Simulation of talking faces in the human brain improves auditory speech recognition

Katharina von Kriegstein, Özgür Dogan, Martina Grüter, Anne-Lise Giraud, Christian A Kell, Thomas Grüter, Andreas Kleinschmidt, Stefan J Kiebel

Simulation of talking faces in the human brain improves auditory speech recognition

Katharina von Kriegstein*†‡, Özgür Dogan§, Martina Grüter¶, Anne-Lise Giraud‖, Christian A. Kell§‖, Thomas Grüter¶, Andreas Kleinschmidt**††, and Stefan J. Kiebel*

*Wellcome Trust Centre for Neuroimaging, University College London, Queen Square, London WC1N 3BG, United Kingdom; †Medical School, University of Newcastle, Framlington Place, Newcastle-upon-Tyne NE2 4HH, United Kingdom; §Department of Neurology, J. W. Goethe University, Schleusenweg, 60528 Frankfurt am Main, Germany; ¶Department of Psychological Basic Research, University of Vienna, Liebiggasse, 1010 Vienna, Austria; ‖Departement d'Etudes Cognitives, École Normale Supérieure, 75005 Paris, France; **CEA, NeuroSpin, 91401 Gif-sur-Yvette, France; and ††Institut National de la Santé et de la Recherche Médicale, U562, 91401 Gif-sur-Yvette, France

Edited by Dale Purves, Duke University Medical Center, Durham, NC, and approved March 15, 2008 (received for review November 15, 2007)

Human face-to-face communication is essentially audiovisual. Typically, people talk to us face-to-face, providing concurrent auditory and visual input. Understanding someone is easier when there is visual input, because visual cues like mouth and tongue movements provide complementary information about speech content. Here, we hypothesized that, even in the absence of visual input, the brain optimizes both auditory-only speech and speaker recognition by harvesting speaker-specific predictions and constraints from distinct visual face-processing areas. To test this hypothesis, we performed behavioral and neuroimaging experiments in two groups: subjects with a face recognition deficit (prosopagnosia) and matched controls. The results show that observing a specific person talking for 2 min improves subsequent auditory-only speech and speaker recognition for this person. In both prosopagnosics and controls, behavioral improvement in auditory-only speech recognition was based on an area typically involved in face-movement processing. Improvement in speaker recognition was only present in controls and was based on an area involved in face-identity processing. These findings challenge current unisensory models of speech processing, because they show that, in auditory-only speech, the brain exploits previously encoded audiovisual correlations to optimize communication. We suggest that this optimization is based on speaker-specific audiovisual internal models, which are used to simulate a talking face.

fMRI | multisensory | predictive coding | prosopagnosia

Human face-to-face communication works best when one can watch the speaker's face (1). This becomes obvious when someone speaks to us in a noisy environment, in which the auditory speech signal is degraded. Visual cues place constraints on what our brain expects to perceive in the auditory channel. These visual constraints improve the recognition rate for audiovisual speech, compared with auditory speech alone (2). Similarly, speaker identity recognition by voice can be improved by concurrent visual information (3). Accordingly, audiovisual models of human voice and face perception posit that there are interactions between auditory and visual processing streams (Fig. 1*A*) (4, 5).

Based on prior experimental (6–8) and theoretical work (9–12) we hypothesized that, even in the absence of visual input, the brain optimizes auditory-only speech and speaker recognition by harvesting predictions and constraints from distinct visual face areas (Fig. 1*B*).

Experimental studies (6, 8) demonstrated that the identification of a speaker by voice is improved after a brief audiovisual experience with that speaker (in contrast to a matched control condition). The improvement effect was paralleled by an interaction of voice and face-identity sensitive areas (8). This finding suggested that the associative representation of a particular face facilitates the recognition of that person by voice. However, it is unclear whether this effect also extends to other audiovisual dependencies in human communication. Such a finding, for example in the case of speech recognition, would indicate that the brain fills-in missing information routinely to make auditory communication more robust.

To test this hypothesis, we asked the following question: What does the brain do when we listen to someone whom we have previously seen talking? Classical speech processing models (the "auditory-only" model) predict that the brain uses auditory-only processing capabilities to recognize speech and speaker (13, 14). Under the "audiovisual" model, we posit that the brain uses previously learned audiovisual speaker-specific information to improve recognition of both speech and speaker (Fig. 1*B*). Even without visual input, face-processing areas could use encoded knowledge about the visual orofacial kinetics of talking and simulate a specific speaker to make predictions about the trajectory of what is heard. This visual online simulation would place useful constraints on auditory perception to improve speech recognition by resolving auditory ambiguities. This constructivist view of perception has proved useful in understanding human vision (15, 16) and may be even more powerful in the context of integration of prior multimodal information. To identify such a mechanism in speech perception would not only have immediate implications for the ecological validity of auditory-only models of speech perception but would also point to a general principle of how the brain copes with noisy and missing information in human communication.

Speech and speaker recognition largely rest on two different sets of audiovisual correlations. Speech recognition is based predominantly on fast time-varying acoustic cues produced by the varying vocal tract shape, i.e., orofacial movements (17, 18). Conversely, speaker recognition uses predominantly time-invariant properties of the speech signal, such as the acoustic properties of the vocal tract length (19). If the brain uses stored visual information for processing auditory-only speech, the relative improvement in speech and speaker recognition could, therefore, be behaviorally and neuroanatomically dissociable. To investigate this potential dissociation, we recruited prosopagnosics who have impaired perception of face identity but seem to have intact perception of orofacial movements (20).

Neurophysiological face processing studies indicate that distinct brain areas are specialized for processing time-varying information [facial movements, superior temporal sulcus (STS)

Author contributions: K.v.K. designed research; K.v.K., Ö.D., M.G., A.-L.G., C.A.K., T.G., and A.K. performed research; S.J.K. contributed new reagents/analytic tools; K.v.K. and Ö.D. analyzed data; and K.v.K. and S.J.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

‡To whom correspondence should be addressed. E-mail: kkriegs@fil.ion.ucl.ac.uk.

This article contains supporting information online at www.pnas.org/cgi/content/full/0710826105/DCSupplemental.

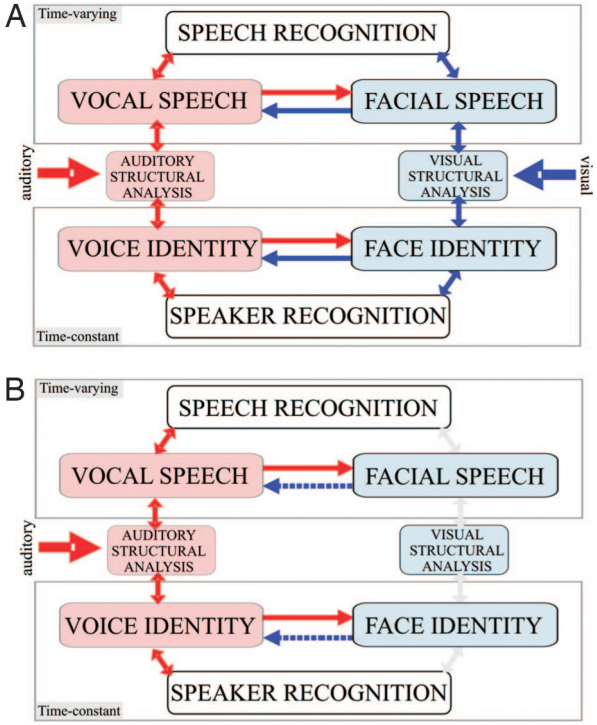

Fig. 1. Model for processing of human communication signals. (*A*) Audiovisual input enters auditory and visual preprocessing areas. These feed into two distinct networks, which process speech and speaker information. Modified from ref. 4. (*B*) Auditory-only input enters auditory preprocessing areas. For speech recognition, facial and vocal speech areas interact while engaging concurrently with higher levels of speech processing. Similarly, for speaker recognition, face and voice identity areas interact with higher levels of speaker identity processing. Note that the interactions between the boxes do not imply direct anatomical connections and that the boxes may represent more than one area, in particular for higher levels of speech and speaker recognition.

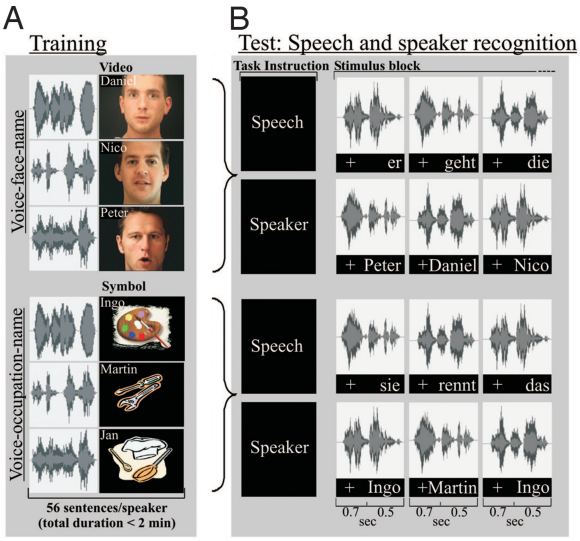

Fig. 2. Experimental design. (*A*) Training phase. All participants learned to identify the same six speakers. Three were learned by voice, video of moving face, and name. Three others were learned by voice, name and a visual symbol of an occupation. (*B*) Test phase (in the MRI-scanner). Participants performed either the speech or speaker recognition task, cued visually before each stimulus block started. The vehicle task and the face area localizer are not displayed in this figure.

(21, 22), and time-constant information (face identity, fusiform face area (FFA) (23–25)] (26, 27). If speech and speaker recognition are neuroanatomically dissociable, and the improvement by audiovisual learning uses learned dependencies between audition and vision, the STS should underpin the improvement in speech recognition in both controls and prosopagnosics. A similar improvement in speaker recognition should be based on the FFA in controls but not prosopagnosics. Such a neuroanatomical dissociation would imply that visual face processing areas are instrumental for improved auditory-only recognition. We used functional magnetic resonance imaging (fMRI) to show the response properties of these two areas.

The study consisted of (*i*) "training phase," (*ii*) "test phase," and (*iii*) "face area localizer." In the training phase (Fig. 2*A*), both groups (17 controls and 17 prosopagnosics) learned to identify six male speakers by voice and name. For three speakers, the voice–name learning was supplemented by a video presentation of the moving face ("voice–face" learning), and for the other three speakers by a symbol of their occupation (no voice–face learning, which we term "voice–occupation" learning).

The test phase (Fig. 2*B*) was performed in the MRI-scanner. Auditory-only sentences from the previously learned speakers were presented in 29-s blocks. These sentences had not been used during the training phase. Before each block, participants received the visual instruction to either perform a speaker or speech recognition task. There were four experimental conditions in total: (*i*) speech task: speaker learned by face; (*ii*) speech task: speaker learned by occupation, (*iii*) speaker task: speaker learned by face; and (*iv*) speaker task: speaker learned by occupation. For the speech tasks, subjects indicated by button press whether a visually presented word occurred during the concurrent auditory sentence. In the speaker tasks, subjects indicated whether the visually presented speaker name corresponded to the speaker of the auditory sentence. A nonspeech control condition with vehicle sounds was included in the test phase. In this condition, subjects indicated whether the visually displayed vehicle name (train, motorcycle, or racing car) corresponded to the concurrently presented vehicle sound.

After the test phase, fMRI data for the face area localizer were acquired. This included passive viewing of faces and objects and was used to localize the face-sensitive FFA and STS (see *Methods*).

Results

Behavior. An overview of the behavioral results is displayed in Table 1. We performed a three-way repeated-measure ANOVA with the within-subject factors "task" (speech, speaker), "learning" (voice–face, voice–occupation) and the between-subject factor "group" (prosopagnosics, controls). There was a main effect of task [$F(1,32) = 74.7, P < 0.001$]; a trend to significance for the main effect of type of learning [$F(1,32) = 4.0, P = 0.053$]; a type of learning and group interaction [$F(1,32) = 4.8, P < 0.04$]; and a three-way interaction between task, type of learning, and group [$F(1,32) = 5.5, P < 0.03$].

In both groups, prior voice–face learning improved speech recognition, compared with voice–occupation learning. In the following, we will call such improvement "face-benefit." For both controls and prosopagnosics there was a significant face-benefit for speech recognition (paired *t* test: speech task/voice–face vs. speech task/voice–occupation learning: t = 2.3, df = 32, $P < 0.03$) [Fig. 3, Table 1 and supporting information (SI) Fig. S1]. Although face-benefits of 1.22% (controls) and 1.53% (prosopagnosics) seem small, these values are expected given the recognition rates were >90% (28). There was no significant difference in the face-benefit between the two groups: An ANOVA for the speech task with the factors learning (voice–face, voice–occupation) and group (prosopagnosic, controls)

Table 1. Behavioral scores for all four experimental conditions and the vehicle control condition

		Controls		Prosopagnosics	
Task	Experimental condition	%	SE	%	SE
Speech	Voice–face	93.50	1.13	95.80	0.53
	Voice–occupation	92.28	1.16	94.27	0.76
	Face benefit	1.22	1.04	1.53	0.58
Speaker	Voice–face	82.41	3.11	78.34	2.52
	Voice–occupation	77.14	3.34	80.15	1.93
	Face benefit	5.27	2.10	−1.81	1.99
Vehicle	—	91.65	1.21	92.76	0.92

Recognition rates (%) are summarized as average over group with standard error (SE). The face-benefit is defined as the task-specific recognition rate after voice–occupation learning subtracted from the recognition rate after voice–face learning.

revealed no interaction [$F(1,32) = 0.06$, $P = 0.8$]. There was a main effect type of learning [$F(1,32) = 5.0$, $P < 0.03$], which is consistent with a face-benefit in both groups for speech recognition.

In the control group, there was a significant face-benefit of 5.27% for speaker recognition (paired t test: speaker task/voice–face vs. speaker task/voice–occupation learning: $t = 2.5$, df = 16, $P < 0.02$). Critically, there was no face-benefit in the prosopagnosics for speaker recognition ($t = -0.9$, df = 16, $P = 0.4$) (Fig. 3, Table 1, and Fig. S1). An ANOVA for the speaker task revealed a significant difference of face-benefit between the controls and prosopagnosics {learning × group interaction in speaker task [$F(1,32) = 6.1$, $P < 0.02$]}.

We also probed whether the face-benefits in speech and speaker recognition were correlated. Neither controls (Pearson: $r = 0.03$, $P = 0.9$) nor prosopagnosics (Pearson: $r = -0.1$, $P = 0.7$) showed a correlation between the two face-benefit scores. This means that a subject with, e.g., a high face-benefit in speaker recognition does not necessarily have a high face-benefit in speech recognition.

Neuroimaging. We performed two separate analyses of blood oxygenation level-dependent (BOLD) responses acquired during the test phase. First, we examined the effect of learning voice–face vs. voice–occupation associations on the responses in face-sensitive STS and FFA (categorical analysis, Fig. 4 *A* and *B* and Fig. S2). In a second analysis, we examined the correlations between behavior and regional activation over subjects in the two face-sensitive areas (correlation analysis, Fig. 4 *C–F*). In both these analyses, we used the face area localizer to localize the STS and FFA (see *Methods*).

Categorical Analysis. In both groups, activity in face-sensitive STS is increased after voice–face learning, for speech recognition (Fig. 4*A*). There was a significant interaction between learning (voice–face vs. voice–occupation) and task (speech vs. speaker)

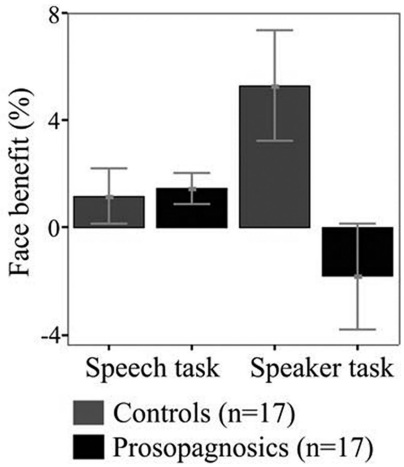

Fig. 3. Behavioral results. Face-benefit in speech and speaker recognition tasks for controls (gray) and prosopagnosics (black). The face-benefit is the percentage difference between correct recognition after voice–face learning minus correct recognition after voice–occupation learning. The error bars represent standard errors.

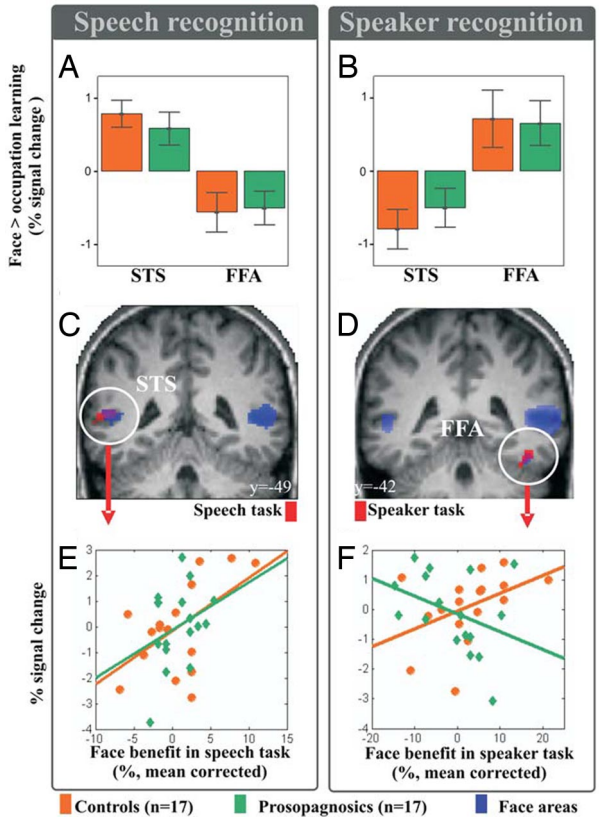

Fig. 4. fMRI results. (*A* and *B*) Difference contrasts between voice–face and voice–occupation learning in speech (*A*) and speaker recognition (*B*). (*C*) Statistical parametric map of positive correlations of BOLD activity with the face-benefit for speech recognition. (*D*) Statistical parametric map of the difference between controls and prosopagnosics in positive correlation of BOLD activity with the face-benefit for speaker recognition. (*E*) Plot of correlation between face-benefit in speech task and STS activity. (*F*) Plot of correlation between face-benefit in speaker task and FFA activity; for controls, the correlation was significant, but not for prosopagnosics. This figure displays the results for the ROI in the left STS. See Table S5 and Fig. S4 for results for the ROI in the right STS.

[ANOVA: $F(1,32) = 24$, $P < 0.0001$)]. In each group, 15 of 17 subjects showed this effect (Table S1).

The activation in the FFA was increased after voice–face learning, but only in the speaker task (Fig. 4*B*). There was a significant interaction between learning (voice–face vs. voice–occupation) and task (speaker vs. speech) [ANOVA: $F(1,32) = 17$, $P < 0.0001$]. Fifteen of 17 subjects in both groups showed this FFA effect (Table S2).

Correlation Analysis. For both groups, a significant positive correlation between activation and face-benefit in speech recognition was found in the left face-sensitive STS ($P = 0.03$, corrected; $n = 34$; statistical maximum at $x = -56$, $y = -44$, $z = 10$; Pearson: $r = 0.5$, $P = 0.006$, two-tailed, $n = 34$) but not in the FFA (both groups, $P = 0.6$). There was no difference between groups in the FFA ($P = 0.5$) (Fig. 4 *C* and *E*). Note that this STS region is a visual face-sensitive area and not active during speech in general; there is no activity in this region when contrasting all conditions containing speech against the control condition with vehicle sounds. Furthermore activity is not higher in the speech task in contrast to the speaker task after voice–occupation learning (Fig. S3).

For controls, we found a significant positive correlation between FFA activity and the face-benefit in speaker recognition (in controls, $P < 0.03$, corrected; statistical maximum at $x = 40$, $y = -42$, $z = -26$; Pearson: $r = 0.6$, $P = 0.012$, two-tailed, $n = 17$). This correlation was significantly greater than in prosopagnosics (controls > prosopagnosics: $P < 0.01$, corrected). There was no significant positive or negative correlation in prosopagnosics ($P = 0.9$; Pearson: $r = -0.4$, $P = 0.15$, two-tailed, $n = 17$) (Fig. 4 *D* and *F*). As expected, no significant correlation between STS activity and face-benefit in speaker recognition was observed (controls > prosopagnosics: $P = 0.9$, corrected; controls: $P = 0.5$, corrected).

Discussion

The results are in line with our prediction that the brain exploits previously acquired speaker-specific audiovisual information to improve both auditory-only speech and speaker recognition.

Importantly, we can discount an alternative explanation for the face-benefits in speech and speaker recognition: Using the auditory-only model, one could argue that subjects, during the training phase, pay more attention to voices presented during the voice–face learning because of the matching visual video. In contrast, voice–occupation learning is based on static stimuli. This difference in stimuli could result in an advantage for auditory learning during voice–face association and, potentially, explain a face-benefit. However, with this argument, one would necessarily expect a correlation between the face-benefits for speaker and speech recognition. There was no such correlation. In addition, the prosopagnosics are unimpaired on auditory learning; we showed that they do as well as normal subjects after voice–occupation learning. Therefore, the auditory-only model predicts that the prosopagnosics show face-benefits in both tasks, which was not observed. Rather, there was a difference in the face-benefit pattern between the controls and prosopagnosics, which confirms a neuropsychological dissociation in terms of the face-benefits of speech and speaker recognition. We can, therefore, rule out a general attention effect under the auditory model as an explanation for our results.

We conclude that subjects must have learned key audiovisual speaker-specific attributes during the training phase. This learning was fast; <2 min of observing each speaker improved subsequent speech recognition, compared with learning based on arbitrary audiovisual cues. A translation of this principle into every day life is improved telephone communication when the speakers have previously engaged in a brief audiovisual exchange, for example during a meeting. The same argument applies to speaker recognition. Control subjects identified a speaker by voice better, if they had seen the speaker talking before. This latter finding confirms two previous studies that show better speaker recognition after voice–face learning (6, 8).

The audiovisual model (Fig. 1) and visual face processing models (26) assumes two separable neural systems for the processing of face motion (STS) and face identity (FFA). A neuroimaging study showed that during speaker recognition FFA activity is increased after voice–face learning (8). Our present findings extend this result in three ways: (*i*) We show that face-movement sensitive STS activity is increased after voice–face learning, but only during speech recognition; (*ii*) activity of the left face-sensitive STS positively correlates with the face-benefit in speech recognition and FFA activity positively correlates with the face-benefit for speaker recognition; and (*iii*) FFA activity correlates positively with the face-benefit in controls but not in prosopagnosics. These results confirm our hypothesis about a neuroanatomical dissociation, in terms of selective task and stimulus-bound response profiles in STS and FFA. We suggest that individual dynamic facial "signatures" (29) are stored in the STS and are involved in predicting the incoming speech content. Note that these dynamic facial signatures might also carry identity information and could therefore be potentially used to improve identity recognition in humans and in primates (30–32). However, our results suggest that neither the controls nor the prosopagnosic subjects employ this information in our experiment to improve their speaker recognition abilities.

Speech recognition during telephone conversations can be improved by video-simulations of an artificial "talking face," which helps especially hearing impaired listeners to understand what is said (33). This creation of an artificial talking face uses a phoneme recognizer and a face synthesizer to recreate the facial movements based on the auditory input. We suggest that our results reflect that the human brain routinely uses a similar mechanism: Auditory-only speech processing is improved by simulation of a talking face. How can such a model be explained in theoretical terms? In visual and sensory-motor processing, "internal forward" models have been used to explain how the brain encodes complex sensory data by relatively few parameters (34, 35). Here, we assume the existence of audiovisual forward models, which encode the physical causal relationship between a person talking and its consequences for the visual and auditory input. Critically, these models also encode the causal dependencies between the visual and auditory trajectories. Perception is based on the "inversion" of models; i.e., the brain identifies causes (Mr. Smith says, "Hello") that explain the observed audiovisual input best. Given that robust communication is of utmost importance for us, we posit that the human brain can quickly and efficiently learn "a new person" by adjusting key parameters in existing internal audiovisual forward models that are already geared toward perception of talking faces. Once parameters for an individual person are learned, auditory speech processing is improved because the brain learned parameters of an audiovisual forward model with strong dependencies between internal auditory and visual trajectories. This enables the system to simulate visual trajectories (via the auditory trajectories) when there is no visual input. The talking face simulation is the better the stronger and more veridical the learned coupling between auditory and visual input is. The visual simulation is fed back to auditory areas thereby improving auditory recognition by providing additional constraints. This mechanism can be used iteratively until the inversion of the audiovisual forward model converges on a percept. The scheme to employ forward models to encode and exploit dependencies in the environment by simulation is in accordance with general theories of brain function, which posit that neural mechanisms are tuned for efficient prediction of relevant stimuli (9, 10, 12, 16, 36, 37).

We suggest that the simulation of facial features is reflected in our results by the recruitment of visual face areas in response to auditory stimulation. Our findings imply that there are distinct audiovisual models for time-varying and time-constant audiovisual dependencies. We posit that the simulation of a face in response to auditory speech is a general mechanism in human communication. We predict that the same principle also applies to other information that is correlated in the auditory and visual domains, such as recognition of emotion from voice and face (38, 39). Furthermore, this scheme might be a general principle of how unisensory tasks are performed when one or more of the usual input modalities are missing (8, 40).

In summary, we have shown that the brain uses previously encoded visual face information to improve subsequent auditory-only speech and speaker recognition. The improvement in speech and speaker recognition is behaviourally and neuroanatomically dissociable. Speech recognition is based on selective recruitment of the left face-sensitive STS, which is known to be involved in orofacial movement processing (21, 22). Speaker recognition is based on selective recruitment of the FFA, which is involved in face-identity processing (23–25). These findings challenge auditory-only models for speech processing and lead us to conclude that human communication involves at least two distinct audiovisual networks for auditory speech and speaker recognition. The existence of an optimized and robust scheme for human speech processing is a key requirement for efficient communication and successful social interactions.

Methods

Participants. In total, 17 healthy volunteers (10 females, 14 right handed, 22–52 years of age, mean age 37.4 years, median 38 years) and 17 prosopagnosics (11 females, 17 right handed, 24–57 years of age, mean age 37.2 years, median 34 years) were included into the study (*SI Methods*, *Participants*).

Prosopagnosia Diagnosis. The diagnosis of hereditary prosopagnosia was based on a standardized semistructured interview (Tables S3 and S4) (41, 42), which has been validated with objective face recognition tests in previous studies (41, 43).

Stimuli. For a detailed description of the stimuli, see *SI Methods*, *Stimuli*.

Experimental Design. ***Training phase.*** All participants were trained outside the MRI-scanner. In each trial the name of the speaker was first presented (for 1 s) followed by presentation of a sentence spoken by that speaker (≈1.3 s). For three of the speakers, the sentences were presented together with the video of the speaking face (voice–face learning). Three other speakers' voices were presented together with static symbols for three different occupations (painter, craftsman, and cook) (voice–occupation learning). The two sets of speakers were counterbalanced over participants: In each group, nine participants learned the first set of speakers with the faces and the second set with the symbols, whereas the other eight participants learned the reverse order. Total exposure to audiovisual information about a speaker was <2 min (*SI Methods*, *Experimental Design*).

Test phase. The test phase consisted of three 15-min MRI-scanning sessions and included four speech and one nonspeech condition (see Introduction). Before the first session, participants were briefly familiarized, inside the scanner, with the setting by showing them a single trial of each task. Stimuli (auditory sentences or vehicle sounds) were presented in a block design. Blocks were presented fully randomized. There were 12 blocks per condition in total. Each block lasted 29 s and contained eight trials. One trial lasted ≈3.6 s and consisted of two consecutive sentences spoken by the same person or two vehicle sounds. In the last second of each trial, a written word (speech task), person name (speaker task), or vehicle name (vehicle task) was presented. Subjects indicated via button press whether the shown word was present in the spoken sentence (speech task) and whether the shown person name matched the speaker's voice (speaker task) or not. Similarly in the vehicle task, subjects indicated whether the vehicle name matched the vehicle sound or not. Between blocks, subjects looked at a fixation cross lasting 12 s.

Face area localizer. The visual localizer study consisted of two 6-min MRI-scanning sessions and included four conditions of passive viewing of face or object pictures: (*i*) faces from different persons with different facial gestures (speaking), (*ii*) different facial gestures of the same person's face, (*iii*) different objects in different views, and (*iv*) same object in different views. Conditions were presented as blocks of 25-s duration. Within the blocks, single stimuli were presented for 500 ms without pause between stimuli. This fast stimulus presentation induced a movement illusion in the condition where the same person's face was presented (moving face), but not in those with faces from different persons (static faces). A fixation cross was introduced between the blocks for 18 s.

Data Acquisition and Analysis. MRI was performed on a 3-T Siemens Vision scanner (*SI Methods*, *Data acquisition*), and the data were analyzed with SPM5 (www.fil.ion.ucl.ac.uk/spm), using standard procedures (*SI Methods*, *Analysis of MRI data*).

Behavioural data were analyzed by using SPSS 12.02 (SPSS). All *P* values reported are two-tailed.

Localization of face-sensitive areas. We defined the regions of interest (ROI) by using the face area localizer. The STS-ROI was defined by the contrast moving face vs. static faces. In the group analysis, this contrast was used to inclusively mask the contrast face vs. object (maximum for both groups in left STS: $x = -52$, $y = -56$, $z = 6$, cluster size 19 voxels). The localizer contrast also included a region in the right STS ($x = 54$, $y = -40$, $z = 6$, cluster size 737 voxels). We report analyses within this region in *SI Methods*, Table S5, and Fig. S4. The FFA-ROI was defined by the contrast faces vs. objects (maximum for both groups was in the right FFA: $x = 44$, $y = -44$, $z = -24$, cluster size 20 voxels) (*SI Methods*, *Face area localizer*). There was no homologous significant activity in the left hemisphere. The statistical maxima for individual subjects are displayed in (Tables S1 and S2).

Categorical analysis for test phase. In the categorical analysis, contrasts of interest were the interactions (*i*) (speech task/voice–face learning − speech task/voice–occupation learning) − (speaker task/voice–face learning − speaker task/voice–occupation learning) and (*ii*) (speaker task/voice–face learning − speaker task/voice–occupation learning) − (speech task/voice–face learning − speech task/voice–occupation learning). These contrasts were computed at the single subject level. For each subject's FFA and STS (as determined by the face area localizer), parameter estimates were extracted from the voxel, at which we found the maximum statistic (*SI Methods*, *Categorical analysis*, and Tables S1 and S2). These values were then entered into a repeated measures ANOVA and plotted (Fig. 4 *A* and *B*) by using SPSS 12.02.

Correlation analysis for test phase. In the correlation analysis, the fMRI signal in FFA and STS after voice–face learning was correlated with the behavioral face-benefit, i.e., recognition rate (%) after voice–face learning minus recognition rate (%) after voice–occupation learning, as determined separately for speech and speaker task. This group analysis was performed by using the MarsBaR ROI toolbox (http://marsbar.sourceforge.net) (*SI Methods*, *Correlation analysis*). To estimate Pearson's *r* values, parameter estimates were extracted at the group maximum and entered into SPSS 12.02.

ACKNOWLEDGMENTS. We thank Chris Frith, Karl Friston, Peter Dayan, and Tim Griffiths for comments on the manuscript and Stefanie Dahlhaus for providing the voice–face videos. This study was supported by the VolkswagenStiftung and Wellcome Trust.

1. Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. *J Acoust Soc Am* 26:212–215.
2. van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. *Proc Natl Acad Sci USA* 102:1181–1186.
3. Schweinberger SR, Robertson D, Kaufmann JM (2007) Hearing facial identities. *Q J Exp Psychol (Colchester)* 60:1446–1456.
4. Belin P, Fecteau S, Bedard C (2004) Thinking the voice: Neural correlates of voice perception. *Trends Cognit Sci* 8:129–135.
5. Braida LD (1991) Crossmodal integration in the identification of consonant segments. *Q J Exp Psychol A* 43:647–677.
6. Sheffert SM, Olson E (2004) Audiovisual speech facilitates voice learning. *Percept Psychophys* 66:352–362.
7. von Kriegstein K, Kleinschmidt A, Giraud AL (2006) Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. *Cereb Cortex* 16:1314–1322.
8. von Kriegstein K, Giraud AL (2006) Implicit multisensory associations influence voice recognition. *PLoS Biol* 4:e326.
9. Deneve S, Duhamel JR, Pouget A (2007) Optimal sensorimotor integration in recurrent cortical networks: A neural implementation of Kalman filters. *J Neurosci* 27:5744–5756.
10. Friston K (2005) A theory of cortical responses. *Philos Trans R Soc London Ser B* 360:815–836.
11. Halle M (2002) *From Memory to Speech and Back: Papers on Phonetics and Phonology, 1954–2002* (de Gruyter, Berlin).
12. Knill D, Kersten D, Yuille A (1998) in *Perception as Bayesian Inference*, eds Knill D, Richards W (Cambridge Univ Press, Cambridge, UK), pp 1–21.

13. Ellis HD, Jones DM, Mosdell N (1997) Intra- and inter-modal repetition priming of familiar faces and voices. *Br J Psychol* 88:143–156.
14. Hickok G, Poeppel D (2007) The cortical organization of speech processing. *Nat Rev Neurosci* 8:393–402.
15. Dayan P (2006) Images, frames, and connectionist hierarchies. *Neural Comput* 18:2293–2319.
16. 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.
17. Fant G (1960) *Acoustic Theory of Speech Production* (Mouton, Paris).
18. Yehia H, Rubin P, Vatikiotis-Bateson E (1998) Quantitative association of vocal-tract and facial behavior. *Speech Commun* 26:23–43.
19. Lavner Y, Gath I, Rosenhouse J (2000) The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Commun* 30:9–26.
20. Humphreys K, Avidan G, Behrmann M (2007) A detailed investigation of facial expression processing in congenital prosopagnosia as compared to acquired prosopagnosia. *Exp Brain Res* 176:356–373.
21. Calvert GA, *et al.* (1997) Activation of auditory cortex during silent lipreading. *Science* 276:593–596.
22. Puce A, Allison T, Bentin S, Gore JC, McCarthy G (1998) Temporal cortex activation in humans viewing eye and mouth movements. *J Neurosci* 18:2188–2199.
23. Eger E, Schyns PG, Kleinschmidt A (2004) Scale invariant adaptation in fusiform face-responsive regions. *NeuroImage* 22:232–242.
24. Loffler G, Yourganov G, Wilkinson F, Wilson HR (2005) fMRI evidence for the neural representation of faces. *Nat Neurosci* 8:1386–1390.
25. Rotshtein P, Henson RN, Treves A, Driver J, Dolan RJ (2005) Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nat Neurosci* 8:107–113.
26. Haxby JV, Hoffman EA, Gobbini MI (2000) The distributed human neural system for face perception. *Trends Cognit Sci* 4:223–233.
27. Kanwisher N, Yovel G (2006) The fusiform face area: A cortical region specialized for the perception of faces. *Philos Trans R Soc London Ser B* 361:2109–2128.
28. Dupont S, Luettin J (2000) Audio-visual speech modeling for continous speech recognition. *IEEE Trans Multimedia* 2:141–151.
29. O'Toole AJ, Roark DA, Abdi H (2002) Recognizing moving faces: A psychological and neural synthesis. *Trends Cognit Sci* 6:261–266.
30. Ghazanfar AA, Maier JX, Hoffman KL, Logothetis NK (2005) Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *J Neurosci* 25:5004–5012.
31. Lander K, Davies R (2007) Exploring the role of characteristic motion when learning new faces. *Q J Exp Psychol (Colchester)* 60:519–526.
32. Kamachi M, Hill H, Lander K, Vatikiotis-Bateson E (2003) Putting the face to the voice: Matching identity across modality. *Curr Biol* 13:1709–1714.
33. Siciliano C, Williams G, Beskow J, Faulkner A (2002) Evaluation of a multilingual synthetic talking face as a communication aid for the hearing-impaired. *Speech Hear Lang Work Prog* 14:51–61.
34. Ballard DH, Hinton GE, Sejnowski TJ (1983) Parallel visual computation. *Nature* 306:21–26.
35. Kawato M, Hayakawa H, Inui T (1983) A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network-Comput Neural Syst* 4:415–422.
36. Bar M (2007) The proactive brain: Using analogies and associations to generate predictions. *Trends Cognit Sci* 11:280–289.
37. Wolpert DM, Ghahramani Z, Jordan MI (1995) An internal model for sensorimotor integration. *Science* 269:1880–1882.
38. de Gelder B, Pourtois G, Weiskrantz L (2002) Fear recognition in the voice is modulated by unconsciously recognized facial expressions but not by unconsciously recognized affective pictures. *Proc Natl Acad Sci USA* 99:4121–4126.
39. de Gelder B, Morris JS, Dolan RJ (2005) Unconscious fear influences emotional awareness of faces and voices. *Proc Natl Acad Sci USA* 102:18682–18687.
40. Amedi A, *et al.* (2007) Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. *Nat Neurosci* 10:687–689.
41. Gruter M, *et al.* (2007) Hereditary prosopagnosia: The first case series. *Cortex* 43:734–749.
42. Kennerknecht I, *et al.* (2006) First report of prevalence of non-syndromic hereditary prosopagnosia (HPA). *Am J Med Genet A* 140:1617–1622.
43. Carbon CC, Grueter T, Weber JE, Lueschow A (2007) Faces as objects of non-expertise: Processing of thatcherised faces in congenital prosopagnosia. *Perception* 36:1635–1645.