (cnp * −1)). As the regressor was extracted from the Statistical Parametric Mapping design matrix, it was already convolved with the canonical HRF (Gitelman et al., 2003). The regressors ry for familiarity and task were used in two separate analyses per region to test for psychophysiological interactions, that is, voxels where the contribution of the sampled region changed significantly as a function of familiarity or task, respectively. In addition to ry, the design matrices also contained the regressors r and y as covariates of no interest (confounds). Responses were considered significant at $p < .05$ (corrected), 3-voxel minimum.

Acknowledgments

We thank our colleagues for their participation and Christian Büchel for comments on an earlier version of the manuscript. K. von Kriegstein, A. Kleinschmidt, and P. Sterzer are funded by the Volkswagenstiftung and A.-L.Giraud by the BMBF (Germany). The sound delivery system was acquired from a BMBF grant.

Reprint requests should be sent to Katharina von Kriegstein, Cognitive Neurology Unit, Brain Imaging Center, Johann Wolfgang Goethe University, Schleusenweg 2-16, 60528 Frankfurt am Main, Germany, or via e-mail: v.kriegstein@bic.uni-frankfurt.de.

The data reported in this experiment have been deposited in the fMRI Data Center (www.fmridc.org). The accession number is 2-2004-117M1.

REFERENCES

Bavelier, D., & Neville, H. J. (2002). Cross-modal plasticity: Where and how? *Nature Reviews Neuroscience, 3,* 443–452.

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Brain Research Cognitive Brain Research, 13,* 17–26.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403,* 309–312.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77,* 305–327.

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology, 81,* 361–380.

Cabeza, R., & Nyberg, L. (2000). Imaging cognition II: An empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience, 12,* 1–47.

Courtney, S. M., Ungerleider, L. G., Keil, K., & Haxby, J. V. (1997). Transient and sustained activity in a distributed neural system for human working memory. *Nature, 386,* 608–611.

Damasio, A. R., Tranel, D., & Damasio, H. (1990). Face agnosia and the neural substrates of memory. *Annual Review of Neuroscience, 13,* 89–109.

de Gelder, B., & Rouw, R. (2001). Beyond localisation: A dynamical dual route account of face recognition. *Acta Psychologica (Amsterdam), 107,* 183–207.

De Renzi, E., Perani, D., Carlesimo, G. A., Silveri, M. C., & Fazio, F. (1994). Prosopagnosia can be associated with damage confined to the right hemisphere—An MRI and PET study and a review of the literature. *Neuropsychologia, 32,* 893–902.

Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra- and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology, 88,* 143–156.

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

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience, 3,* 191–197.

George, N., Dolan, R. J., Fink, G. R., Baylis, G. C., Russell, C., & Driver, J. (1999). Contrast polarity and face recognition in the human fusiform gyrus. *Nature Neuroscience, 2,* 574–580.

Giraud, A. L., Price, C. J., Graham, J. M., Truy, E., & Frackowiak, R. S. (2001). Cross-modal plasticity underpins language recovery after cochlear implantation. *Neuron, 30,* 657–663.

Giraud, A. L., & Truy, E. (2002). The contribution of visual areas to speech comprehension: A PET study in cochlear implants patients and normal-hearing subjects. *Neuropsychologia, 40,* 1562–1569.

Gitelman, D. R., Penny, W. D., Ashburner, J., & Friston, K. J. (2003). Modeling regional and psychophysiologic interactions in fMRI: The importance of hemodynamic deconvolution. *Neuroimage, 19,* 200–207.

Gorno-Tempini, M. L., & Price, C. J. (2001). Identification of

famous faces and buildings: A functional neuroimaging study of semantically unique items. *Brain, 124,* 2087–2097.
Gorno-Tempini, M. L., Price, C. J., Josephs, O., Vandenberghe, R., Cappa, S. F., Kapur, N., Frackowiak, R. S., & Tempini, M. L. (1998). The neural systems sustaining face and proper-name processing. *Brain, 121,* 2103–2118.
Haxby, J. V., Grady, C. L., Horwitz, B., Ungerleider, L. G., Mishkin, M., Carson, R. E., Herscovitch, P., Schapiro, M. B., & Rapoport, S. I. (1991). Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proceedings of the National Academy of Sciences, U.S.A., 88,* 1621–1625.
Haxby, J. V., Horwitz, B., Ungerleider, L. G., Maisog, J. M., Pietrini, P., & Grady, C. L. (1994). The functional organization of human extrastriate cortex: A PET–rCBF study of selective attention to faces and locations. *Journal of Neuroscience, 14,* 6336–6353.
Ishai, A., Haxby, J. V., & Ungerleider, L. G. (2002). Visual imagery of famous faces: Effects of memory and attention revealed by fMRI. *Neuroimage, 17,* 1729–1741.
Ishai, A., Ungerleider, L. G., & Haxby, J. V. (2000). Distributed neural systems for the generation of visual images. *Neuron, 28,* 979–990.
Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17,* 4302–4311.
Leveroni, C. L., Seidenberg, M., Mayer, A. R., Mead, L. A., Binder, J. R., & Rao, S. M. (2000). Neural systems underlying the recognition of familiar and newly learned faces. *Journal of Neuroscience, 20,* 878–886.
Neuner, F., & Schweinberger, S. R. (2000). Neuropsychological impairments in the recognition of faces, voices, and personal names. *Brain and Cognition, 44,* 342–366.
O'Craven, K. M., & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience, 12,* 1013–1023.
Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh Inventory. *Neuropsychologia, 9,* 97–113.
Pallis, C. A. (1955). Impaired identification of faces and places with agnosia for colors. *Journal of Neurology, Neurosurgery and Psychiatry, 18,* 218–224.
Puce, A., Allison, T., Gore, J. C., & McCarthy, G. (1995). Face-sensitive regions in human extrastriate cortex studied by functional MRI. *Journal of Neurophysiology, 74,* 1192–1199.
Schweinberger, S. R., & Burton, A. M. (2003). Covert recognition and the neural system for face processing. *Cortex, 39,* 9–30.
Sergent, J., Ohta, S., & MacDonald, B. (1992). Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain, 115,* 15–36.
Shah, N. J., Marshall, J. C., Zafiris, O., Schwab, A., Zilles, K., Markowitsch, H. J., & Fink, G. R. (2001). The neural correlates of person familiarity. A functional magnetic resonance imaging study with clinical implications. *Brain, 124,* 804–815.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270,* 303–304.
Tarr, M. J., & Gauthier, I. (2000). FFA: A flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience, 3,* 764–769.
Tong, F., Nakayama, K., Vaughan, J. T., & Kanwisher, N. (2000). Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron, 21,* 753–759.
Van Lancker, D. R., Cummings, J. L., Kreiman, J., & Dobkin, B. H. (1988). Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex, 24,* 195–209.
von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Brain Research Cognitive Brain Research, 17,* 48–55.
Vuilleumier, P., Mohr, C., Valenza, N., Wetzel, C., & Landis, T. (2003). Hyperfamiliarity for unknown faces after left lateral temporo-occipital venous infarction: A double dissociation with prosopagnosia. *Brain, 126,* 889–907.
Wojciulik, E., Kanwisher, N., & Driver, J. (2000). Covert visual attention modulates face-specific activity in the human fusiform gyrus: fMRI study. *Journal of Neurophysiology, 79,* 1574–1578.