Cross-modal Interactions during Perception of Audiovisual Speech and Nonspeech Signals: An fMRI Study

Ingo Hertrich, Susanne Dietrich, and Hermann Ackermann

Abstract

■ During speech communication, visual information may interact with the auditory system at various processing stages. Most noteworthy, recent magnetoencephalography (MEG) data provided first evidence for early and preattentive phonetic/phonological encoding of the visual data stream—prior to its fusion with auditory phonological features [Hertrich, I., Mathiak, K., Lutzenberger, W., & Ackermann, H. Time course of early audiovisual interactions during speech and non-speech central-auditory processing: An MEG study. *Journal of Cognitive Neuroscience, 21,* 259–274, 2009]. Using functional magnetic resonance imaging, the present follow-up study aims to further elucidate the topographic distribution of visual–phonological operations and audiovisual (AV) interactions during speech perception. Ambiguous acoustic syllables—disambiguated to /pa/ or /ta/ by the visual channel (speaking face)—served as test materials, concomitant with various control conditions (nonspeech AV signals, visual-only and acoustic-only speech, and nonspeech stimuli). (i) Visual speech yielded an AV-subadditive activation of primary auditory cortex and the anterior superior temporal gyrus (STG), whereas the posterior STG responded both to speech and nonspeech motion. (ii) The inferior frontal and the fusiform gyrus of the right hemisphere showed a strong phonetic/phonological impact (differential effects of visual /pa/ vs. /ta/) upon hemodynamic activation during presentation of speaking faces. Taken together with the previous MEG data, these results point at a dual-pathway model of visual speech information processing: On the one hand, access to the auditory system via the anterior supratemporal "what" path may give rise to direct activation of "auditory objects." On the other hand, visual speech information seems to be represented in a right-hemisphere visual working memory, providing a potential basis for later interactions with auditory information such as the McGurk effect. ■

INTRODUCTION

Speakers, both during formal and colloquial conversation, usually talk to each other face to face. As a consequence, visual information arising from articulatory movements, especially lip gestures, complements the acoustic speech signal and may have a significant impact upon spoken language comprehension. For example, "lip reading" can considerably enhance the intelligibility of verbal utterances (Sumby & Pollack, 1954). Psychoacoustic experiments demonstrated, furthermore, a particularly pronounced impact of the visual channel on speech perception in case of ambiguous acoustic signals (e.g., addition of noise to the test materials; Sekiyama, Kanno, Miura, & Sugita, 2003; Sekiyama & Tohkura, 1991). Computational models of audiovisual (AV) speech perception based upon psychoacoustic data primarily addressed the controversial issue of an early versus late fusion of these two afferent data streams (e.g., Schwartz, Robert-Ribes, & Escudier, 1998). However, more recent electrophysiological studies—focusing upon the time course of the underlying cerebral processes—indicate that AV speech signals give rise to cross-modal interactions at various latencies—pointing at the involvement of several successive processing stages (Hertrich, Mathiak, Lutzenberger, Menning, & Ackermann, 2007; Van Wassenhove, Grant, & Poeppel, 2005; Möttönen, Schurmann, & Sams, 2004; Möttönen, Krause, Tiippana, & Sams, 2002), even as early as the auditory P50 potential (Lebib, Papo, De Bode, & Baudonniere, 2003) and the magnetic M50 field (Hertrich, Mathiak, Lutzenberger, & Ackermann, 2009). Interestingly, a further study—based upon an apparent motion design switching the visual mouth display exactly at the time of the onset of acoustic stimulation—did not find a significant impact of visual information upon auditory M100 fields. As a consequence, the natural delay of ca. 150 msec between visual and acoustic cues (see Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009) seems to represent an important prerequisite for early AV interactions in that visual cues may act as short-term primes for central-auditory processing. If the onset of the acoustic signal is shifted against the onset of the video sequence during application of AV speech stimuli, the tolerance for fusion processes is larger for delayed acoustic than for delayed visual signals. As a potential neurophysiological explanation of these effects, Schroeder, Lakatos, Kajikawa, Partan, and Puce (2008) assume visual input to act upon the phase, for instance, in terms of resetting operations, of slow oscillatory activity of the central-auditory system.

Besides electrophysiological studies, functional magnetic resonance imaging (fMRI) was also able to document

University of Tübingen, Germany

early as well as later stages of AV interactions during speech perception. In particular, these studies showed the inferior frontal gyrus (IFG) of the left hemisphere to support perceptually relevant levels of AV speech processing (Hasson, Skipper, Nusbaum, & Small, 2007; Skipper, van Wassenhove, Nusbaum, & Small, 2007). With respect to the central-auditory system, these studies indicate the planum polare, that is, a region anterior to primary auditory cortex (PAC), to participate in the representation of phonological data structures, whereas the left planum temporale, posterior to PAC, rather seems to contribute to AV attention processes (Pekkola et al., 2006).

Recent magnetoencephalography (MEG) investigations provided first electrophysiological evidence for phonetic/linguistic encoding of the "visual stream" of spoken language prior to its integration with auditory input, supporting the assumption of "separate identification" (see Schwartz et al., 1998) of auditory and visual phonetic features during speech perception (Hertrich, Mathiak, et al., 2009; Hertrich et al., 2007; see below). More specifically, two preceding whole-head MEG studies of our group, focusing upon early preattentive AV phenomena, were able to delineate three subsequent AV interactions during syllable perception:

(1) a "preparatory baseline shift," in terms of an early, unspecific response of the central-auditory system both to visual speech or nonspeech events,
(2) a subadditive impact of visual motion onto the auditory M100 field, and
(3) interactions at the level of auditory sensory memory, related to a late component of the visually induced phonological mismatch response (/pa/–/ta/ contrast) (Hertrich, Mathiak, et al., 2009; Hertrich et al., 2007).

Most noteworthy, the applied visual cues (display of a speaker's face) were found to be represented in a phonetic/categorical manner already at a latency of 250 to 300 msec after the onset of speech movements. This activity can be assigned to the time domain of the auditory M100 field as the acoustic stimuli lagged behind the visual cues by ca. 150 msec. Phonetic encoding of the "visual stream" of spoken language, thus, precedes the late interaction level, that is, the fusion of auditory and visual information into a common sound percept.

Dipole analyses were able to assign the magnetic sources of the visually evoked field bound to early categorical speech representation to posterior insular cortex of both hemispheres, that is, to a location outside the central-auditory system. However, this pair of dipoles was derived under the assumption of a single point-like source at either side of the brain. Although this "insular" activity accounts for a considerable amount of variance, its location, as an alternative, might reflect the "center of gravity" of a group of distributed sources. To overcome these principal spatial limitations of electrophysiological surface data, the present study tries, as its first aim, to further specify, using fMRI, the site of origin of the neural responses associated with the representation of visual phonological features. Therefore, this functional imaging experiment used—more or less—the same test materials (bimodal and unimodal speech and nonspeech stimulus configurations) and a similar design as the preceding MEG investigation (Hertrich, Mathiak, et al., 2009). Again, visual stimuli varied across three levels of movement range (see Table 1): no movement (=baseline), small concentric movements of a circle (or /ta/ in case of speech), and larger circle movements (or /pa/, respectively). In order to address early preattentive processing of AV events with respect to the /pa/–/ta/ distinction, subjects were asked to monitor pitch changes (final rise or decline) imposed upon the acoustic stimulus components (bimodal selective attention task; see Johnson & Zatorre, 2006). As a further major finding, the previous whole-head MEG study was able to document a subadditive influence of visual speech on the strength of the auditory M100 field. As a second aim, therefore, the present study tries to further elucidate the differential impact of visual stimuli on the various functional–neuroanatomic subsystems of the supratemporal plane. Primary sensory regions of the cortex, characterized by the distinct cytoarchitectural features of "koniocortex" (see, e.g., Rockel, Hiorns, & Powell, 1980), have been assumed to process predominantly modality-specific input and, therefore, to be insensitive to afferent information transmitted via other perceptual channels (for reviews, see, e.g., Bernstein, Auer, & Moore, 2004; Schroeder et al., 2003). More recent data, however, point at a significant impact of visual information upon the activity of PAC both in monkeys (Brosch, Selezneva, & Scheich, 2005; Ghazanfar, Maier, Hoffman, & Logothetis, 2005) and in humans (Lehmann et al., 2006; Foxe & Schroeder, 2005; Molholm et al., 2002). With regard to the domain of spoken language, several functional imaging studies revealed visual speech information to elicit significant responses of PAC—even in the absence of an auditory signal (Pekkola

Table 1. Stimulus Types: Each of the Two Sets (Speech and Nonspeech, Applied in Different Runs) Comprises Three Silent Stimuli and Three Stimuli including an Acoustic Signal

	Speech	*Nonspeech*
Baseline condition	Static face	Static circles
Visual-only I	Video /ta/	Small motion
Visual-only II	Video /pa/	Large motion
Acoustic-only	Static face + Syl	Static circles + Tone
AV I	Video /ta/ + Syl	Small motion + Tone
AV II	Video /pa/ + Syl	Large motion + Tone

The silent static condition can be considered as an "empty" stimulus because it is identical with the display during the interstimulus and baseline intervals.

Syl = synthetic acoustic syllable, ambiguous between /ta/ and /pa/; tone = acoustic tone signal.

et al., 2005; Calvert & Lewis, 2004; but see Bernstein et al., 2002, for negative findings). For example, Calvert et al. (1997) were able to document hemodynamic activation both of primary and secondary acoustic cortex during silent lip reading.

The anterior and posterior components of secondary auditory cortex within the supratemporal plane show functional differences that have been interpreted in terms of a dual-pathway model of central-auditory processing (e.g., Rauschecker & Tian, 2000). The anterior pathway, the so-called "what" stream, appears to be associated with object- and content-related processing stages, whereas the posterior "where" path subserves auditory orientation and object localization in space. For example, investigations of the processing of species-specific vocalizations in Rhesus macaques found neurons of the anterior belt to be sensitive to call type, whereas the caudal belt predominantly responds to the spatial cues of these signals (Tian, Reser, Durham, Kustov, & Rauschecker, 2001). Within the human domain, the anterior "what" path has been postulated to play an important role in early stages of speech perception (Scott, 2005; Specht & Reul, 2003; Scott, Blank, Rosen, & Wise, 2000). Assuming visual speech information to enter the auditory system via secondary areas that are also engaged in auditory speech perception, it can be expected that the speaking face gives rise to hemodynamic activation particularly in anterior regions of secondary auditory cortex.

Taken together, this fMRI study tries to further define the topographical and functional characteristics of the brain mechanisms engaged in early and preattentive processing stages of AV speech perception. Two major research questions will be addressed:

(a) To what extent do uni- and bimodal speech and nonspeech stimuli show a differential impact upon primary and secondary auditory areas of the supratemporal plane?

(b) Which cerebral structures support the early phonetic/linguistic encoding of visual speech information?

METHODS

Subjects

Twenty right-handed native speakers of German (mean age = 28 years, *SD* = 8 years; 10 women) participated in this fMRI experiment. Self-reported right-handedness was confirmed by means of a short questionnaire (German version of the Edinburgh Handedness Inventory; see Oldfield, 1971), predicting hemispheric left lateralization of language functions in more than 90% of right-handers (Pujol, Deus, Losilla, & Capdevila, 1999). None of the subjects reported a history of relevant neurological or audiological disorders. The study had been approved by the Ethics Committee of the University of Tübingen.

Experimental Procedure

By and large, the present fMRI investigation was based upon the same stimulus materials and the same experimental design as a preceding whole-head MEG study (Hertrich, Mathiak, et al., 2009): Acoustic-only, visual-only, and AV stimuli—either of a speech (/pa/ or /ta/) or a nonspeech format (moving circle patterns and acoustic tone signals)—were applied in randomized order, speech and nonspeech items being presented in different runs. In order to direct the subjects' attention toward the auditory modality, a previous investigation of our group had asked the participants to respond to final pitch changes implemented in the acoustic component of both AV speech and AV nonspeech stimuli (Hertrich, Mathiak, et al., 2009). By contrast, the present study did not include behavioral responses in order to avoid any confounding impact of motor preparation or execution processes upon the evoked hemodynamic responses. Again, however, subjects were instructed to listen to the final pitch changes, and they consistently reported that these stimulus characteristics could be monitored during the experiment. Because falling and rising F0 contours (see below) were randomly distributed across the visual conditions, the extent of visible motion did not predict the forthcoming pitch shift.

Stimuli

Speech Stimuli

A synthetic syllable utterance (Figure 1A and D)—comprising an ambiguous voiceless stop consonant between /t/ and /p/ followed by the vowel /a/ (formant frequencies F1–F5 = 800, 1240, 2300, 3800, and 4500 Hz)—was generated by means of a formant synthesizer (Hertrich & Ackermann, 1999, 2007). Apart from high-frequency energy content associated with the burst of the stop consonants, the onset frequency of the second formant (F2)—starting below the F2 of vowel /a/ in case of /pa/ and above the F2 of /a/ in case of /ta/—represents the major difference between /pa/ and /ta/ at the acoustic level. Thus, the F2 transition within the aspiration period was modeled by a flat contour. A preliminary listening experiment was able to demonstrate that, indeed, the synthesized signal was perceived either as a /pa/ or /ta/ syllable, depending upon the synchronized visual speech information, that is, the video displaying /pa/ or /ta/ utterances. Fundamental frequency (F0) of the speech signal amounted to 120 Hz during the initial part of the vowel, extending across a time interval of up to 200 msec after stimulus onset. Following this stationary phase, F0 either began to rise or to fall (randomized order) by six semitones to either 170 or 85 Hz at stimulus offset (syllable duration = 300 msec). These stimulus-final pitch movements approximately correspond to the range of natural intonation of a male voice during speech production.

The visual speech stimuli were produced by a male person (Figure 1F) and consisted of two different video

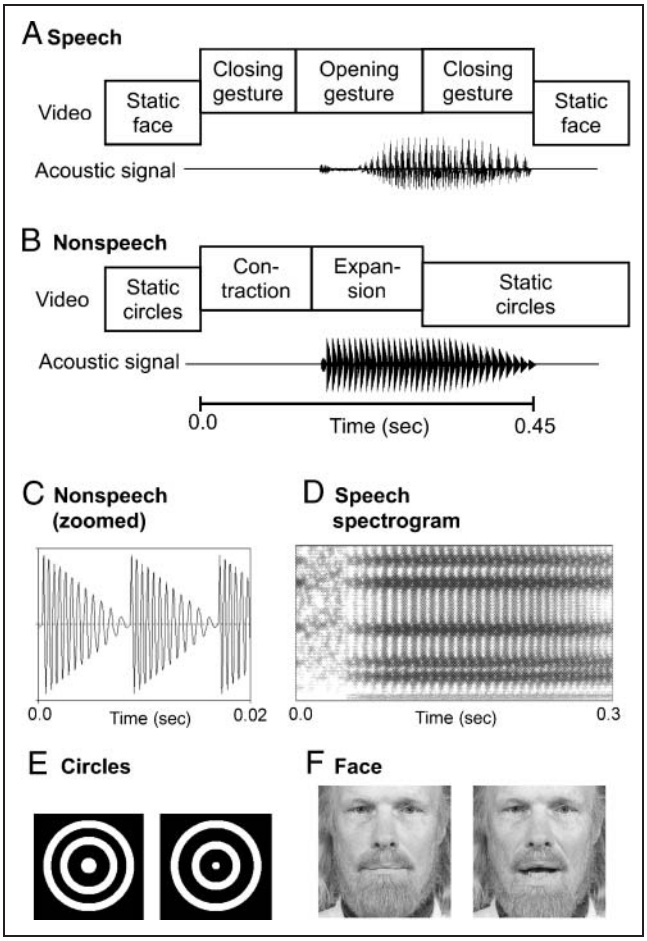

Figure 1. (A and B) Time course of audiovisual speech and nonspeech stimuli. (C) Zoomed 20-msec interval of the nonspeech oscillogram showing the down-tuned single-formant sweeps within each pitch period. (D) Spectrogram of the ambiguous speech signal that can be perceived as /pa/ or /ta/, depending on visual stimulation. (E) Example display of the visual nonspeech stimuli showing the interstimulus configuration (left circle pattern) and a contracted state of the inner circles during the visual motion stimuli at ca. 75% of the large movement excursion (right). (F) Face video stimuli showing the bilabial closure of the consonant /p/ (left) and the open mouth configuration during the vowel /a/ (right).

recordings (utterance of /pa/ and /ta/). These movement sequences were embedded into a larger frame, which displayed the speaker's static face during the interstimulus intervals (ISIs). Thus, the /pa/ and /ta/ videos could be concatenated into coherent stimulus series lacking any gaps or interruptions. The size of the speaker's face displayed during the experiment approximately corresponded to the visual impression of a person at a distance of 1 m.

Disregarding the final pitch movements, the speech runs comprised the following five stimulus configurations: (1) silent video of /ta/, (2) silent video of /pa/, (3) static face paired with the synthetic speech signal (= acoustic-only), (4) AV /ta/ video, (5) AV /pa/ video. Furthermore, null events were inserted showing the static face for the duration of an entire stimulus.

Nonspeech Stimuli

Repetitions of single-formant sweeps, giving rise to a perceived tone signal with a strong pitch (Figure 1B and C), served as the nonspeech auditory events. Within each pitch period, the formant was down-tuned from 2000 to 500 Hz and dampened to zero at its offset as shown in Figure 1C. Again, F0 amounted to 120 Hz across an initial time interval of 200 msec, followed by the same pitch movements as in case of the speech stimuli. Thus, regarding the perceived "intonational pattern," that is, the attentional focus of the stimuli, the nonspeech stimuli were quite similar to the synthetic speech signals. In contrast to the speech stimuli, the subjective auditory percept of the nonspeech tone signals did not change depending on visual stimulation.

Concentric circles (light blue color on a black background) served as the nonspeech analogues of the /pa/ and /ta/ video sequences. Visual motion comprised concentric contraction (duration = 150 msec; Figure 1B) and subsequent expansion (150 msec) of the inner structures of the circle pattern (Figure 1E), varying in amplitude and velocity ("small" vs. "large" nonspeech motion), in analogy to the visible mouth excursions during production of the syllables /pa/ (= large) and /ta/ (= small lip opening). In line with the speech condition, the same static display (left panel of Figure 1E) preceded and followed the movement sequences. The diameter of the entire circle pattern was approximately 80% of the horizontal head size in the video sequences.

Similar to the speech conditions, the nonspeech runs also comprised five stimulus configurations: (1) silent small movement, (2) silent large movement of the circle pattern, (3) static display paired with the acoustic tone signal (= acoustic-only), (4) small movement paired with the tone signal, and (5) large movement paired with the tone signal. Furthermore, null events were inserted showing the static circle pattern for the duration of an entire stimulus.

In terms of visibility of motion, the artificial nonspeech stimuli can be assumed to be at least as salient as the face stimuli due to the strong contrast between the circle pattern and the black background. Furthermore, due to the setting in the scanner, the display was always in the center of the subjects' visual field, and in case of both speech and nonspeech runs, the visual signal was temporally related to the onset of the acoustic signal. Therefore, the subjects must not be supposed to disregard the visual signals. However, the speaking face can be expected to be more salient regarding, first, its property as a natural sound source and, second, its function as a carrier of phonological information.

fMRI Sessions

Each fMRI session encompassed three speech and three nonspeech runs. During each single run, 20 repetitions of six different stimulus types (including the null event),

altogether 120 events, were presented in a pseudorandomized order (see Table 1). A video projector, in combination with a mirror system, presented the videos within the scanner.

The altogether 120 stimuli per run were presented at ISIs of 3.2–4.8 sec (randomly jittered in steps of 0.4 sec) during continuous image acquisition by means of a 3-Tesla scanner (TRIO, Siemens; TR = 2 sec, TE = 30 msec, 34 horizontal slices, interleaved, 64 × 64 voxels, resolution = 3.3 × 3.3 × 4 mm), using an event-related design (Figure 2). At the end of each fMRI session, a complete anatomical MRI dataset was obtained from each subject (256 × 256 × 256 voxels, resolution = 1 mm in each dimension).

Analysis of the fMRI Data

The evaluation of the functional imaging data encompassed the following steps of signal analysis:

(a) In order to display all brain regions engaged in the processing of the stimulus categories considered, the contrast of hemodynamic responses versus baseline was computed separately for the acoustic-only, visual-only, and AV events.

(b) A whole-head ANOVA was performed in order to identify cerebral areas sensitive to statistical interactions of the three experimental factors: (i) *type* of signal (speech/nonspeech), (ii) size of *visual motion* (no/small/large excursion), and (iii) presence/absence of an *acoustic* signal. For space reasons, the results of these analyses—primarily performed in order to obtain criteria for the selection of regions of interest (ROIs)—will be presented as supplementary materials (Suppl. 2 and 3).

(c) For the sake of identifying regions particularly sensitive to the visual /pa/–/ta/ distinction, a whole-head SPM analysis was performed, based upon the subtraction contrast visual /pa/ versus visual /ta/, pooled across visual-only and AV events. Again, this part of signal analysis has been added to the supplementary materials (Suppl. 4 and 5).

(d) ROI analyses were performed in order to characterize the differential impact of the experimental conditions on central-auditory processing and phonological encoding of the two video sequences displaying production of /pa/ and /ta/. The selected ROIs included (i) three areas within the supratemporal plane, separating hemodynamic responses of primary (Heschl's gyrus, HG) and secondary (anterior and posterior to HG) regions of the central-auditory system; (ii) two regions within the fusiform gyrus (FG), assuming this structure to represent a phonological interface of the visual system toward language areas, engaged in the perception of visual speech; and (iii) two neuroanatomically defined components of the IFG, exhibiting differences between visual /pa/ and /ta/. The exact boundaries of these ROIs were determined on the basis of neuroanatomical criteria as documented in the Results section. As far as possible, anatomical masks, defined in MNI space, were used, implemented in the MarsBaR Toolbox for SPM5 (Tzourio-Mazoyer et al., 2002) in order to allow for a comparison of the obtained data with other studies. Subsequent hypothesis-driven statistical analyses (ANOVAs and *t* tests), based upon these ROIs, tried to further delineate the (i) motion effects of (silent) visual-only stimuli, (ii) differential responses to speech versus nonspeech events, (iii) hyper- and subadditive AV interactions, and (iv) the visual /pa/–/ta/ contrast. The selected ROIs approximately correspond to already significant clusters of hemodynamic activation as determined on the basis of the preceding whole-head analyses. Therefore, no correction for multiple testing was considered necessary. The main purpose of the ROI analysis was to provide quantitative data about the relative strength of hemodynamic activation across conditions and brain regions. In order to address hemispheric lateralization effects, this analysis considered all selected ROIs at either side.

(e) A further ROI analysis was performed in order to determine eventual tonotopic effects of visual /pa/ and /ta/, that is, a differential impact of these stimuli upon medial and lateral parts of PAC (see Results section).

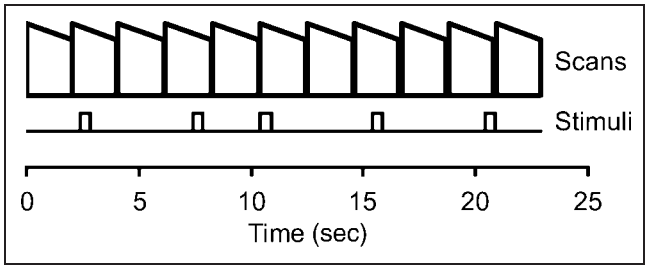

Figure 2. Time course of stimulus presentation.

RESULTS

Whole-head Subtraction Analyses: Uni- and Bimodal Stimulus Configurations versus Baseline

The first step of statistical data analysis included the calculation of the spatial distribution of the event-related BOLD responses—separately for speech and nonspeech runs—under the various experimental conditions (vs. the null event, i.e., the silent, static picture during continuous scanning) (Figure 3). Stimuli were pooled across large and small movement amplitudes (nonspeech signals) and /ta/ and /pa/ videos (speech events), respectively. Thus, three speech and three nonspeech contrasts (vs. baseline) could be computed: (1) acoustic-only, (2) visual-only, and (3) AV stimuli. A table listing all significant activation clusters is provided in the supplementary materials (Suppl. 1).

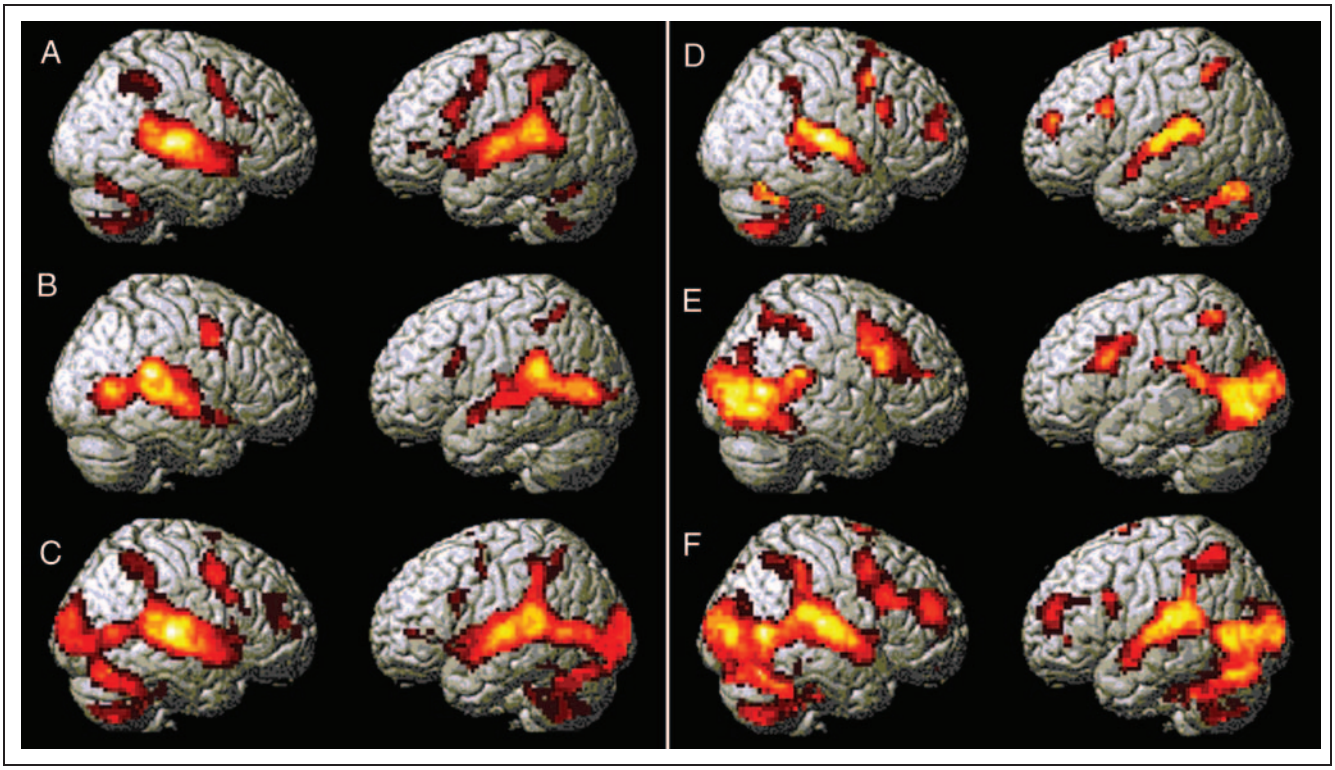

Figure 3. Activation versus baseline (= static display) shown for speech (A, B, C) and nonspeech (D, E, F) stimuli, separately for acoustic-only (A, D), visual-only (B, E), and audiovisual trials (C, F). Threshold: voxel-level $p < .001$ (uncorrected); cluster-level $p < .05$ (corrected).

(i) Acoustic-only trials yielded bilateral hemodynamic responses within the upper part of the temporal lobe both to speech (Figure 3A) and nonspeech stimulus configurations (Figure 3D). Furthermore, parietal and frontal regions showed significant activation foci under both conditions. Because cluster size depends upon the selected significance level, these data must be considered with some precautions. Nevertheless, the synthetic speech signals gave rise to more extensive BOLD signal changes than nonspeech signals within the temporal lobe of either side as well as within left inferior frontal and parietal regions—at any given threshold.

(ii) In the absence of acoustic stimulation, nonspeech silent visual motion yielded a broader hemodynamic activation pattern at the level of visual cortex (Figure 3E) than the articulatory gestures (Figure 3B). A reversed pattern could be observed within the superior temporal lobe. Inferior frontal activation clusters emerged both during the speech and nonspeech conditions, characterized in both instances by a more extensive distribution within the right hemisphere.

(iii) As expected, AV stimuli were found associated with a more widespread pattern of hemodynamic activation as compared to the two unimodal conditions, encompassing both the areas responsive to silent visual motion as well as acoustic-only stimuli (Figure 3C and F). By contrast to the visual-only speech condition, occipital AV responses reached the significance threshold, indicating that the acoustic component of the bimodal events has an enhancing effect on—at least some—cerebral visual areas (Figure 3C).

Impact of Visual Speech Information upon the Central-auditory System (ROI Analysis)

ROI Selection

At the cortical level, the processing of acoustic input is bound predominantly to the superior surface of the temporal lobe (supratemporal plane) which—in terms of its gyral organization—separates into three major components (from rostral to caudal): planum polare, the transverse gyrus or gyri of Heschl (HG), and planum temporale (e.g., Di Salle et al., 2003). Whereas HG houses PAC, the adjacent rostral and caudal regions (planum polare and planum temporale) serve as secondary acoustic areas involved in higher-order sensory and/or phonetic operations. Thus, three ROIs related to central-auditory processing were considered for further analysis (upper panels of Figure 4): PAC was defined as the anatomical HG mask of the MarsBaR toolbox, and two adjacent regions in rostral (anterior superior temporal gyrus, aSTG) and caudal (posterior superior temporal gyrus, pSTG) direction from PAC, were modeled as spheres (diameter = 1 cm). The centers of these regions (Table 2) were determined on the basis of a whole-head ANOVA (see Suppl. 2 and 3), which evaluated the interactions among the experimental factors.

Table 2. Center Coordinates (MNI, x = Lateral, y = Front/Back, z = Height) and Radius (mm) of the Spherical ROIs Shown in Figures 4 and 6

	Left Hemisphere				*Right Hemisphere*			
Region	*x*	*y*	*z*	*Radius*	*x*	*y*	*z*	*Radius*
aSTG	−57	−9	−3	10	63	−3	−3	10
pSTG	−60	−45	12	10	66	−39	15	10
medial HG	−51	−15	3	7	51	−15	3	7
lateral HG	−63	−6	6	7	63	−6	6	7
FG_{37}	−30	−47	−15	13	33	−46	−16	13
FG_{19}	−34	−75	−16	13	31	−74	−14	13

The aSTG ROI was derived from the interaction of *acoust* (presence or absence of an acoustic signal) with *type* (speech vs. nonspeech stimuli), whereas the pSTG region corresponds to the interaction of *acoust* with *visual motion* (static, small, large).

Heschl's Gyrus

The upper panels of Figure 5 display the strength of the hemodynamic responses across the 10 stimulus conditions (percent signal change, baseline = 0). Since a previous study had found gender effects of visually induced PAC activation (Ruytjens, Albers, van Dijk, Wit, & Willemsen, 2007), male and female data are displayed separately. Most noteworthy, the visual-only condition yielded a significant main effect [$F(1, 18) = 8.87$, $p < .01$] of the between-subjects factor *gender* (column pairs 1, 2, 6, 7, from the left). More specifically, only male participants showed significant hemodynamic PAC activation, restricted, however, to visual speech (!) stimuli (t tests; significance level $p < .01$ for visual /ta/ within both hemispheres and for /pa/ at the left side, $p < .05$ for /pa/ within the right hemisphere). Although the differential syllable effect (visual /pa/ versus /ta/) did not reach the significance level, hemodynamic activation tended to be higher in response to visual /ta/ as compared to /pa/, in spite of larger lip movements in the case of /pa/. This /ta/–/pa/ difference could reflect the broader acoustic spectrum of syllable /ta/ in terms of an enlarged content of higher frequencies in natural speech. In addition to these subtle differences in response strength, the spatial distribution of hemodynamic activity, as a consequence, must also be expected to differ between /pa/ and /ta/. Therefore, a small-volume analysis was performed, using HG as an anatomical mask. At a significance level of $p < .001$ (uncorrected), the right hemisphere showed activation clusters in association both with visual /pa/ and /ta/. Most noteworthy, the peak of the /ta/ cluster had a more medial position ($x = 51$, $y = -15$, $z = 3$) as compared to visual /pa/ ($x = 63$, $y = -6$,

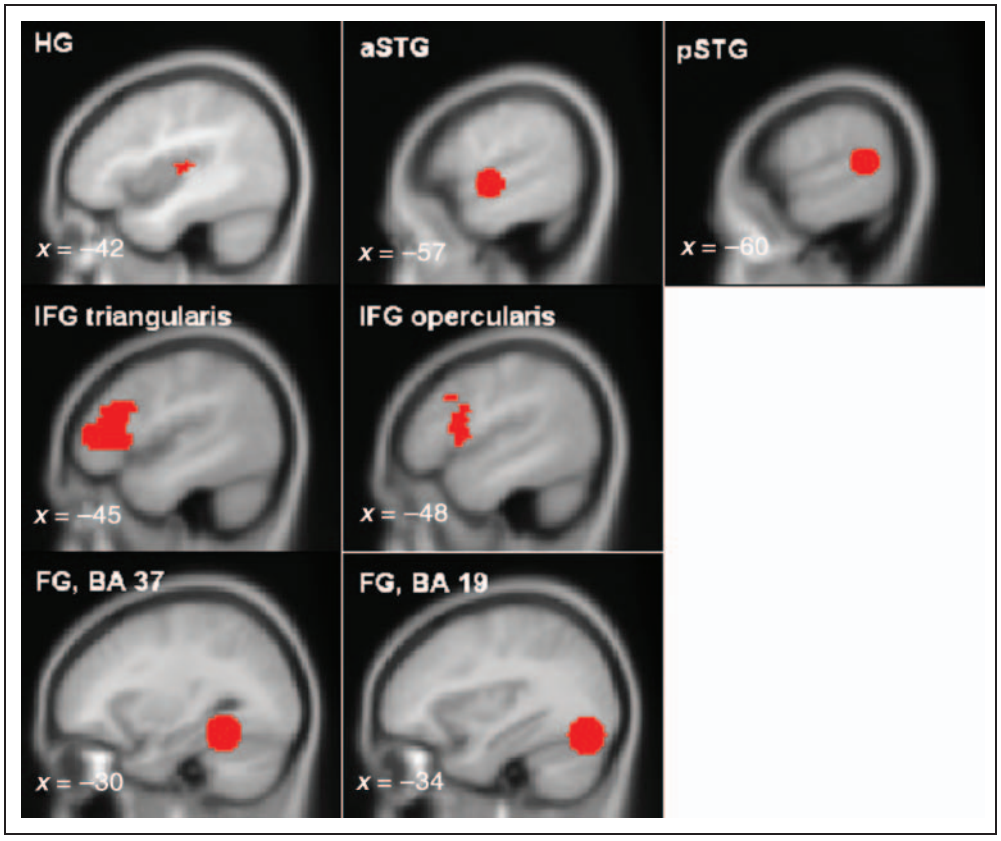

Figure 4. Definition of the ROIs, shown in sagittal slices for the left hemisphere: FG = fusiform gyrus; IFG = inferior frontal gyrus; HG = Heschl's gyrus; aSTG = anterior superior temporal gyrus; pSTG = posterior superior temporal gyrus, BA = Brodmann's area. Three regions (HG, IFG triangularis, and IFG opercularis) were taken from anatomical masks (see text), the remaining four ROIs were defined as spheres specified in Table 2.

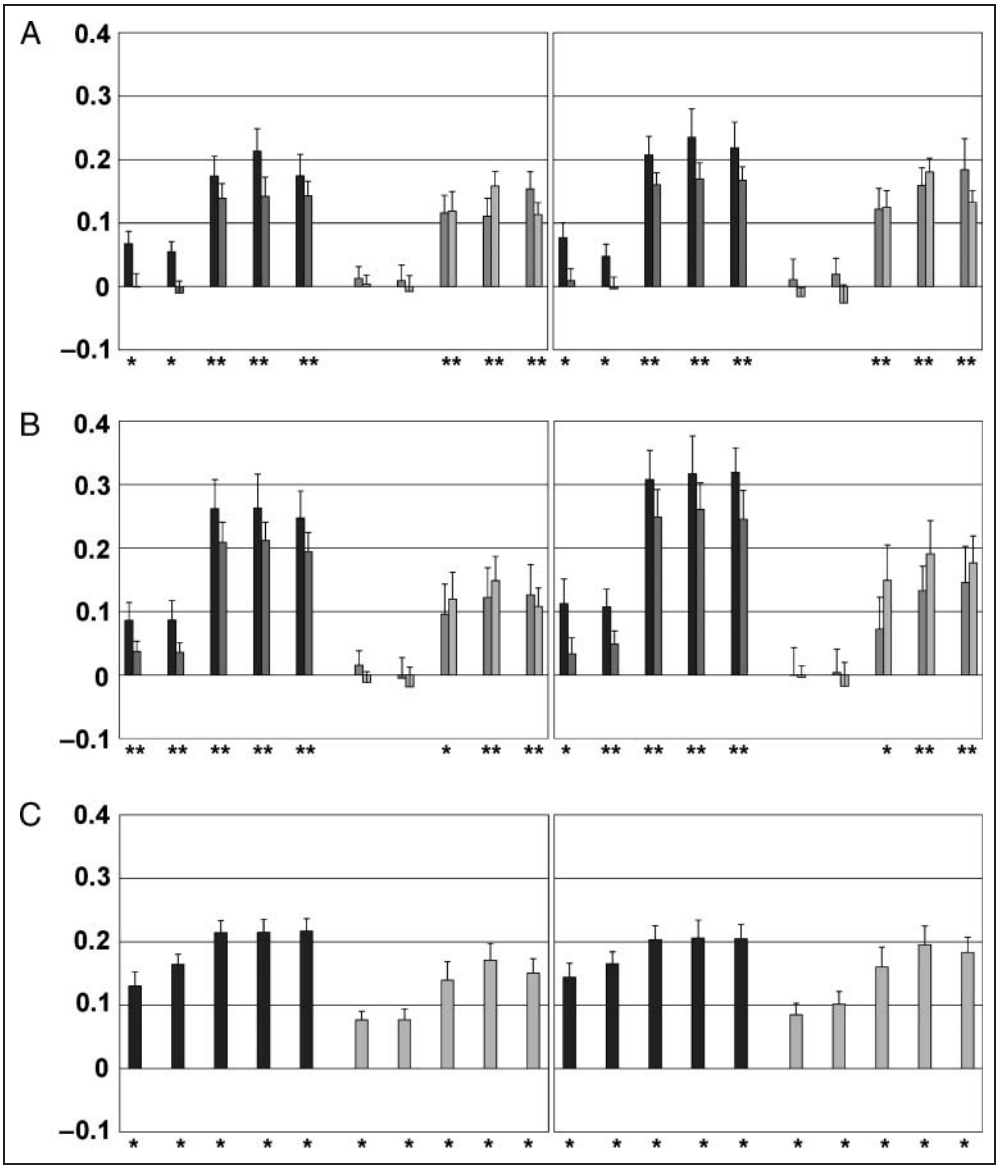

Figure 5. Condition-dependent strength of hemodynamic responses in STG. (A) PAC, (B) aSTG, (C) pSTG. Left and right panels refer to the respective hemisphere. Within each panel, the left five columns or pairs of columns correspond to the speech and the right five columns to the nonspeech conditions in the following order (from left to right): (1) silent small (or /ta/ in case of speech), (2) silent large movement (or /pa/ in case of speech), (3) acoustic-only, (4) AV small, (5) AV large movement. For PAC and aSTG, the columns are split into pairs, corresponding to the male (left) and female (right) subgroups. Error bars indicate the standard error of the mean across subjects; asterisks below the bars indicate significant activation versus baseline ($p < .05$).

$z = 6$). Both maxima then were used as the centers for an additional ROI analysis, based upon spherical ROIs of a diameter of 14 mm each. Because the left hemisphere did not show any significant clusters even at a significance level of $p < .01$, mirror-symmetric ROIs were applied to the left hemisphere. A repeated measures ANOVA with the factors *location* (medial/lateral), *hemisphere*, and *syllable* (/pa/ vs. /ta/) revealed, apart from a significant main

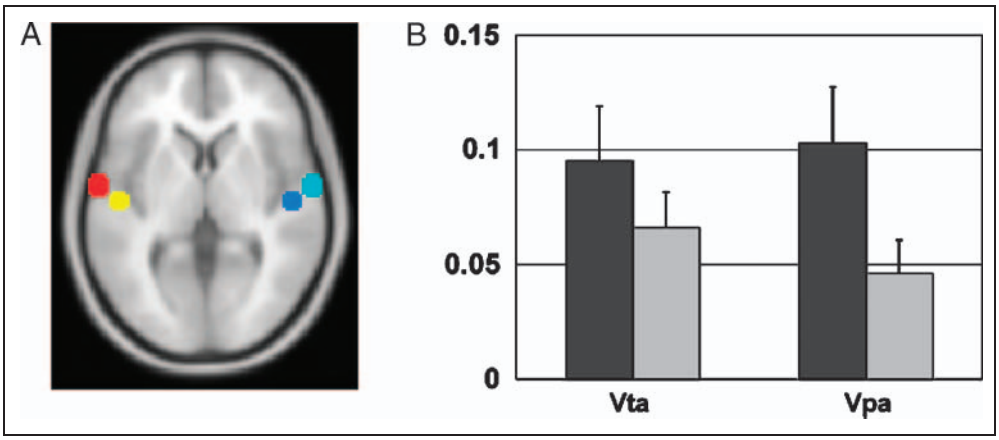

Figure 6. (A) Definition of bilateral symmetric ROIs representing lateral and medial parts of PAC (coordinates are presented in Table 2). (B) Percent signal change in lateral (black) and medial (gray) portions of HG in response to visual /ta/ and /pa/. Note the more medial activation in case of /ta/, reflecting tonotopic characteristics of the different consonants induced by visual stimulation.

effect of *location* [$F(1, 19) = 5.44, p < .05$], a tendency toward an interaction between the two factors [$F(1, 19) = 3.78$, $p = .067$; medial: ta > pa, lateral: pa > ta; see Figure 6B].

An often addressed aspect in previous studies on AV interactions was whether auditory and visual stimulation give rise to additive, subadditive, or hyperadditive (superadditive) brain responses (Stevenson, Geoghegan, & James, 2007; Calvert, Hansen, Iversen, & Brammer, 2001). Lehmann et al. (2006), for example, reported primary sensory regions to respond in a subadditive manner, whereas secondary regions may show more complex activity patterns. In order to test for AV subadditivity effects at the level of PAC, a repeated measures ANOVA was performed, based upon the term (V + A − AV), that is, the sum of unimodal responses minus activation by AV stimuli, as the dependent variable. This measure revealed a significant main effect of type [$F(1, 18) = 4.39, p < .05$] with a positive mean value for the speech and a negative one for the nonspeech condition, indicating subadditivity in case of AV speech perception, and hyperadditivity in case of nonspeech stimulus processing. In particular, the PAC response to the acoustic tone signal tended to be enhanced by visual nonspeech motion, although silent nonspeech motion alone did not yield any hemodynamic PAC activation.

Anterior Component of the Superior Temporal Gyrus

Figure 5B displays the stimulus-related hemodynamic effects within aSTG, demonstrating a similar response pattern to visual-only events as described for HG (see preceding paragraph). Similar to PAC, therefore, male and female data are shown separately. Although failing significance, a tendency for a gender effect [$F(1, 18) = 3.09, p = .096$] showed the same direction as in HG (male > female). Furthermore, a significant main effect of speech/nonspeech was observed [$F(1, 18) = 9.02, p < .01$; speech > nonspeech]. As compared to baseline, the speech stimuli—but not the moving cycle patterns—gave rise to significant activation in aSTG (see asterisks in Figure 5).

In order to assess the differential influence of visual motion at the level of the aSTG in the presence of an acoustic signal, an additional ANOVA was performed, including all responses to speech and nonspeech acoustic-only and AV stimuli. The intersubject factor Gender and the intrasubject factors Type (speech/nonspeech), Motion (static, small, large), and Hemisphere served as independent variables. Apart from main effects of Type [$F(1, 18) = 21.58, p < .001$; speech > nonspeech], Hemisphere [$F(1, 18) = 5.05, p < .05$; right > left], and Motion [$F(2, 17) = 3.72, p < .05$; small > static], a significant Hemisphere × Motion interaction could be noted [$F(2, 17) = 11.29, p < .01$]. Figure 5B shows that visual motion—in the presence of an acoustic signal—gave rise to enhanced right-hemisphere aSTG activity, particularly in case of the nonspeech events, although silent visual nonspeech motion did not elicit any significant aSTG responses versus baseline.

Similar again to HG, subadditivity at the level of the aSTG depended upon signal type [main effect of Type, $F(1, 19) = 8.65, p < .01$], with larger values of the term (V + A − AV) for the speech as compared to the nonspeech condition. Furthermore, a significant main effect of Hemisphere emerged [$F(1, 19) = 5.38, p < .05$; left > right]. As concerns the speech domain, both visual /pa/ and /ta/ showed significant subadditivity at the left and /pa/ also at the right side—as indicated by a positive sign of the term (V + A − AV) (*t* tests, $p < .05$). By contrast, hyperadditivity (negative sign) for nonspeech stimuli tended to be more pronounced within the right than the left hemisphere.

Posterior Component of the Superior Temporal Gyrus

The lower panels of Figure 5 display the stimulus-related activity (vs. baseline) within pSTG. Silent visual motion elicited—both within the context of speech and nonspeech events—significant hemodynamic activation of this area (*t* tests, $p < .001$), in the absence of any gender effects. Apart from a main effect of Type [$F(1, 19) = 9.18$, $p < .01$; speech > nonspeech], a significant three-way Type × Hemisphere × Motion (small/large) interaction could be noted: The left hemisphere showed a larger /pa/–/ta/ difference, whereas nonspeech motion was associated with the reverse effect, that is, a larger impact of movement size within the right hemisphere (Figure 5C).

The stimuli with an acoustic signal component, that is, acoustic-only events and AV configurations, yielded a significant interaction between TYPE (speech/nonspeech) and Hemisphere [$F(1, 18) = 9.15, p < .01$; speech: left > right, nonspeech: right > left]. Finally, the subadditivity measure (V + A − AV) consistently showed a positive sign, both for speech ($p < .001$) and nonspeech ($p < .05$) stimuli. However, verbal utterances were associated with higher values than the nonspeech signals [main effect of *type*: $F(1, 19) = 6.53, p < .05$].

Impact of Visual Motion Size upon Inferior Occipito-temporal and Frontal Areas (ROI Analyses)

ROI Selection

A preceding MEG study of our group had shown an overproportionally strong magnetic field in response to visual /pa/ as compared to /ta/, the calculated source lying outside the central-auditory system (Hertrich, Mathiak, et al., 2009). These findings were interpreted in terms of a suprathreshold representation of the bilabial phonological gesture, whereas the visual /ta/ gesture, by contrast, appeared to be suppressed. In order to further delineate the brain regions accounting for these visual–phonological effects, a whole-head ANOVA was performed, addressing the subtraction contrast visual /pa/ versus /ta/. Considering the overall similarity of these two stimulus categories, only

subtle effects can be expected. Therefore, the significance threshold was set to $p < .005$ at voxel and $p < .05$ at cluster level, with a minimum cluster size of 59. The supplementary materials (Suppl. 4 and 5) include the detailed results of this step of analysis; the reversed contrast (i.e., visual ta > pa) did not yield any significant suprathreshold clusters at the selected significance level. As the most relevant aspect of this ANOVA with respect to the subsequent ROI analysis, activation clusters emerged within two cortical areas that can be assumed to engage in the processing of visual speech information, namely, IFG and FG. For the sake of comparability with other studies, neuroanatomical masks implemented in the MarsBaR toolbox (see Methods) were used for the delineation of two subregions of the IFG (pars triangularis and pars opercularis)—instead of a determination of ROIs on the basis of activation clusters (middle panels of Figure 4).

Previous studies reported different phonology-related functions in association with anterior and posterior parts of FG, corresponding to Brodmann's area (BA) 37 and BA 19, respectively (e.g., Dietz, Jones, Gareau, Zeffiro, & Eden, 2005). For example, there is some evidence that these two FG subregions connect visual face processing and the extraction of visual phonological information (Dien, 2009; Blonder et al., 2004). Thus, two ROIs were defined as the intersections of FG with BA 19 (FG_{19}) and BA 37 (FG_{37}), modeled as spheres with a radius of 13 mm (Table 2 and lower panels of Figure 4). To detect eventual lateralization effects of hemodynamic activation, both hemispheres were considered for analysis, although significant clusters for /pa/ > /ta/ at the applied threshold—as determined on the basis of the whole-head ANOVA—were restricted to the right side.

Intersection of Fusiform Gyrus with Brodmann's Area 37 (FG_{37})

Hemodynamic activation of anterior FG (intersection with BA 37) showed a significant main effect of motion size [large > small; $F(1, 19) = 4.54, p < .05$] as well as a strong impact of the factor hemisphere [right > left; $F(1, 19) = 19.06, p < .001$; ANOVA based upon the visual-only stimuli]. Neither visual /ta/ nor AV /ta/ utterances yielded any significant hemodynamic responses (t test, $p > .2$, for both hemispheres; see Figure 7A, asterisks below the bars), whereas the response to visual /pa/ reached the significance level within the right hemisphere. Acoustic-only stimuli did not yield even a tendency of BOLD signal changes within this area. Finally, acoustic stimulation did not interact with visual motion in terms of the subadditivity measure ($p > .1$ for all subconditions).

Intersection of Fusiform Gyrus with Brodmann's Area 19 (FG_{19})

The more posterior FG ROI (intersection with BA 19) also showed a significant visual motion effect [$F(1, 19) = 5.08$, $p < .05$], in the absence, however, of any differences between the two hemispheres (Figure 7B). Acoustic stimulation gave rise to hemodynamic activation above baseline (t test, $p < .05$) at the right side in case of speech stimuli

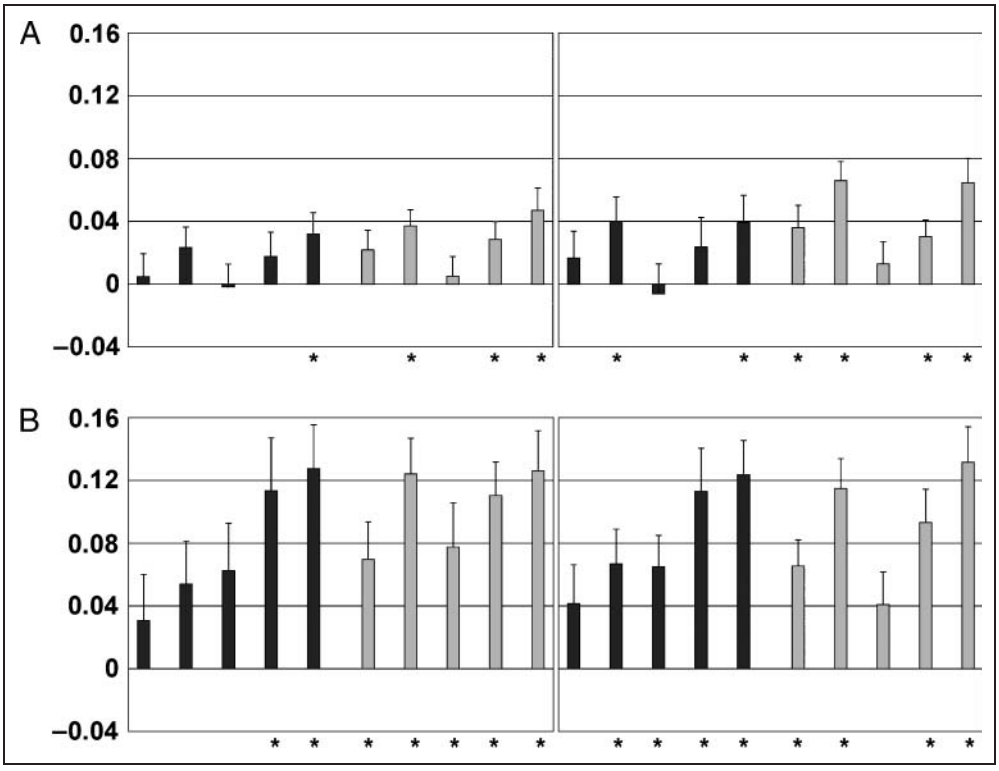

Figure 7. Condition-dependent strength of hemodynamic responses in two regions of FG. (A) Intersection of FG with BA 37 (FG_{37}); (B) intersection of FG with BA 19 (FG_{19}). Left and right panels refer to the respective hemisphere. Within each panel, the left group of five columns corresponds to the speech and the right one to the nonspeech conditions in the following order (from left to right): (1) silent small (or /ta/ in case of speech), (2) silent large movement (or /pa/ in case of speech), (3) acoustic-only, (4) AV small, (5) AV large movement. Error bars indicate the standard error of the mean across subjects; asterisks below the bars indicate significant activation versus baseline ($p < .05$).

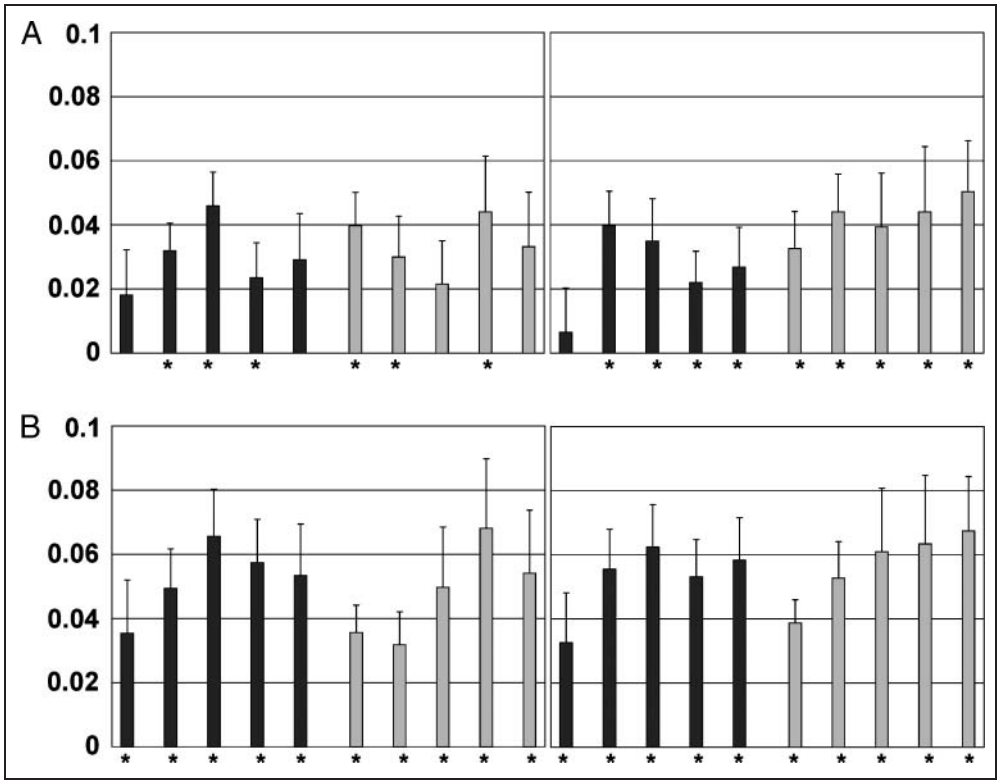

Figure 8. Condition-dependent strength of hemodynamic responses in inferior frontal ROIs. (A) IFG pars triangularis, (B) IFG pars opercularis. Left and right panels refer to the respective hemisphere. Within each panel, the left group of five columns corresponds to the speech and the right one to the nonspeech conditions in the following order (from left to right): (1) silent small (or /ta/ in case of speech), (2) silent large movement (or /pa/ in case of speech), (3) acoustic-only, (4) AV small, (5) AV large movement. Error bars indicate the standard error of the mean across subjects; asterisks below the bars indicate significant activation versus baseline ($p < .05$).

and within the left hemisphere in case of nonspeech events. An interaction with visual motion—in terms of the subadditivity measure (t test, $p > .1$ for all subconditions)—did not emerge.

Inferior Frontal Gyrus, Pars Triangularis (IFGtr) and Pars Opercularis (IFGop)

Hemodynamic responses within the more anterior ROI in the IFG (IFGtr) to visual-only stimuli were characterized by an interaction of Motion (small/large) and Hemisphere [$F(1, 19) = 10.49, p < .01$; Figure 8A]. Post hoc t tests revealed the significant motion effect to be exclusively bound to the right hemisphere (/pa/ > /ta/; $p < .05$). In addition, a significant subadditivity (V + A < AV) effect of the speech stimuli emerged [$t(19) = 3.17, p < .01$]. Hemodynamic responses to AV speech stimuli even tended to be smaller than those to acoustic-only trials, indicating a suppressive effect of visual motion on auditory-evoked neural activity.

Similar to IFGtr, the opercular component showed also a Hemisphere × Motion interaction of the responses to silent motion stimuli [$F(1, 19) = 4.49, p < .05$] and a differential motion effect restricted to the right hemisphere [$F(1, 19) = 5.23, p < .05$; large > small, pooled across speech and nonspeech; Figure 8B]. Only speech stimuli yielded a significant subadditivity effect (V + A – AV; $p < .001$), the nonspeech events showed, however, a similar tendency ($p = .08$).

DISCUSSION

Summary of Results

As its major aim, the present fMRI study sought to extend a previous whole-head MEG investigation, based—more or less—upon the same experimental design and the same test materials (see Hertrich, Mathiak, et al., 2009). More specifically, this approach was expected to provide further information on the topographic and functional characteristics of the cerebral network subserving the processing and, more specifically, the phonetic/phonological encoding of visual speech stimuli. In order to distinguish speech-related effects from an unspecific supramodal influence of visual motion, the follow-up study included a nonspeech control condition. The obtained findings point at (i) a significant impact of articulatory speech gestures—but not the nonspeech visual objects considered—upon aSTG and (in male subjects) PAC; (ii) differential scaling effects of speech and nonspeech visual motion within inferior frontal areas; and (iii) predominantly subadditive AV interactions across most brain areas taken into account (except, e.g., FG).

Impact of Speech/Nonspeech Visual Information upon Auditory Cortex

Primary Auditory Cortex (Heschl's Gyrus)

Significant hemodynamic responses of PAC to silent (!) displays of movement sequences were found restricted to the speech domain (i.e., the speaking face). In line with a

preceding fMRI investigation (Bunzeck, Wuestenberg, Lutz, Heinze, & Jäncke, 2005), the nonspeech conditions of the present study did not elicit any significant BOLD signal changes within this area. This discrepancy might reflect differences in cross-modal ecological validity of the two AV stimulus categories: In case of speech events, the auditory signal is causally linked to the visual cues, that is, the speaking face represents the natural source of the speech sounds, whereas the association of moving circle patterns with tones lacks a similar causal connection in our experience of the physical world.

Experiments in monkeys found visual scenes to elicit, among others, neural activity in distinct fields of auditory cortex including PAC (Kayser, Petkov, Augath, & Logothetis, 2007). The present study was able to detect a more medial localization of hemodynamic responses to silent (!) /ta/ utterances as compared to the respective /pa/ syllables within primary auditory areas of the supratemporal plane. As a rule, the acoustic (!) signal of /ta/ productions is characterized by more pronounced high-frequency components. The distribution of visually induced PAC responses to /ta/ and /pa/ utterances, thus, is compatible with the tonotopic organization of primary auditory areas, that is, a more medial representation of higher and a more lateral "location" of lower-frequency bands (Bendor & Wang, 2006). Conceivably, this pattern of hemodynamic activity reflects auditory imagery processes as, for example, discussed in Kraemer, Macrae, Green, and Kelley (2005). Differential hemodynamic PAC effects could not, however, be observed if the visual /ta/ and /pa/ sequences were paired with the ambiguous acoustic signal, although the two different AV configurations give rise to distinct auditory percepts each /pa/ and /ta/. The absence of a visual /pa/ versus /ta/ effect in case of AV stimuli may be due to the fact that the acoustic signal was always the same ambiguous intermediate item between /pa/ and /ta/. Obviously, thus, the actual acoustic stimulus component represented the dominant factor at the level of PAC, and the perceptually relevant AV fusion processes did not rely on visually induced modulation of PAC activity.

Ruytjens et al. (2007) had observed significant HG activation during silent lip reading in women, but not in men. In line with these findings, behavioral studies reported a higher impact of the visual channel on auditory speech perception in female subjects (Traunmüller & Öhrström, 2007). Presumably, these effects are due to gender differences in interhemispheric connectivity (see, e.g., De Gennaro et al., 2004). For example, a dichotic listening study (Hertrich, Mathiak, Lutzenberger, & Ackermann, 2002) had revealed a stronger tendency toward fusion errors, that is, the combination of phonetic features presented to the left and right ear to a single perceived phoneme, in women, concomitant with a less strong right-ear advantage. By contrast, the present study found hemodynamic PAC responses to silent visual articulation to be restricted to male participants—an observation at variance with the data of the preceding investigations referred to. However, gender effects during speech perception appear to interact with a multitude of experimental conditions such as attentional setting and processing strategies (see, e.g., Schirmer, Kotz, & Friederici, 2005). In contrast to those previous studies which had documented higher transmodal signal integration in women, the present experiment did not require explicit attention or behavioral responses to the segmental phonetic structure of the stimuli. This design could have kept female subjects from the application of a phonological integration strategy.

As an alternative explanation, hemodynamic PAC activation during silent speech motion might reflect uncontrolled central-auditory processing in case of directed attention toward an "empty" acoustic channel. The subjects' high familiarity with speaking faces during everyday communication in association with the highly redundant stimulus design of the present study might facilitate such effects. Under these conditions, PAC responses could reflect higher-order (e.g., phonological) imagery rather than a direct impact of visual information upon the central-auditory system.

Secondary Auditory Areas of the Supratemporal Plane

Differential hemodynamic activation patterns emerged rostral (aSTG) and caudal (pSTG) to PAC. Similar to PAC, aSTG exclusively responded to visual speech stimuli, whereas pSTG was found sensitive to silent nonspeech motion as well. Previous studies have shown that the processing of meaningful auditory events such as species-specific calls of nonhuman primates or human speech signals specifically engage aSTG (Altmann, Bledowski, Wibral, & Kaiser, 2007; Rauschecker & Tian, 2000). Left-hemisphere anterolateral STG, furthermore, has been shown to particularly respond to intelligible consonantal bursts as compared to incomprehensible control sounds matched for spectro-temporal complexity (Obleser, Zimmermann, Van Meter, & Rauschecker, 2007). In the light of the dual-pathway model ("what" vs. "where" stream) of central-auditory processing (e.g., Rauschecker & Tian, 2000), the results of the present study indicate the neural responses to the visual display of articulatory gestures to recruit—besides PAC—the rostral ("what") projections of the central-auditory system. This suggestion is also in line with the observation that auditory imagery, for instance, in case of music, engages the anterior rather than the posterior parts of the central-auditory system (Rauschecker, 2001) and that hemodynamic activation of rostral STG (planum polare) is related to the representation of phonological information (Hasson et al., 2007).

In contrast to aSTG, both speech and nonspeech visual stimuli yielded significant hemodynamic pSTG activation. Griffiths and Warren (2002) postulated the planum temporale to be involved in the spectro-temporal analysis of any complex auditory stimulus, among others, in the assignment of those signals to visual objects and events in

space. Conceivably, thus, activation of pSTG, operating across both the speech and nonspeech domains, reflects the tight temporal association between the acoustic and visual data stream imposed by the experimental design, for example, the fact that visual motion predicted the time of acoustic stimulus onset, building up a respective "expectation" even in the case of the interspersed visual-only stimuli. These suggestions might explain the bilateral pattern of hemodynamic activation—restricted to the supratemporal regions posterior to HG—during auditory verbal "imagery" as reported by a recent fMRI study (Jäncke & Shah, 2004): The authors had asked their subjects to produce auditory verbal imagery in response to light flashes (after some training), that is, visual stimuli that do not represent a valid sound source, whereas the verbal utterances of the present study were associated with a speaking face, that is, the display of a realistic source of a speech signal. Thus, pSTG activation might be related to cross-modal temporal coordination of afferent inputs, an assumption compatible with the idea of oscillatory synchronization across channels (Kayser, Petkov, & Logothetis, 2008; Kayser et al., 2007). Based upon the observation of hierarchically coupled neuronal oscillations, Schroeder et al. (2008) suggest, in a similar vein, that temporally correlated nonacoustic input may shift the ongoing activity of the central-auditory system toward an "ideal excitability phase," giving rise to enhanced processing of the incoming acoustic signals.

The observed differential hemodynamic activation patterns at the level of the supratemporal plane indicate the information flow from the visual system toward the anterior pathway of the central-auditory system to depend upon the ecological validity of the visual signal as a potential sound source. Thus, neural activity of aSTG (and PAC in some subjects) seems to be associated with highly automatized AV processes, depending upon cross-modal experience. So far, it must remain unsettled whether this process is specific to speech or whether it is bound to long-term learning with respect to the visual recognition of any ecologically valid sound sources such as environmental events, animal sounds, or musical instruments.

The hemodynamic responses to visual speech movements at the level of PAC and aSTG showed a strong subadditive effect and were found largely suppressed or inhibited in the presence of an actual acoustic speech signal. Our previous MEG study (Hertrich, Mathiak, et al., 2009) had shown a similar impact of speaking faces upon the auditory M100 field. Tentatively, thus, the observed subadditive effect could be assigned to early, preattentive central-auditory processes underlying, for example, auditory M100 responses. It should be reminded here that the visual cues preceded the onset of the acoustic signal—or the respective imagined auditory event in case of visual-only stimuli—by ca. 150 msec. Considering, first, the strong subadditivity of AV speech interactions and, second, its early emergence in relation to the beginning of the acoustic speech signal, these early cross-modal operations cannot be held responsible for higher-order phonological phenomena such as the McGurk effect (MacDonald & McGurk, 1978; see below).

Cerebral Networks Subserving the Scaling of Visual Speech Movements

As a major aim, the present study tried to further delineate early preattentive stages of the phonetic/phonological encoding of visual speech stimuli. In order to distinguish these specific processes from more general effects of visual motion, a nonspeech control condition had been introduced. Based on our previous MEG study (Hertrich, Mathiak, et al., 2009), it was hypothesized that a visual–phonological representation should translate into an overproportional response to visual /pa/ as compared to /ta/, based on the assumption of "coronal underspecification" in the mental lexicon (De Lacy, 2006; Harris & Lindsey, 1995; Lahiri & Marslen-Wilson, 1991; Avery & Rice, 1989). The concept of underspecification postulates the presence of "default" features that do not need to be specified. As a crucial argument in favor of coronal underspecification, the respective sounds show the tendency to become assimilated by the following labial or dorsal sounds as, for example, in the word "encode," which often is pronounced with a velar nasal [əŋkoud] instead of [ənkoud]. Our MEG study had documented a strong field component evoked by the video displaying articulation of /pa/ while responses to visual /ta/ were suppressed. Tentatively, the respective source could be assigned to posterior insular cortex. The present fMRI study revealed the most pronounced /pa/ > /ta/ effects within inferior frontal areas (IFGtr) as well as FG. Furthermore, this effect was found to be lateralized toward the right hemisphere.

Besides the phonological loop, acting upon verbally encoded data, Baddeley (2003) proposed a second modality-specific component of working memory, namely, the visual sketch pad, subserving short-term storage of feature-based visual information. More specifically, a positron emission tomography study found, among others, FG and IFG to support a "face working memory" (Courtney, Ungerleider, Keil, & Haxby, 1996). Furthermore, repetitive transcranial magnetic stimulation provided some evidence for lateralization of the visual working memory toward the right hemisphere (Hong, Lee, Kim, Kim, & Nam, 2000). Against this background, the observed hemodynamic activation of right-hemisphere FG and IFGtr, in response to the (silent) lip movements associated with the production of /pa/, might reflect transient maintenance of visual speech information in working memory.

A variety of data point at a significant contribution of FG—part of the ventral, object-related data stream of the visual system—to the encoding of visual speech features. In particular, the intersection of FG with BA 37 is known to operate as an interface between the visual system and the

mental lexicon during reading—as evidenced by clinical and neurolinguistic studies (Cao, Bitan, & Booth, 2008; Dietz et al., 2005; McCandliss, Cohen, & Dehaene, 2003; Sakurai et al., 2000). Furthermore, right IFG is found activated in subjects performing phonological short-term memory tasks based upon visual (orthographic) stimuli (Sumiyoshi, Matsuo, Nakai, & Kato, 2003) and in psychotic subjects during verbal hallucinations (Sommer et al., 2008), that is, another form of nonauditory phonological representation of speech material. A further argument in favor of this right-hemisphere visual working memory hypothesis can be derived from the observation that subjects paying attention to AV speech are particularly sensitive to the speaker's right hemiface, located in the recipient's left visual hemifield (Guo, Meints, Hall, Hall, & Mills, 2009; Jordan & Thomas, 2007). Such a perceptual asymmetry might have an additional functional relevance as articulatory movements tend to be asymmetric as well, with higher expressivity in the speaker's right hemiface (Nicholls & Searle, 2006). Finally, a recent study investigating the recruitment of visual cortex in a blind subject during the perception of ultra-fast synthetic speech found task-related activations in right rather than left primary visual cortex (Hertrich, Dietrich, Moos, Trouvain, & Ackermann, 2009), indicating a contralateral mechanism linking the auditory and the visual system during speech perception.

Investigations of the mechanisms of the McGurk effect ("McGurk fusion": visual /ga/ paired with acoustic /ba/ is perceived as /da/) provide further insights into the contribution of right hemisphere structures to the processing of visual speech information. Under these conditions, Diesch (1995) observed visual hemifield effects which point at two segregated and differently lateralized data streams during visual speech perception. If the visual component of the incongruent AV signal was presented to the left hemifield/right visual cortex, subjects more likely experienced the McGurk effect, that is, a single, fused consonant. By contrast, if the visual information was transmitted via the right hemifield (= left hemisphere), subjects tended to directly insert both visual and acoustic information into their percept, resulting in a heard consonant cluster ("combination case"). These findings are consistent with the assumption that in the "combination case," the visually presented articulatory movements give rise to direct auditory imagery, whereas the McGurk fusion requires an intermediate right hemisphere representation in visual working memory. Accordingly, the MEG study of Hertrich et al. (2007) investigating responses to visual speech deviants in an AV oddball design found a visual mismatch response lateralized to the right hemisphere followed by a later left-hemisphere component with a source in or near the auditory system. Based on the present results as well as our previous MEG study (Hertrich, Mathiak, et al., 2009), it can be assumed that at least part of the direct impact of visual features on auditory cortex activity is inhibited when an actual acoustic signal is presented.

In contrast to FG and IFG, visual /pa/ and /ta/ stimuli yielded hemodynamic activation of an approximately equal strength in regions associated with central-auditory processing (STG and HG). Most presumably, these findings can be explained by the fact that the size of visual speech motion was irrelevant with respect to both the prediction of acoustic onset and "loudness" of an imagined auditory event. Hemodynamic responses in the more posterior one of the two FG regions considered (intersection with BA 19) displayed proportional scaling of motion extent and strict additivity of AV interactions, pointing at an early stage of visual processing.

Taken together, these differential effects of visual motion size, namely, visual /ta/ versus /pa/, indicate visual speech information to be processed by two different mechanisms: (i) a more or less direct access to the auditory system via the anterior "what" path (aSTG) toward PAC and (ii) a right hemisphere frontal loop which appears to subserve visual working memory, structured in terms of visual phonological features. Presumably, the latter pathway precedes AV fusion into a common percept, for example, during speech communication in an acoustically disturbed environment when missing acoustic information has to be restored. Such cross-modal effects enhancing phonemic restoration have been demonstrated for congruent AV speech signals when parts of the acoustic component were replaced by noise (Shahin & Miller, 2009).

Left Inferior Frontal Hemisphere Speech Processing Network

At variance with other studies (Hasson et al., 2007; Skipper et al., 2007), the present experiment found visual–phonological effects to be associated with hemodynamic responses of right-hemisphere inferior frontal regions rather than the left-hemisphere speech generation network. Furthermore, acoustic-only signals yielded even more pronounced hemodynamic activation of the left IFG than AV stimuli (see Figure 7). First, these findings might be related to the particular stimulus configuration of the present study, that is, the processing of unresolved phonetic ambiguity of the acoustic stimulus. Because, second, subjects had been instructed to attend to the stimulus-final pitch changes rather than the phonetic content of the stimuli, the absence of any relevant visual effects at the level of the left IFG might reflect the preattentive setting of the experiment in this respect. Thus, the brain responses of the present study seem to be restricted to early and highly automatic speech processing stages. In contrast to sensory memory, verbal working memory operations must not be expected under these conditions.

Conclusions

The results of the present study—concomitant with our preceding whole-head MEG investigations—point at a rather multifaceted scenario of AV interactions during

speech perception, encompassing both early as well as late processing stages:

(i) Even in the absence of an acoustic signal component, visual speech information has access both to PAC and aSTG. Furthermore, visual display of /pa/ and /ta/ utterances were found to elicit differential activation patterns at the level of PAC in at least some (male) subjects. These effects appear to be inhibited or masked in the presence of an actual acoustic signal—as indicated by strong subadditive AV interactions.

(ii) FG and IFG of the right hemisphere displayed enhanced hemodynamic responses to visual /pa/ and suppressed reactions in case of /ta/. Most presumably, these areas participate in the transient storage of phonetically encoded visual speech information (visual working memory). Although not explicitly tested, this right-hemisphere network might be involved in phonological AV fusion processes and, thus, should contribute to the McGurk effect as well as visually induced restoration of meaningful speech, for example, if the acoustic signal is ambiguous or disturbed.

(iii) A third—and more basic—AV interaction seems to be bound to supratemporal regions caudal to PAC and appears to support (speech-unspecific) temporal coordination of the auditory and the visual channel under these conditions (i.e., visual motion acts as a predictor of the forthcoming acoustic signal). This mechanism can be expected to support functional connectivity between modality-specific brain regions.

Acknowledgments

This study was supported by the German Research Foundation (DFG; SFB 550/B1). We thank Maike Borutta for excellent technical assistance.

Reprint requests should be sent to Ingo Hertrich, Department of General Neurology, University of Tübingen, Hoppe-Seyler-Str. 3, D-72076 Tübingen, Germany, or via e-mail: ingo.hertrich@uni-tuebingen.de.

REFERENCES

Altmann, C. F., Bledowski, C., Wibral, M., & Kaiser, J. (2007). Processing of location and pattern changes of natural sounds in the human auditory cortex. *Neuroimage, 35,* 1192–1200.

Avery, P., & Rice, K. (1989). Segment structure and coronal underspecification. *Phonology, 6,* 179–200.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4,* 829–839.

Bendor, D., & Wang, X. (2006). Cortical representations of pitch in monkeys and humans. *Current Opinion in Neurobiology, 16,* 391–399.

Bernstein, L., Auer, E. T., & Moore, J. K. (2004). Audiovisual speech binding: Convergence or association? In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 203–223). Cambridge, MA: MIT Press.

Bernstein, L. E., Auer, E. T., Moore, J. K., Ponton, C. W., Don, M., & Singh, M. (2002). Visual speech perception without primary auditory cortex activation. *NeuroReport, 13,* 311–315.

Blonder, L. X., Smith, C. D., Davis, C. E., West, M. L., Garrity, T. F., Avison, M. J., et al. (2004). Regional brain response to faces of humans and dogs. *Cognitive Brain Research, 20,* 384–394.

Brosch, M., Selezneva, E., & Scheich, H. (2005). Nonauditory events of a behavioral procedure activate auditory cortex of highly trained monkeys. *Journal of Neuroscience, 25,* 6797–6806.

Bunzeck, N., Wuestenberg, T., Lutz, K., Heinze, H. J., & Jäncke, L. (2005). Scanning silence: Mental imagery of complex sounds. *Neuroimage, 26,* 1119–1127.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science, 276,* 593–596.

Calvert, G. A., Hansen, P. C., Iversen, S. D., & Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage, 14,* 427–438.

Calvert, G. A., & Lewis, J. W. (2004). Hemodynamic studies of audiovisual interactions. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 483–502). Cambridge, MA: MIT Press.

Cao, F., Bitan, T., & Booth, J. R. (2008). Effective brain connectivity in children with reading difficulties during phonological processing. *Brain and Language, 107,* 91–101.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology, 5,* e1000436.

Courtney, S. M., Ungerleider, L. G., Keil, K., & Haxby, J. V. (1996). Object and spatial visual working memory activate separate neural systems in human cortex. *Cerebral Cortex, 6,* 39–49.

De Gennaro, L., Bertini, M., Pauri, F., Cristiani, R., Curcio, G., Ferrara, M., et al. (2004). Callosal effects of transcranial magnetic stimulation (TMS): The influence of gender and stimulus parameters. *Neuroscience Research, 48,* 129–137.

De Lacy, P. (2006). Markedness: Reduction and preservation in phonology. *Cambridge studies in linguistics* (Vol. 112). Cambridge, UK: Cambridge University Press.

Di Salle, F., Esposito, F., Scarabino, T., Formisano, E., Marciano, E., Saulino, C., et al. (2003). fMRI of the auditory system: Understanding the neural basis of auditory gestalt. *Magnetic Resonance Imaging, 21,* 1213–1224.

Dien, J. (2009). A tale of two recognition systems: Implications of the fusiform face area and the visual word form area for lateralized object recognition models. *Neuropsychologia, 47,* 1–16.

Diesch, E. (1995). Left and right hemifield advantages of fusions and combinations in audiovisual speech perception. *Quarterly Journal of Experimental Psychology A, 48,* 320–333.

Dietz, N. A. E., Jones, K. M., Gareau, L., Zeffiro, T. A., & Eden, G. F. (2005). Phonological decoding involves left posterior fusiform gyrus. *Human Brain Mapping, 26,* 81–93.

Foxe, J. J., & Schroeder, C. E. (2005). The case for feedforward multisensory convergence during early cortical processing. *NeuroReport, 16,* 419–423.

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience, 25,* 5004–5012.

Griffiths, T. D., & Warren, J. D. (2002). The planum temporale as a computational hub. *Trends in Neurosciences, 25,* 348–353.

Guo, K., Meints, K., Hall, C., Hall, S., & Mills, D. (2009). Left gaze bias in humans, rhesus monkeys and domestic dogs. *Animal Cognition, 12,* 409–418.

Harris, J., & Lindsey, G. (1995). The elements of phonological representation. In J. Durand & F. Katamba (Eds.), *Frontiers of phonology: Atoms, structures, derivations* (pp. 34–79). Harlow, Essex: Longman.
Hasson, U., Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2007). Abstract coding of audiovisual speech: Beyond sensory representation. *Neuron, 56,* 1116–1126.
Hertrich, I., & Ackermann, H. (1999). A vowel synthesizer based on formant sinusoids modulated by fundamental frequency. *Journal of the Acoustical Society of America, 106,* 2988–2990.
Hertrich, I., & Ackermann, H. (2007). Modelling voiceless speech segments by means of an addditive procedure based on the computation of formant sinusoids. In P. Wagner, J. Abresch, S. Breuer, & W. Hess (Eds.), *Proceedings of the 6th ISCA Workshop on Speech Synthesis* (pp. 178–181). Bonn: University of Bonn.
Hertrich, I., Dietrich, S., Moos, A., Trouvain, J., & Ackermann, H. (2009). Enhanced speech perception capabilities in a blind listener are associated with activation of fusiform gyrus and primary visual cortex. *Neurocase, 15,* 163–170.
Hertrich, I., Mathiak, K., Lutzenberger, W., & Ackermann, H. (2002). Hemispheric lateralization of the processing of consonant–vowel syllables (formant transitions): Effects of stimulus characteristics and attentional demands on evoked magnetic fields. *Neuropsychologia, 40,* 1902–1917.
Hertrich, I., Mathiak, K., Lutzenberger, W., & Ackermann, H. (2009). Time course of early audiovisual interactions during speech and non-speech central-auditory processing: An MEG study. *Journal of Cognitive Neuroscience, 21,* 259–274.
Hertrich, I., Mathiak, K., Lutzenberger, W., Menning, H., & Ackermann, H. (2007). Sequential audiovisual interactions during speech perception: A whole-head MEG study. *Neuropsychologia, 45,* 1342–1354.
Hong, K. S., Lee, S. K., Kim, J. Y., Kim, K. K., & Nam, H. (2000). Visual working memory revealed by repetitive transcranial magnetic stimulation. *Journal of the Neurological Sciences, 181,* 50–55.
Jäncke, L., & Shah, N. J. (2004). Hearing syllables by seeing visual stimuli. *European Journal of Neuroscience, 19,* 2603–2608.
Johnson, J. A., & Zatorre, R. J. (2006). Neural substrates for dividing and focusing attention between simultaneous auditory and visual events. *Neuroimage, 31,* 1673–1681.
Jordan, T. R., & Thomas, S. M. (2007). Hemiface contributions to hemispheric dominance in visual speech perception. *Neuropsychology, 21,* 721–731.
Kayser, C., Petkov, C. I., Augath, M., & Logothetis, N. K. (2007). Functional imaging reveals visual modulation of specific fields in auditory cortex. *Journal of Neuroscience, 27,* 1824–1835.
Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerebral Cortex, 18,* 1560–1574.
Kraemer, D. J. M., Macrae, C. N., Green, A. E., & Kelley, W. M. (2005). Musical imagery: Sound of silence activates auditory cortex. *Nature, 434,* 158.
Lahiri, A., & Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition, 38,* 245–294.
Lebib, R., Papo, D., De Bode, S., & Baudonniere, P. M. (2003). Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the human P50 event-related brain potential modulation. *Neuroscience Letters, 341,* 185–188.
Lehmann, C., Herdener, M., Esposito, F., Hubl, D., Di Salle, F., Scheffler, K., et al. (2006). Differential patterns of multisensory interactions in core and belt areas of human auditory cortex. *Neuroimage, 31,* 294–300.
MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics, 24,* 253–257.
McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences, 7,* 293–299.
Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory–visual interactions during early sensory processing in humans: A high-density electrical mapping study. *Cognitive Brain Research, 14,* 115–128.
Möttönen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cognitive Brain Research, 13,* 417–425.
Möttönen, R., Schurmann, M., & Sams, M. (2004). Time course of multisensory interactions during audiovisual speech perception in humans: A magnetoencephalographic study. *Neuroscience Letters, 363,* 112–115.
Nicholls, M. E. R., & Searle, D. A. (2006). Asymmetries for the visual expression and perception of speech. *Brain and Language, 97,* 322–331.
Obleser, J., Zimmermann, J., Van Meter, J., & Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related fMRI. *Cerebral Cortex, 17,* 2251–2257.
Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9,* 97–113.
Pekkola, J., Ojanen, V., Autti, T., Jaaskelainen, I. P., Möttönen, R., & Sams, M. (2006). Attention to visual speech gestures enhances hemodynamic activity in the left planum temporale. *Human Brain Mapping, 27,* 471–477.
Pekkola, J., Ojanen, V., Autti, T., Jaaskelainen, I. P., Mottonen, R., Tarkiainen, A., et al. (2005). Primary auditory cortex activation by visual speech: An fMRI study at 3 T. *NeuroReport, 16,* 125–128.
Pujol, J., Deus, J., Losilla, J. M., & Capdevila, A. (1999). Cerebral lateralization of language in normal left-handed people studied by functional MRI. *Neurology, 52,* 1038–1043.
Rauschecker, J. P. (2001). Cortical plasticity and music. *Annals of the New York Academy of Sciences, 930,* 330–336.
Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences, U.S.A., 97,* 11800–11806.
Rockel, A. J., Hiorns, R. W., & Powell, T. P. S. (1980). The basic uniformity in structure of the neocortex. *Brain, 103,* 221–244.
Ruytjens, L., Albers, F., van Dijk, P., Wit, H., & Willemsen, A. (2007). Activation in primary auditory cortex during silent lipreading is determined by sex. *Audiology and Neuro-Otology, 12,* 371–377.
Sakurai, Y., Takeuchi, S., Takada, T., Horiuchi, E., Nakase, H., & Sakuta, M. (2000). Alexia caused by a fusiform or posterior inferior temporal lesion. *Journal of the Neurological Sciences, 178,* 42–51.
Schirmer, A., Kotz, S. A., & Friederici, A. D. (2005). On the role of attention for the processing of emotions in speech: Sex differences revisited. *Cognitive Brain Research, 24,* 442–452.
Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences, 12,* 106–113.
Schroeder, C. E., Smiley, J., Fu, K. G., McGinnis, T., O'Connell, M. N., & Hackett, T. A. (2003). Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. *International Journal of Psychophysiology, 50,* 5–17.

Schwartz, J. L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception. In R. Campbell, B. Dodd, & D. Burnam (Eds.), *Hearing by eye II* (pp. 85–108). Hove, UK: Psychology Press.

Scott, S. K. (2005). Auditory processing—Speech, space and auditory objects. *Current Opinion in Neurobiology, 15,* 197–201.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123,* 2400–2406.

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neuroscience Research, 47,* 277–287.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America, 90,* 1797–1805.

Shahin, A. J., & Miller, L. M. (2009). Multisensory integration enhances phonemic restoration. *Journal of the Acoustical Society of America, 125,* 1744–1750.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex, 17,* 2387–2399.

Sommer, I. E. C., Diederen, K. M. J., Blom, J. D., Willems, A., Kushan, L., Slotema, K., et al. (2008). Auditory verbal hallucinations predominantly activate the right inferior frontal area. *Brain, 131,* 3169–3177.

Specht, K., & Reul, J. (2003). Functional segregation of the temporal lobes into highly differentiated subsystems for auditory perception: An auditory rapid event-related fMRI-task. *Neuroimage, 20,* 1944–1954.

Stevenson, R., Geoghegan, M., & James, T. (2007). Superadditive BOLD activation in superior temporal sulcus with threshold non-speech objects. *Experimental Brain Research, 179,* 85–95.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215.

Sumiyoshi, C., Matsuo, K., Nakai, T., & Kato, N. (2003). Basic brain activities for phonologically ambiguous syllables: An fMRI study using Japanese speakers. *Proceedings of the International Society for Magnetic Resonance in Medicine, 11,* 565.

Tian, B., Reser, D., Durham, A., Kustov, A., & Rauschecker, J. P. (2001). Functional specialization in rhesus monkey auditory cortex. *Science, 292,* 290–293.

Traunmüller, H., & Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics, 35,* 244–258.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage, 15,* 273–289.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences, U.S.A., 102,* 1181–1186.