Physical and Perceptual Factors Shape the Neural Mechanisms That Integrate Audiovisual Signals in Speech Comprehension

HweeLing Lee and Uta Noppeney

Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany

Face-to-face communication challenges the human brain to integrate information from auditory and visual senses with linguistic representations. Yet the role of bottom-up physical (spectrotemporal structure) input and top-down linguistic constraints in shaping the neural mechanisms specialized for integrating audiovisual speech signals are currently unknown. Participants were presented with speech and sinewave speech analogs in visual, auditory, and audiovisual modalities. Before the fMRI study, they were trained to perceive physically identical sinewave speech analogs as speech (SWS-S) or nonspeech (SWS-N). Comparing audiovisual integration (interactions) of speech, SWS-S, and SWS-N revealed a posterior–anterior processing gradient within the left superior temporal sulcus/gyrus (STS/STG): Bilateral posterior STS/STG integrated audiovisual inputs regardless of spectrotemporal structure or speech percept; in left mid-STS, the integration profile was primarily determined by the spectrotemporal structure of the signals; more anterior STS regions discarded spectrotemporal structure and integrated audiovisual signals constrained by stimulus intelligibility and the availability of linguistic representations. In addition to this “ventral” processing stream, a “dorsal” circuitry encompassing posterior STS/STG and left inferior frontal gyrus differentially integrated audiovisual speech and SWS signals. Indeed, dynamic causal modeling and Bayesian model comparison provided strong evidence for a parallel processing structure encompassing a ventral and a dorsal stream with speech intelligibility training enhancing the connectivity between posterior and anterior STS/STG. In conclusion, audiovisual speech comprehension emerges in an interactive process with the integration of auditory and visual signals being progressively constrained by stimulus intelligibility along the STS and spectrotemporal structure in a dorsal fronto-temporal circuitry.

Introduction

In natural face-to-face communication, the speaker’s intent is conveyed through both auditory (voice) and visual (facial movement) cues. To provide the most likely interpretation of the complex time-varying audiovisual (AV) signals, the human brain is challenged to integrate AV speech signals with higher order linguistic (e.g., phonological, semantic, syntactic) representations. Given the dominant links between speech perception and production, a role in speech perception has also been invoked for articulatory-gestural representations (Liberman and Mattingly, 1985; Wilson et al., 2004; Skipper et al., 2007). Speech processing may thus rely on multiple parallel pathways, most prominently a “what” stream along the anterior temporal lobe transforming the sensory inputs into semantic representations and a dorsal fronto-temporal “how” circuitry interfacing between AV and motor rep-

resentations (Scott and Johnsrude, 2003; Rauschecker and Scott, 2009).

A fundamental question is whether integration of auditory and visual speech signals is “special” or governed by generic principles (Massaro et al., 1996; Stekelenburg and Vroomen, 2007). AV integration in speech perception may be considered special in terms of (1) the specific complex spectrotemporal structure of speech and (2) the linguistic or articulatory representations that provide top-down constraints on the integration and interpretation of the sensory signals (Davis and Johnsrude, 2007). Indeed, the classical McGurk illusion depends crucially on the availability of phonological representations (Tuomainen et al., 2005).

At the neural level, AV integration emerges in a widespread system encompassing subcortical, primary sensory, and association areas (Schroeder and Foxe, 2005; Driver and Noesselt, 2008). This multitude of multisensory integration sites raises the question whether different stimulus properties may be integrated at distinct levels of the cortical hierarchy. Specifically, cortical systems may have evolved that are specialized for integrating AV speech inputs. Previous studies have shown that AV speech and nonspeech signals were integrated in different regions within the left posterior (Stevenson and James, 2009) and anterior (Hertrich et al., 2011) superior temporal sulcus/gyrus (STS/STG). Yet, since these initial studies compared speech to action sequences or formant sweeps that differ from speech along physical (e.g., spectrotemporal structure and complexity) and linguistic dimensions

Received Dec. 14, 2010; revised June 8, 2011; accepted June 15, 2011.

Author contributions: H.L. and U.N. designed research; H.L. performed research; H.L. and U.N. analyzed data; H.L. and U.N. wrote the paper.

This work was supported by the Max Planck Society. We thank Joost Maier, Karin Pilz, Pammi Chandrasekhar, and Johannes Tuennerhoff for help with data acquisition and/or stimuli creation, and Ruth Adam for comments on a previous version of this manuscript.

Correspondence should be addressed to HweeLing Lee, Max Planck Institute for Biological Cybernetics, Speemannstrasse 41, 72076 Tübingen, Germany. E-mail: hwee-ling.lee@tuebingen.mpg.de.

DOI:10.1523/JNEUROSCI.6510-10.2011

Copyright © 2011 the authors 0270-6474/11/3111338-13\$15.00/0

that are linked with an intelligible speech percept (e.g., semantics, syntax, phonology), they were not able to identify the determinants of putative speech-selective AV integration.

To dissociate the contributions of physical and perceptual factors to the integration of audiovisual speech signals, we presented participants with speech and sinewave speech analogs (SWS). Before the fMRI study, we trained one group of participants to perceive SWS as speech, and the other group to perceive SWS as nonspeech. We dissociated three processing levels in audiovisual speech comprehension. First, we identified AV integration common to all stimulus classes regardless of their physical or perceptual properties. Second, we determined where AV integration depended on the spectrotemporal structure of the sensory signals by comparing speech and SWS. Third, we identified AV integration processes that were molded by representational learning and reflected participants' ability to integrate sinewave inputs into intelligible speech percepts. Our study demonstrates that both spectrotemporal complexity and the availability of intelligible speech percepts determine the neural processes integrating AV speech signals.

Materials and Methods

Subjects

Thirty-one healthy right-handed German native speakers (16 females; mean age, 24.7 years) gave informed consent to participate in the study. The study was approved by the human research ethics committee of the medical faculty at the University of Tübingen. All volunteers participated in both psychophysics and functional imaging experiments. One participant was excluded from the study as a result of <10% performance accuracy in the target detection task during the fMRI study.

Stimuli

Stimuli included AV, auditory (A), and visual (V) signals of speech and SWS (Remez et al., 1983; Tuomainen et al., 2005; Benson et al., 2006; Möttönen et al., 2006; Desai et al., 2008). Stimulus material was taken from close-up video recordings of a female face looking straight into the camera, uttering short sentences. Audio and video were recorded with a digital camera (DCR-TRV900E; Sony Corporation; video at 30 frames per second, 720×480 pixels; audio at 44 kHz sampling rate, 16 bit resolution). Sentences were four-word-long neutral statements in German. The AV movies of speech were first cropped to one single complete sentence, preceded and followed by 15 frames of neutral facial expression during which no sound was presented (Adobe Premiere Pro) (for further information on the stimuli, see Maier et al., 2011).

To transform auditory speech into sinewave speech analogs, the audio tracks were separated from the video tracks. The auditory speech was transformed into sinewave speech by replacing the first three formants with sine wave analogs (www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/SWS). The auditory SWS tracks were recombined with the video tracks to create AV SWS movies.

Training to manipulate SWS intelligibility (before the fMRI study)

The intelligibility of SWS stimuli was manipulated between participants by exposing the SWS-S (intelligible SWS speech percept) and the SWS-N (SWS nonspeech percept) groups to different types of speech training ~1 d before the fMRI study. The SWS-S group was presented with SWS sentences preceded by the corresponding speech sentence to facilitate understanding of SWS as speech. The SWS-N group was exposed only to the SWS sentences. We manipulated the speech percept for SWS stimuli between rather than within participants, because an initial pilot study showed that participants quickly generalize from a few intelligible SWS stimuli to the entire SWS stimulus set rendering a within-subject manipulation of the speech percept impossible.

More specifically, we applied the following training procedure to control for effects of spectrotemporal structure and stimulus exposure. First, 20 sentence stimuli were divided in two stimulus sets A and B that were

assigned either to speech or SWS conditions. The assignment was counterbalanced across participants. In the training session, each SWS and speech sentence was presented five times in the AV condition only. The SWS stimuli were preceded by the corresponding AV speech sentence in the SWS-S, but not the SWS-N groups. On the sixth presentation, participants typed the words they recognized from the speech or SWS sentences that were successively presented in each of the three modalities (V, A, and AV). Based on their performance, four SWS and four speech stimuli were selected for each subject. In the SWS-S group, the SWS stimuli were selected if all words were correctly recognized in the auditory and audiovisual modalities. In the SWS-N group, SWS stimuli were selected if none of the words was correctly recognized in any of the three modalities. The speech stimuli were selected such that approximately identical stimuli were presented as speech and SWS across participants to control for stimulus effects (e.g., length, number of syllables). Participants were then further trained on this restricted set of stimuli (in the same fashion as described above). After every fifth presentation, their word recognition performance was evaluated for speech and SWS sentences in each of the three modalities. The training for the SWS-S group was terminated when the auditory and audiovisual SWS stimuli were correctly recognized. To control for exposure effects across the SWS-S and SWS-N groups, the training for the SWS-N group was terminated to match the number of presentations and exposure duration applied in the participants in the SWS-S group. The training session was completed after ~1 h.

This rotation of stimuli over speech and SWS conditions controlled optimally for stimulus effects as measured by number of syllables and length. For the SWS-S group, the average stimulus length was 2.6 s (SD, 0.14 s) for speech stimuli and 2.6 s (SD, 0.13 s) for SWS stimuli. The average number of syllables was 5.4 (SD, 0.15) for speech stimuli and 5.4 (SD, 0.35) for SWS stimuli. For the SWS-N group, the average stimulus length was 2.6 s (SD, 0.14 s) for speech stimuli and 2.5 s (SD, 0.08 s) for SWS stimuli. The average number of syllables was 5.4 (SD, 0.13) for speech stimuli and 5.5 (SD, 0.28) for SWS stimuli. Repeated measures of ANOVAs with stimulus class (speech vs SWS) as within-subject variable and group (SWS-S vs SWS-N) as between-subject variable were performed separately for stimulus length and number of syllables. Neither of the two ANOVAs revealed significant main effects of group or stimulus or a significant interaction between stimulus and group, indicating that our rotation procedure successfully controlled for stimulus confounds.

Since the rotation of stimuli did not fully ensure that identical stimuli were presented for the SWS-S and SWS-N groups, both the behavioral and fMRI analyses were also repeated for a subset of trials in which the stimuli for the SWS-S and SWS-N groups were identical. In this way, we could confirm that effects of speech intelligibility were not confounded by changes in spectrotemporal structure.

Psychophysics experiments to evaluate participants' SWS comprehension (before and after fMRI study)

On the day of the fMRI study, the intelligibility of speech and SWS stimuli in A, V, and AV conditions were evaluated three times: (1) Outside scanner (before and after the fMRI study), participants from the SWS-S and SWS-N groups were presented with the four SWS and the four speech stimuli in each of the three modalities (A, V, and AV). After each stimulus presentation, they typed the words they recognized. (2) To control for effects of scanner noise on stimulus intelligibility, participants were presented with all speech and SWS stimuli in all three modalities (A, V, and AV) while being scanned (but before the actual fMRI experiment). They indicated the number of words they recognized by a five-choice button press (max, 4; min, 0) (self-report).

For each evaluation, performance accuracy was computed as the number of words correctly recognized [typed out correctly (objective) outside the scanner or indicated via button response (subjective) inside the scanner] divided by the total number of words of the sentences.

Experimental design (fMRI study)

During the fMRI study, participants listened and viewed A, V, and AV movies of speech and SWS. The $2 \times 2 \times 2$ ($\times 2$) factorial design manipulated the following: (1) visual input (present, absent), (2) auditory

input (present, absent) and (3) stimulus class (speech, SWS) as within-participants factors, and (4) stimulus percept (SWS-S group perceived SWS as speech, SWS-N group perceived SWS as nonspeech) as a between-participants factor (Fig. 1). The between-subject factor (i.e., whether participants perceived SWS as speech or nonspeech) was selectively manipulated by a training session before the fMRI study. The training session ensured that the sentences were intelligible for the SWS-S group and unintelligible for the SWS-N group while controlling for effects of spectrotemporal structure, number of syllables, stimulus length, stimulus exposure, and familiarity (for further details, see section above). Please note that speech and SWS videos are physically identical and differ only via the AV associative context in which they were presented during the training and the fMRI study; in the AV-SWS conditions, the SWS video is always paired with an auditory SWS stimulus.

For each subject, four SWS and four speech stimuli were selected based on his/her word recognition scores (from the training session before fMRI). Stimulus intelligibility was evaluated before (with and without scanner noise) and after the fMRI experiment. During the fMRI study, each of these eight sentences was presented 48 times in each modality (V, A, AV), amounting to 192 presentations. The stimuli had an average duration of 2.6 s (SD, 0.07 s) and were presented with a fixed inter-trial interval of 0.5 s. The speech and SWS stimuli were presented in periods of 12 stimuli interspersed with fixation periods of 4.8, 9.6, and 14.4 s. The stimulus modality was randomized in an event-related fashion, and the stimulus class was manipulated across blocks.

To maintain participants' attention to both auditory and visual modalities and avoid task-modulatory effects on speech perception (Hickok and Poeppel, 2000), participants responded to simple visual (circle), auditory (beep), and audiovisual (circle plus beep) targets that were presented randomly interspersed during each session. Approximately 18.75% of the trials were targets. Targets were presented for 300 ms.

fMRI data acquisition

A 3T Siemens TRIO TIM MRI scanner (Siemens Medical) was used to acquire both T1 structural volume images (TR/TE/TI, 2300/9.38/1100 ms; 176 slices; matrix, 256 × 240; spatial resolution, 1 × 1 × 1 mm³ voxels) and T2*-weighted axial echo-planar images with blood oxygenation level-dependent (BOLD) contrast (gradient-echo; TR/TE, 3200/40 ms; 40 axial slices; acquired in ascending direction; matrix, 64 × 64; slice thickness, 2.5 mm; interslice gap, 0.5 mm; spatial resolution, 3 × 3 × 3 mm³ voxels). There were six sessions with a total of 135 volume images per session. The first three volumes were discarded to allow for T1 equilibration effects.

fMRI analysis: conventional SPM analysis

The data were analyzed with statistical parametric mapping (SPM5; Wellcome Center for Neuroimaging, London, UK; <http://www.fil.ion.ucl.ac.uk/spm>). Scans from each participant were realigned using the first as a reference. The EPI images were unwrapped, spatially normalized into MNI standard space using parameters from segmentation of the T1 structural image (Ashburner and Friston, 2005), resampled to 2 × 2 × 2 mm³ voxels, and spatially smoothed with a Gaussian kernel of 8 mm FWHM. The time series in each voxel was high-pass filtered to 1/128 Hz. The fMRI experiment was modeled in an event-related fashion with regressors entered into the design matrix after convolving each event-related unit impulse (indexing sentence onset) with a canonical hemodynamic response function and its first temporal derivative. In addition

Figure 1. The experimental paradigm manipulated: visual input (present, absent), auditory input (present, absent), and stimulus class (speech, SWS) as within-subject factors. The intelligibility of SWS were manipulated across participants by training the SWS-S group to perceive the SWS stimuli as speech and the SWS-N group to perceive the SWS stimuli as nonspeech.

to modeling the six conditions in our experiment (i.e., A, V, AV for speech and SWS stimuli), the statistical model included two regressors for each target type (i.e., V, A, and AV targets). Realignment parameters were included as nuisance covariates to account for residual motion artifacts. Condition-specific effects for each subject were estimated according to the general linear model and passed to a second-level analysis as contrasts. This involved creating six contrast images (i.e., each of the six conditions relative to fixation summed over the six sessions) for each subject and entering them into a second-level ANOVA. The ANOVA modeled the 12 conditions (i.e., 6 conditions for the SWS-S and SWS-N group each).

Inferences were made at the second level to allow a random-effects analysis and inferences at the population level (Friston et al., 1995). Unless otherwise stated, we report activations at $p < 0.05$ at the cluster level corrected for multiple comparisons within the AV sentence processing system (all stimuli > fixation at $p < 0.05$, whole-brain corrected, and extent threshold, >850 voxels) using an auxiliary (uncorrected) voxel threshold of $p < 0.001$. This auxiliary threshold defines the spatial extent of activated clusters, which forms the basis of our (corrected) inference.

At the random effects level, we identified AV integration processes using an interaction approach that tests for nonlinear response combinations (i.e., superadditive [AV > (A+V)] and subadditive [AV < (A+V)] AV interactions). More specifically, we identified AV interactions that (1) were common to all stimulus classes, (2) differed for speech and SWS (physical effect), and (3) differed for SWS-S and SWS-N (perceptual effect). Significant AV interactions were characterized as multisensory enhancement [i.e., AV > max(A,V)] or multisensory suppression [i.e., AV < max(A,V)] by comparing the AV response to its maximal unisensory (i.e., A or V) response. These additional tests are not statistically independent from the interaction, yet serve in-depth data characterization and dissociation of multiple mechanisms underlying, for example, subadditive AV interactions. At a descriptive level, we also use conjunction approaches by displaying the auditory- and visual-selective activations (and their intersection) pertaining to the relevant contrasts. Combining multiple complementary methodological approaches reduces the interpretational ambiguities that are associated

Table 1. Performance accuracy and reaction times of visual, auditory, and audiovisual targets for the SWS-S and SWS-N groups

	SWS-S group			SWS-N group		
	V	A	AV	V	A	AV
Performance accuracy	98.3 (1.2)	98.3 (1.2)	99.4 (0.6)	99.2 (0.5)	98.9 (0.5)	99.7 (0.3)
Reaction times (s)	0.71 (0.15)	0.69 (0.19)	0.61 (0.15)	0.73 (0.13)	0.69 (0.18)	0.63 (0.16)

Values represent across participants' mean $\pm$ SEM.

A Before fMRI experiment, outside scanner

B Before fMRI experiment, inside scanner

C After fMRI experiment, outside scanner

Figure 2. Bar plots representing percentage word recognized (across participants' mean $\pm$ SEM) in the visual (light gray), auditory (dark gray), and audiovisual conditions (black), separately for speech (left) and SWS-S perceived as speech (middle) and SWS-N perceived as nonspeech (right). **A**, Before fMRI experiment, outside scanner. **B**, Before fMRI experiment, inside scanner (self-report). **C**, After fMRI experiment, outside scanner. The asterisks indicate significance of direct comparisons at $*p < 0.05$ and $**p < 0.001$, corrected for multiple comparisons using the Holm–Bonferroni method (Holm, 1979).

with each approach when applied in isolation [for further methodological discussion, see a recent chapter focusing on potential and limitations of the various analysis approaches for identification of MSI sites (Noppeney, 2011)].

We dissociated the following three levels of AV integration.

AV integration regardless of spectrotemporal structure or intelligible speech percept (common to all stimulus classes). AV integration processes that emerge regardless of the spectrotemporal structure and the availability of an intelligible speech percept were identified by testing for subadditive (or superadditive) AV interactions that are common to all stimulus classes.

AV integration that depends on the physical stimulus properties (different for speech and SWS). AV integration that depends on the specific spec-

trotemporal structure of speech was identified by comparing subadditive (resp. superadditive) AV interactions for speech and SWS. To increase estimation efficiency, we pooled over SWS-S and SWS-N.

AV integration that depends on the intelligible speech percept (different for SWS-S and SWS-N). AV integration that depends on the speech percept was identified by comparing subadditive (or superadditive) AV interactions for SWS-S and SWS-N that are identical in terms of spectrotemporal structure (across participants) but differ in the availability of the intelligible speech percept. This statistical comparison identified AV integration processes where higher-order linguistic representations (e.g., semantic, phonological, syntactic) influence AV integration.

Analysis of hemispheric lateralization of A, V, AV speech processing

To directly assess whether the reported brain activations were lateralized, we tested for contrast-by-hemisphere interactions. For this purpose, the T1 structural image was segmented and normalized into MNI standard space using symmetrical tissue probability maps that were created by averaging the SPM tissue probability maps with their left–right flipped version. The motion-corrected and unwarped fMRI data were normalized into MNI standard space using the parameters from segmentation of the T1 structural image and spatially smoothed with a Gaussian kernel of 8 mm FWHM. The fMRI experiment was modeled and condition-specific effects for each subject were estimated according to the general linear model as described above. Contrast images, together with their left–right flipped version, were created for each subject and entered into a second-level ANOVA. The activation peaks of all AV interaction contrasts (reported above) were examined in the ANOVA for an effect of hemisphere. For complete characterization of the data, we also investigated the lateralization of subadditive interactions separately for speech and SWS stimuli in the SWS-S and the SWS-N groups (i.e., the lateralization of the subadditive interactions that are shown in Fig. 3A).

fMRI analysis: dynamic causal modeling

For each subject, three dynamic causal models (DCMs) were constructed. Each DCM included three regions: (1) left posterior superior temporal sulcus/gyrus (pSTS/STG) ($x = -52$, $y = -46$, $z = +20$) showing AV interactions common to all stimulus classes as the input region for auditory and visual inputs, (2) left inferior frontal gyrus (IFG) ($x = -46$, $y = +18$, $z = +12$) showing an AV interaction that differs for SWS and speech, (3) left anterior mid-STS (aSTS) ($x = -52$, $y = -16$, $z = -6$) exhibiting an AV interaction that depends on an intelligible speech percept. In all three models, visual (V plus AV) and auditory (A plus AV) speech and SWS stimuli entered as extrinsic inputs to pSTS/STG as the input region. The timings of the onsets were individually adjusted for each region to match the specific time of slice acquisition (Kiebel et al., 2007). The three DCMs manipulated the intrinsic and modulatory connectivity structure that connects the three regions (for the candidate DCMs, see Fig. 8A). Model 1 conforms to a serial processing structure with bidirectional connections from pSTS/STG to aSTS, and from aSTS to IFG. Both connections allow for modulatory effects of sensory modality (e.g., visual modality) and stimulus class (e.g., speech). Model 2 conforms to a parallel processing structure with bidirectional connections

Figure 3. *A*, Subadditive AV interactions, separately for speech (top row) and SWS (middle row) in the SWS-S (left) and SWS-N (right) groups are rendered on a flattened surface template brain. Height threshold: $p < 0.001$, uncorrected, and extent threshold, > 100 voxels, and inclusively masked with all stimuli $>$ fixation at $p < 0.001$, uncorrected, for illustration purposes. *B*, Subadditive AV interactions that are enhanced for speech $>$ SWS, separately for the SWS-S and SWS-N groups are rendered on a flattened surface template brain. Height threshold, $p < 0.05$, uncorrected, and extent threshold, > 150 voxels for illustration purposes, and inclusively masked with all stimuli $>$ fixation at $p < 0.001$, uncorrected.

from pSTS/STG to aSTS (ventral stream), and pSTS/STG to IFG (dorsal stream). Again, both connections allow for modulatory effects of sensory modality (e.g., visual modality) and stimulus class (e.g., speech). Model 3 extends model 2 by including additional bidirectional connections between aSTS and IFG. In addition, in all models, we modeled the AV interaction by allowing AV stimuli to change the self-modulatory connections in the three areas jointly (e.g., pSTS/STG) or separately for speech and SWS (e.g., aSTS). It may be desirable to specify DCMs with many more regions both in the left and right hemispheres. However, since the regional effects of speech percept and physical structure were observed only in the left hemisphere, we limited our analysis to models including only left-hemispheric regions. This is because the aim of DCM is to investigate how regionally specific effects emerge from interactions among brain areas, and therefore it is not advised to select brain regions that do not show significant activations (Stephan et al., 2010). Furthermore, given the complexity of our experimental design, we limited our analysis to address primarily one central question in speech perception. We asked whether speech processing follows a serial or parallel processing structure in the left hemisphere. Future studies with a simpler experimental design are needed to extend this left hemispheric processing model to a more complete model integrating both hemispheres. Specifically, in these more extensive DCMs one may then also investigate lateralization effects in terms of connection strength and how speech perception emerges from interactions between the two hemispheres.

The left-hemispheric regions in our DCM were selected using the regional subcluster maxima of the relevant contrasts from our whole-brain random-effects analysis. Region-specific time series (concatenated over the six runs and adjusted for confounds) comprised the first eigenvariate of all voxels within a sphere of 4 mm^3 radius centered on the peak subcluster voxel from each anatomical region as identified by the relevant second level contrast (or directly adjacent to the peak voxel if it was on the subcluster border).

Bayesian model comparison

To determine the most likely of the three DCMs given the observed data from all subjects, we implemented fixed- and random-effects group analyses (Penny et al., 2004, 2010). The fixed-effects group analysis was implemented by taking the product of the subject-specific Bayes factors over subjects (this is equivalent to the exponentiated sum of the log model evidences of each subject-specific DCM) (Kass and Raftery, 1995;

Penny et al., 2004). Since the fixed-effects group analysis can be distorted by outlier subjects, Bayesian model selection was also implemented in a random-effects group analysis using a hierarchical Bayesian model that estimates the parameters of a Dirichlet distribution over the probabilities of all models considered. To characterize our Bayesian model selection results at the random-effects level, we report (1) the expectation of the posterior probability (i.e., the expected likelihood of obtaining the k th model for any randomly selected subject) and (2) the exceedance probability of one model being more likely than any other model tested (Penny et al., 2010). The exceedance probability quantifies our belief about the posterior probability that is itself a random variable. Thus, in contrast to the expected posterior probability, the exceedance probability also depends on the confidence in the posterior probability. Model comparison and statistical analysis of connectivity parameters of the optimal model enabled us to dissociate serial and parallel (i.e., dorsal vs ventral) model structures for speech and SWS processing.

For the optimal model, the subject-specific modulatory, extrinsic, and intrinsic connection strengths were entered into t tests at the group level separately for the SWS-S and SWS-N groups. This allowed us to summarize the consistent findings from the subject-specific DCMs using classical statistics separately for each group.

We then investigated whether SWS-S and SWS-N groups use the neural systems differently by comparing the connectivity parameters between the SWS-S and SWS-N groups. This across-group comparison allowed us to determine how the prior SWS intelligibility training changes coupling among brain regions.

Results

Behavioral results

Target detection task performance

Table 1 displays performance accuracy and reaction times (across participants' mean $\pm$ SEM) for the SWS-S and SWS-N groups in the target detection task. A repeated-measures ANOVA on performance accuracy with modality (V, A, AV) as within-subject factor and group (SWS-S group, SWS-N group) as between-subject factor revealed a significant main effect of modality ($F_{(1.98,55.5)} = 3.49$; $p < 0.05$), no effect of group ($F_{(1,28)} < 1$; n.s.), and no interaction ($F_{(1.98,55.5)} < 1$; n.s.). *Post hoc* paired-samples

Figure 4. AV integration common to all stimulus classes. **A**, Activations pertaining to visual (red) and auditory (blue) stimuli > fixation (**i**), and subadditive AV interactions (yellow) (**ii**) that are common to all stimulus classes are rendered on a template brain. Height threshold, $p < 0.001$, uncorrected, and extent threshold, > 100 voxels, and inclusively masked with all stimuli > fixation at $p < 0.001$, uncorrected, for illustration purposes. **B**, Left, Subadditive AV interactions are displayed on coronal slices of a normalized structural image (averaged across participants). Height threshold, $p < 0.001$, uncorrected, and extent threshold, > 100 voxels, and inclusively masked with all stimuli > fixation at $p < 0.001$, uncorrected, for illustration purposes. Right, Parameter estimates (across participants' mean $\pm$ SEM) for V, A, and AV conditions of speech, SWS-S, and SWS-N at a given coordinate. Z scores pertain to $AV > \max(A, V)$. [†]n.b., Parameter estimates of speech conditions were averaged across SWS-S and SWS-N groups.

t tests indicated that participants were more accurate for AV targets compared with auditory ($t_{(29)} = 2.47$; $p < 0.05$) or visual targets ($t_{(29)} = 2.21$; $p < 0.05$). Likewise, a repeated-measures ANOVA on reaction times with modality (V, A, AV) as within-subject factor and group (SWS-S group, SWS-N group) as between-subject factor also revealed a significant main effect of modality ($F_{(1.65,46.2)} = 35.1$; $p < 0.001$), no effect of group ($F_{(1,28)} < 1$; n.s.), and no interaction ($F_{(1.65,46.2)} < 1$; n.s.). *Post hoc* paired-samples t tests indicated that participants responded to AV targets faster compared with auditory ($t_{(29)} = 6.33$; $p < 0.001$)

or visual targets ($t_{(29)} = 9.68$; $p < 0.001$). Collectively, the results demonstrate that participants equally attended to both visual and auditory modalities and benefited from integrating inputs from multiple modalities.

Evaluation of intelligibility of speech and SWS sentences

Figure 2 displays percentage words correctly recognized (across participants' mean $\pm$ SEM) for speech and SWS stimuli (1) before (outside the scanner), (2) before (inside the scanner; self-report), and (3) after the fMRI experiment (outside the scanner) in V, A, and AV modalities. Significant correlations were obtained in all conditions between participants' subjective self-report inside the scanner and their objective word recognition scores outside the scanner before and after the fMRI study ($p < 0.001$) indicating that their self-report can be used as a reliable estimate of their SWS sentence comprehension during scanning.

A repeated-measures ANOVA on percentage words recognized with test (before outside, after outside), modality (V, A, AV), and stimulus class (speech, SWS) as within-subject factors and group (SWS-S group, SWS-N group) as between-subject factor revealed significant main effects of group ($F_{(1,28)} = 622.2$; $p < 0.001$), stimulus class ($F_{(1,0,28,0)} = 374.9$; $p < 0.001$), and modality ($F_{(1,0,3,28,9)} = 49.7$; $p < 0.001$). Interactions were observed between (1) stimulus class and modality ($F_{(1,0,3,28,9)} = 30.7$; $p < 0.001$) and (2) stimulus class and group ($F_{(1,0,28,0)} = 546.4$; $p < 0.001$) indicating that our training procedure selectively manipulated the intelligibility of SWS stimuli. Only the SWS-S group was able to comprehend SWS stimuli. The absence of a main effect of time ($F_{(1,28)} < 1$; n.s.) demonstrates that sentence intelligibility was consistent across the entire study. None of the other interactions was significant.

Similarly, a repeated-measures ANOVA on percentage words recognized limited to participants' subjective report inside the scanner (i.e., with scanner noise present) with modality (V, A, AV) and stimulus class (speech, SWS) as within-subject factors and group (SWS-S group, SWS-N group) as between-subject factor revealed significant main effects of group ($F_{(1,28)} = 149.0$; $p < 0.001$), stimulus class ($F_{(1,0,28,0)} = 226.1$; $p < 0.001$), and modality ($F_{(1,23,34,4)} = 74.1$; $p < 0.001$). Interactions were observed between (1) stimulus class and modality ($F_{(1,21,33,8)} = 36.6$; $p < 0.001$) and (2) stimulus class and group ($F_{(1,0,28,0)} = 268.1$; $p < 0.001$), indicating that our training procedure selectively manipulated the intelligibility of SWS stimuli even in the presence of scanner noise. In addition, a trend toward a significant three-way interaction between stimulus

Table 2. Summary of subadditive AV interactions that are (1) common to all stimulus classes, depend on (2) the physical stimulus properties or (3) the intelligible speech percept

Brain regions	MNI coordinates			Cluster size	<i>p</i> value	<i>Z</i> score (peak)	<i>Z</i> score ^a
	<i>x</i>	<i>y</i>	<i>z</i>				
Conjunction across subadditive AV interactions of all stimulus classes							
(A + V) > AV _{[speech in SWS-S group] ∩ [speech in SWS-N group] ∩ [SWS-S] ∩ [SWS-N]}							
Right posterior superior temporal sulcus/gyrus	52	−40	10	1186	<0.001	5.3	—
	60	−38	6			5.1	—
	44	−36	4			4.7	3.1
	52	−30	2			4.6	3.2
Left inferior frontal sulcus/gyrus	−40	14	22	154	0.03	4.4	—
Subadditive AV interactions that depend on the physical stimulus properties (speech > SWS-S and SWS-N)							
(A + V) > AV _{[speech in SWS-S group] + [speech in SWS-N group] > [SWS-S] + [SWS-N]}							
Left mid superior temporal sulcus/middle temporal gyrus	−52	−38	2	323	0.001	4.3	2.2
Left inferior frontal gyrus	−46	18	12	256	0.004	4.0	2.6
Subadditive AV interactions that depend on the intelligibility of speech percept (SWS-S > SWS-N)							
(A + V) > AV _{[SWS-S] > [SWS-N]}							
Left anterior mid-superior temporal sulcus	−52	−16	−6	153	0.03	4.5	—

^a*Z* scores pertaining to the lateralization of each effect (i.e., the three-way interaction of hemisphere by subadditive interaction contrast); height threshold, *p* < 0.05, uncorrected.

class, modality, and group ($F_{(1.21,33.8)} = 3.62$; $p = 0.06$) was observed. The group by modality interaction was not significant ($F_{(1.23,34.4)} < 1$; n.s.). Hence, these results based on participants' subjective report inside the scanner are again consistent with their objective word recognition performance outside the scanner.

For completeness, the behavioral profile was basically equivalent when constraining the analysis to only those SWS trials that were matched across the SWS-S and SWS-N groups (i.e., main effects and interactions were replicated in this analysis).

fMRI results: conventional SPM analysis

Using an interaction analysis, we identified AV integration processes that (1) are common to all stimulus classes, depend on (2) spectrotemporal structure or (3) on participants' speech percept. We further characterized the response profile in each region according to the magnitude of the AV response relative to the maximal unisensory response (i.e., multisensory enhancement vs suppression). Figure 3 shows AV interactions separately for speech and SWS in the SWS-S and SWS-N groups. For all stimulus classes, we observed only subadditive (but no superadditive) AV interactions.

AV integration regardless of spectrotemporal structure and intelligible speech percept (common to all stimulus classes)

Subadditive AV interactions common to speech, SWS-S, and SWS-N were observed in bilateral posterior STS/STG and left inferior frontal sulcus (IFS)/IFG (BA 44) (Figs. 4*Aii*, 7; Table 2). The left posterior STS/STG was only marginally significant at the statistical threshold of 0.07 using the very stringent and overly conservative conjunction null conjunction analysis (Friston et al., 2005). As shown in the parameter estimate plots, the AV response was suppressed relative to the maximal unisensory visual response in left IFS/IFG but not in bilateral posterior STS/STG (Fig. 4*B*). In line with recent neuroimaging studies (Beauchamp et al., 2004; Wallace et al., 2004; Dahl et al., 2009) showing multisensory integration in the transition zones between unisensory cortical domains, AV interactions in the posterior STS/STG were located in the intersection of auditory and visual activations relative to fixation (Fig. 4*Ai*). Thus, both interaction and conjunction analyses revealed convergent results in the posterior STG.

Since a subset of previous neurophysiological and fMRI studies have also reported AV integration processes in primary sen-

sory cortices and subcortical structures, we also characterized the response properties selectively in those structures that were identified based on anatomical landmarks. Neither of those structures integrated sensory inputs in superadditive or subadditive AV interactions. Instead, the primary auditory cortex was selective for auditory inputs with equal responses to A or AV signals (i.e., no response enhancement or suppression either). This is consistent with our recent study demonstrating BOLD superadditive AV interactions only for stimulus transients but not for sustained stimulation (as by our 3–5 s sentences that were organized in blocks) (Werner and Noppeney, 2011).

Similarly, the superior colliculi showed additive response combinations. Yet, here, the AV response was enhanced relative to the unisensory A or V response. This response profile accords with the neurophysiological findings showing enhancement in superior colliculus (Meredith and Stein, 1983). Since our fMRI acquisition parameters were not optimized for subcortical processes, our discussion focuses on AV integration observed at the cortical level.

For completeness, we did not observe any superadditive AV interactions within the AV sentence processing system. This lack of superadditive responses most likely results from several factors: First, to maximize the perceptual difference between SWS-S and SWS-N stimuli, we did not degrade the speech stimuli. According to the principle of inverse effectiveness, the probability of superadditivity and response enhancement decreases with stimulus reliability and efficacy (Stein and Meredith, 1993; Stein et al., 2009; Stevenson and James, 2009; Werner and Noppeney, 2010a,b). In particular, a recent fMRI study demonstrated that AV interactions in STS for speech are primarily subadditive for high signal-to-noise ratio (Stevenson and James, 2009). Second, we presented continuous speech (blocks of sentences) rather than brief stimulus transients, while superadditive responses at the level of the BOLD response were primarily observed for rapid stimulus transients possibly mediating low-level stimulus salience effects (Werner and Noppeney, 2010a, 2011).

AV integration that depends on the physical stimulus properties (different for speech and SWS)

Subadditive AV interactions that were increased for speech relative to SWS were observed in left mid STS/middle temporal gyrus and the pars opercularis of left IFG (BA 44/45) (Figs. 5*Aii*, 7; Table 2). Interestingly, AV integration selective for speech relative to SWS were again located in the intersection of speech-

Figure 5. AV integration that depends on the physical stimulus properties. **A**, Activations pertaining to visual (red) and auditory (blue) speech > SWS (*i*), and subadditive AV interactions (yellow) (*ii*) that are enhanced for speech > SWS are rendered on a template brain. Height threshold, $p < 0.001$, uncorrected, and extent threshold, >100 voxels, and inclusively masked with all stimuli > fixation at $p < 0.001$, uncorrected, for illustration purposes. **B**, Left, Subadditive AV interactions are displayed on coronal slices of a normalized structural image (averaged across participants). Height threshold, $p < 0.001$, uncorrected, and extent threshold, >100 voxels, and inclusively masked with all stimuli > fixation at $p < 0.001$, uncorrected, for illustration purposes. Right, Parameter estimates (across participants' mean $\pm$ SEM) for V, A, and AV conditions of speech, SWS-S, and SWS-N at a given coordinate. Z scores pertain to $AV > \max(A, V)$. [†]n.b., Parameter estimates of speech conditions were averaged across SWS-S and SWS-N groups.

selective auditory and visual activations in the left mid-STS (Fig. 5*Ai*). In left IFG, the AV response was strongly suppressed relative to the maximal unisensory visual response during the speech condition, but comparable with the maximal unisensory visual response during the SWS conditions (Fig. 5*B*). As shown in Figure 5*Bii*, the gradual transition from a suppressive profile for speech and to a nearly additive profile for SWS-N resulted from two factors: (1) Visual facial movements alone elicited the greatest response if they had previously been paired with auditory

speech. In other words, the visual response is greatest for the natural visual speech. (2) Auditory speech was most potent in suppressing visual evoked responses. Hence, during the natural speech condition, the strong response to a visual facial movement is completely suppressed by a concurrent natural auditory speech stimulus resulting in a suppressive AV interaction for the speech stimulus (Fig. 5*Bii*). In contrast, the smaller visual response to SWS-N paired with facial movements were not suppressed by auditory SWS-N.

For completeness, no subadditive AV interactions that were increased for SWS relative to speech were observed.

AV integration that depends on the intelligible speech percept (different for SWS-S and SWS-N)

Increased subadditive AV interactions for SWS-S relative to SWS-N were observed in left middle STS, anterior to the left mid-STS region where the AV integration profile depended on spectrotemporal structure (Figs. 6*A*, 7; Table 2). This left anterior mid-STS region showed a suppressive AV interaction profile for SWS-S and speech, and a multi-sensory enhancement for SWS-N (Fig. 6*B*). These results suggest that the availability of the intelligible speech percept determines the mode of integration within left anterior mid-STS. As a consequence of our initial stimulus selection procedure, one may argue that the differences may not be caused by changes in speech percept alone, but by spectrotemporal differences between SWS-S and SWS-N. To ensure that changes in AV interaction profile in left anterior mid-STS are indeed due to subject's percept, we performed additional first- and second-level analyses that separately modeled the sentences that were matched (identical) and unmatched for the SWS-S and SWS-N groups. Limiting the statistical comparison only to the matched (i.e., identical stimuli presented equally often in SWS-S and SWS-N groups) basically replicated the effect in left anterior mid-STS at ($x = -52$, $y = -14$, $z = -6$; Z score = 3.2) at a lower level of statistical significance. In other words, when we included only identical sentences in the SWS-N and SWS-S conditions, we equally observed the significant modulation of the AV interaction profile by group. Even more convincingly, the interaction profile again changed from a suppressive profile for SWS-S to an additive profile for SWS-N. This suggests that the lower statistical significance when including only the matched trials results from a reduction in estimation efficiency because of the lower number of trials and the correlations between the matched and unmatched regressors. Collectively, these results demonstrate that AV integration in left anterior mid-STS is

selective auditory and visual activations in the left mid-STS (Fig. 5*Ai*). In left IFG, the AV response was strongly suppressed relative to the maximal unisensory visual response during the speech condition, but comparable with the maximal unisensory visual response during the SWS conditions (Fig. 5*B*). As shown in Figure 5*Bii*, the gradual transition from a suppressive profile for speech and to a nearly additive profile for SWS-N resulted from two factors: (1) Visual facial movements alone elicited the greatest response if they had previously been paired with auditory

indeed modulated by speech intelligibility regardless of spectrotemporal structure.

For completeness, no subadditive AV interactions that were increased for SWS-S relative to SWS-N were observed.

Hemispheric lateralization of fMRI activations

First, we examined the activation peaks of all AV interaction contrasts (reported above) for potential hemisphere by contrast interactions. The Z scores for each hemisphere by contrast interaction are added as an additional column in Table 2. In brief, at an uncorrected level of statistical significance, the conjunction across subadditive AV interactions was right-lateralized in the posterior STS/STG, but left-lateralized in the IFS/IFG. In contrast, the difference in subadditivity for speech versus SWS (i.e., the physical effects on AV integration) was left-lateralized in both frontal and temporal cortices. Finally, the difference in subadditivity for SWS-S versus SWS-N was not significantly lateralized. However, since we did not observe any significant effect of speech intelligibility on AV integration in the right hemisphere, we will still refer to the activation as located in the left anterior mid-STS [nota bene (n.b.), this labeling does not imply any significant lateralization].

Second, we investigated the lateralization of subadditive AV interactions separately for speech and SWS stimuli in the SWS-S and the SWS-N groups (i.e., the lateralization of the subadditive AV interactions that are shown in Fig. 3A) (Table 3). In brief, as indicated in the conjunction analysis, this analysis confirmed that the subadditive AV interactions in the posterior mid-STS were right-lateralized independently for speech stimuli in both groups, SWS-S and SWS-N stimuli. A more posterior STS region showed left-lateralized subadditive AV interactions selectively for speech stimuli (primarily in the SWS-N group). In contrast, speech and SWS stimuli induced left-lateralized activations in non-overlapping subregions of the IFG.

fMRI results: DCM

Figure 8B shows the model comparison results for the three DCMs from the fixed-effects group (left) and the random-effects analysis (right). Both analyses demonstrate that model 3 is the optimal model among the three models tested both in the SWS-N and SWS-S groups. The exceedance probability of model 3 being more likely than any other model tested was 0.88 for SWS-N group and 0.86 for SWS-S group. Similarly, in both groups, the second best model was model 2. Hence our DCM results provide strong evidence for a parallel or dual stream model, where speech and SWS stimuli are processed along both the ventral (pSTS/STG-aSTS) and dorsal (pSTS/STG-IFG) streams. Despite being more complex (plus two additional parameters) than the other two models, model 3 even outperforms model 2 suggesting that dorsal and ventral streams do not process the information inde-

Figure 6. AV integration that depends on the intelligible speech percept. **A**, Activations pertaining to visual (red) and auditory (blue) SWS-S > SWS-N (*i*), and subadditive AV interactions (yellow) (*ii*) that are enhanced for SWS-S > SWS-N are rendered on a template brain. Height threshold, $p < 0.001$, uncorrected, and extent threshold, > 100 voxels, and inclusively masked with all stimuli > fixation at $p < 0.001$, uncorrected, for illustration purposes. **B**, Left, Subadditive AV interactions are displayed on coronal slices of a normalized structural image (averaged across participants). Height threshold, $p < 0.001$, uncorrected, and extent threshold, > 100 voxels, and inclusively masked with all stimuli > fixation at $p < 0.001$, uncorrected, for illustration purposes. Right, Parameter estimates (across participants' mean $\pm$ SEM) for V, A, and AV conditions of speech, SWS-S, and SWS-N at a given coordinate. Z scores pertain to $AV > \max(A,V)$. $\dagger$ n.b., Parameter estimates of speech conditions were averaged across SWS-S and SWS-N groups.

Figure 7. Subadditive AV interactions that are common to all stimulus classes (magenta), depend on the physical stimulus properties (cyan), and the intelligible speech percept (yellow) are rendered on a flattened surface of a template brain. Height threshold, $p < 0.001$, uncorrected, and extent threshold, > 100 voxels, and inclusively masked with all stimuli > fixation at $p < 0.001$, uncorrected, for illustration purposes.

pendently, but in an interactive fashion with both streams converging in IFG in line with several current models of speech processing (Hickok and Poeppel, 2007; Rauschecker and Scott, 2009).

Table 3. Summary of lateralization of subadditive AV interactions for each stimulus class in each group (i.e. three-way interaction of subadditive interaction by hemisphere)

Brain regions	MNI coordinates			Cluster size	<i>p</i> value ^a	Z score (peak)
	<i>x</i>	<i>y</i>	<i>z</i>			
Subadditive AV interaction for speech in SWS-S group						
(A + V) > AV _[speech in SWS-S group]						
Right mid superior temporal sulcus/middle temporal gyrus	44	−28	−2	400	<0.001	5.8
Left posterior superior temporal sulcus/middle temporal gyrus	−54	−56	10	104	0.05	3.9
Left inferior frontal gyrus	−48	34	−2	26	n.s.	4.5
Subadditive AV interaction for speech in SWS-N group						
(A + V) > AV _[speech in SWS-N group]						
Right mid superior temporal sulcus/middle temporal gyrus	44	−34	4	91	0.064	4.0
Left inferior frontal gyrus	−44	24	22	408	<0.001	5.4
	−52	12	6			4.8
Subadditive AV interaction for SWS in SWS-S group						
(A + V) > AV _[SWS-S]						
Right mid superior temporal sulcus/middle temporal gyrus	42	−30	−2	487	<0.001	6.0
Left supramarginal gyrus	−48	−36	26	138	0.02	5.0
Left inferior frontal gyrus	−48	12	14	111	0.04	4.7
Subadditive AV interaction for SWN in SWS-N group						
(A + V) > AV _[SWS-N]						
Right mid superior temporal sulcus/middle temporal gyrus	40	−34	2	818	<0.001	5.9
	40	−38	6			5.8
	40	−30	−4			5.4
	56	−40	2			5.1

^aValue of *p* corrected at cluster level for multiple comparisons within all voxels activated for stimuli (averaged across flipped and unflipped) > fixation (height threshold, *p* < 0.05, whole-brain corrected, and extent threshold, >850 voxels). n.s., Nonsignificant (when corrected at the cluster level).

Figure 8. *A*, The three candidate DCMs manipulating the intrinsic connectivity structure. *i*, Model 1 conforms to a serial processing structure. *ii*, Model 2 conforms to a parallel processing structure encompassing a ventral and a dorsal stream. *iii*, Model 3 extends model 2 by adding bidirectional connections between aSTS and IFG. *B*, Bayesian model comparisons for each of the three DCMs in the SWS-N (*i*) and SWS-S (*ii*) groups using a fixed- and random-effects group analyses. In both groups, model 3 shows the highest log model evidence, expected posterior probability and exceedance probability. *C*, The optimal model 3 for the SWS-N group (*i*) and the SWS-S group (*ii*). The numbers are across participants' mean $\pm$ SEM of intrinsic, extrinsic, and modulatory connection strengths [solid lines indicate significance at *p* < 0.05, corrected for multiple comparisons using the Holm–Bonferroni method (Holm, 1979)].

Figure 8C shows the changes in connection strength for the optimal model 3 in the SWS-S and SWS-N groups. The numbers by the intrinsic connections index the influence that one region exerts over another (i.e., responsiveness of the target region to activity in the source region). The numbers by the modulatory effects index the change in coupling induced by stimulus class (e.g., speech vs SWS) or modality (e.g., auditory vs visual). Furthermore, the self-connections are modulated by AV integration. All numbers reflect coupling strengths or changes in coupling averaged across participants (mean $\pm$ SEM). In line with the predominant role of the temporal cortices in speech recognition, the ventral stream (i.e., the connections between pSTS/STG and aSTS) exceeds in strength the dorsal pathway between pSTS/STG and IFG (SWS-N: $t_{(14)} = 2.3$, $p < 0.05$; SWS-S: $t_{(14)} = 2.8$, $p < 0.05$).

Finally, we compared the DCM connectivity across the SWS-S and SWS-N groups in multivariate ANOVAs separately for the intrinsic and modulatory connections. The intrinsic connectivity showed a marginally significant difference across the two groups ($p = 0.08$). *Post hoc* testing demonstrated that the forward connections from pSTS/STG to aSTS were significantly stronger in the SWS-S group than in the SWS-N group ($t_{(28)} = 2.78$; $p < 0.05$). In contrast, the modulatory connections did not differ across the two groups ($p = 0.24$). These results suggest that SWS-S training increases the coupling between posterior and anterior STS areas leading to enhanced SWS intelligibility.

Discussion

Face-to-face communication challenges the human brain to integrate information from auditory and visual senses with higher-order linguistic representations (van Wassenhove et al., 2005). Despite recent suggestions that regions within the STS may be specialized for AV speech perception (Stevenson and James, 2009; Hertrich et al., 2011), it is unclear whether differences between speech and other stimulus classes result from differences in spectrotemporal structure or higher-order linguistic representations associated with an intelligible speech percept. Our functional imaging and effective connectivity (DCM) results suggest that AV speech comprehension emerges in an interactive process with the integration of auditory and visual signals being progressively constrained by stimulus intelligibility along the STS and spectrotemporal structure and articulatory representations in a fronto-temporal circuitry.

Bilateral posterior STS/STG showed an AV integration profile common to all stimulus classes regardless of their spectrotemporal structure or speech intelligibility (Wright et al., 2003; Beauchamp et al., 2004, 2008; van Atteveldt et al., 2004; Amedi et al., 2005; Barraclough et al., 2005; Beauchamp, 2005; Miller and D'Esposito, 2005; Ghazanfar et al., 2008, 2010; Arnal et al., 2009; Szyk et al., 2009; Werner and Noppeney, 2010b). More specifically, the bilateral posterior STS/STG responded equally to A, V, and AV inputs leading to subadditive AV interactions. This subadditive response profile could be caused by nonlinearities of the BOLD response. Alternatively, it may simply reflect amodal higher-order processes that are commonly engaged for all stimulus classes. In other words, information from different senses may have already converged in other areas and STS performs higher-order processing on the A, V, or integrated AV representations. While our study cannot unambiguously implicate the posterior STS/STG in multisensory integration, the specific anatomical location argues for a "true" multisensory mechanism at the neural level. Thus, the AV integration effect was observed selectively in the transition zones between visual and auditory dominant systems (see intersection in Fig. 4Ai) that may be ide-

ally suited to integrate inputs from multiple senses based on theoretical considerations and recent neuroimaging evidence (Beauchamp et al., 2004; Wallace et al., 2004; Dahl et al., 2009).

In contrast, left IFG and left mid-STG formed a circuitry that integrated AV inputs depending on spectrotemporal and perceptual factors. The left IFG [more specifically, left ventral premotor cortex (BA44)] forms part of the mirror neuron system (Nelissen et al., 2005; Petrides et al., 2005) that responds, when participants make or observe a particular action (Rizzolatti and Craighero, 2004; Fadiga et al., 2005) or perceive sounds produced by that action (Kohler et al., 2002). We demonstrate that left IFG (BA44) primarily responded to visual facial movements rather than the corresponding auditory speech (Skipper et al., 2005). Interestingly, the BOLD response elicited by "visual" facial movements was modulated by prior cross-modal learning (Gonzalo et al., 2000; von Kriegstein and Giraud, 2006). It was maximal for facial movements that had been paired with auditory speech during the training and minimal for those paired with unintelligible SWS (Fig. 5). Thus, prior cross-modal exposure to AV speech may induce associations between facial movements and articulatory patterns that amplify the "mirror neuron" response in left IFG to facial movements even when presented alone. This response enhancement is not purely a visual speech intelligibility effect, since speech recognition performance was increased for SWS-S relative to speech. Conversely, auditory speech is the most effective stimulus in suppressing left IFG response in visual speech during AV integration. These two factors [i.e., (1) facial movements alone elicit the greatest response when previously paired with auditory speech, and (2) auditory speech is the most potent stimulus in suppressing visual evoked responses] generate a gradual change in AV integration profile in left IFG (and to some degree in mid-STG) from nearly additive in the SWS-N condition to suppressive AV interactions in the speech condition (Fig. 5B). The suppressive integration profile for AV speech cannot be attributed to a saturation of the BOLD response. Instead, it indexes nonlinearities at the neural level and is consistent with recent neurophysiological reports of suppressive AV interactions in the ventrolateral prefrontal cortex (Sugihara et al., 2006). Hence the AV integration profile (i.e., whether it is suppressive or additive) in this left fronto-temporal circuitry is determined by the spectrotemporal structure of the signals and prior cross-modal training that enables articulatory and higher-order linguistic (e.g., phonological) representations. Recent studies of the McGurk effect have similarly shown that the left inferior frontal cortex codes participants' integrated AV McGurk percept ("da") rather than its visual ("ga") or auditory ("ba") constituents (Hasson et al., 2007; Skipper et al., 2007). Collectively, these results may suggest that the fronto-temporal circuitry integrates auditory and visual inputs into a speech percept guided by prior articulatory-gestural representations. However, since the present study did not selectively manipulate the availability of articulatory movements, this perception-action/production loop perspective is only one tentative interpretation. Alternatively, the left prefrontal cortex may play a more general role in categorically integrating sensory inputs from environmentally relevant stimuli such as speech and matched facial movements onto phonological categories (Werner and Noppeney, 2010a). Finally, given the low temporal resolution of fMRI, we cannot fully exclude the possibility that left IFG retrieves and processes already integrated AV information furnished by posterior STS/STG that exhibits a similar activation profile (Sugihara et al., 2006). Future studies combining M/EEG and DCM may enable us to disentangle these different interpretational perspectives.

In left anterior mid-STS, the AV integration profile depended only on participants' percept (i.e., speech intelligibility regardless of spectrotemporal structure). While both speech and intelligible SWS-S showed subadditive (with a trend toward suppressive) AV interaction profiles, sensory inputs were integrated with an AV enhancement during the SWS-N condition. This invariance of the AV integration profile to physical stimulus features in the anterior portion of left mid-STS converges with previous findings in the auditory domain (Binder et al., 2000; Scott et al., 2000; Davis and Johnsrude, 2003; Davis et al., 2007). In particular, Davis and Johnsrude (2003) demonstrated that anterior temporal areas responded to higher-order linguistic information and intelligible speech stimuli regardless of their specific auditory form. Anterior temporal areas have also been shown to be activated by both written and spoken words indicating a role in amodal lexical and semantic access (Spitsyna et al., 2006). Our study demonstrates that higher-order linguistic information does not only converge but is integrated in these anterior areas as a function of stimulus intelligibility. The change in AV integration profile from response suppression to enhancement follows the principle of inverse effectiveness, whereby less effective inputs are combined with response enhancement and more effective inputs with response suppression (Stein and Meredith, 1993; Stein et al., 2009; Stevenson and James, 2009; Werner and Noppeney, 2010a,b). In line with the neural mechanisms invoked for priming (Dolan et al., 1997; Henson et al., 2000), the response suppression and enhancement in AV integration may serve two distinct aims: If a speech or a SWS stimulus can be easily recognized in at least one (e.g., auditory) modality alone, the suppressive AV interactions may reflect more efficient and faster processing. Conversely, if neither of the unisensory signals can be recognized when presented alone as in the case of SWS-N, AV integration may enable the emergence of intelligible-like speech representations. Even though the word recognition rate as measured inside the scanner increased for AV relative to either unisensory condition only nonsignificantly, it is conceivable that intelligible word fragments may elicit speech-like processing (for a similar argument, see Noesselt et al., 2010; Noppeney et al., 2010; Werner and Noppeney, 2010a). This integration profile extends current theories of auditory speech perception to the audiovisual domain (Scott et al., 2000; Davis and Johnsrude, 2003; Scott and Johnsrude, 2003; Poeppel et al., 2008; Petkov et al., 2009; Rauschecker and Scott, 2009) and suggests that left anterior mid-STS may integrate auditory and visual signals into an intelligible speech percept.

In conclusion, our results suggest that both bottom-up spectrotemporal inputs and the availability of top-down linguistic (and articulatory) representations shape the neural processes underlying AV integration in speech perception (Davis and Johnsrude, 2007; Friston, 2009, 2010). AV speech comprehension emerges in an interactive process with integration of auditory and visual signals being progressively constrained by stimulus intelligibility along the STS and spectrotemporal structure and familiarity within a fronto-temporal circuitry. Indeed, our DCM results provided strong evidence for a parallel processing model encompassing a ventral stream from pSTS/STG to aSTS and a dorsal stream from pSTS/STG to IFG that interactively engage in AV speech integration. Brief prior AV training modified the functional coupling within the network of these three key players and increased the strength of the forward connection from pSTS/STG to aSTS. Furthermore, it dramatically altered the responses to identical visual signals and their AV interaction profile (suppressive vs additive) by which they are being integrated.

How do our results speak to the initially posed question whether audiovisual integration of speech is special? Brain regions and systems may have evolved specialized for (1) integrating complex spectrotemporal signals that map onto phonological representations (mid-STS) or (2) integrating sensory signals into higher-order semantic information that is invariant to the specific surface features of the acoustic or visual signals (anterior mid-STS). However, none of these regions may in itself be specific for speech processing. For instance, the mid-STS region may also be involved in integrating spectrotemporally complex auditory and visual nonspeech signals. Conversely, the anterior mid-STS may well be involved in integrating semantically invested videos with auditory nonspeech tracks. Instead, we propose that AV integration of speech signal emerges from interactions among several brain areas that may each have specialized for a process that is shared by nonspeech signals. In short, AV integration of speech signals may be special in terms of the specific interactions among multiple regions in a fronto-temporal system.

References

- Amedi A, von Kriegstein K, van Atteveldt NM, Beauchamp MS, Naumer MJ (2005) Functional imaging of human crossmodal identification and object recognition. *Exp Brain Res* 166:559–571.
- Arnal LH, Morillon B, Kell CA, Giraud AL (2009) Dual neural routing of visual facilitation in speech processing. *J Neurosci* 29:13445–13453.
- Ashburner J, Friston KJ (2005) Unified segmentation. *Neuroimage* 26:839–851.
- Barracough NE, Xiao D, Baker CI, Oram MW, Perrett DI (2005) Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *J Cogn Neurosci* 17:377–391.
- Beauchamp MS (2005) See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Curr Opin Neurobiol* 15:145–153.
- Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A (2004) Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nat Neurosci* 7:1190–1192.
- Beauchamp MS, Yasar NE, Frye RE, Ro T (2008) Touch, sound and vision in human superior temporal sulcus. *Curr Opin Neurobiol* 18:1011–1020.
- Benson RR, Richardson M, Whalen DH, Lai S (2006) Phonetic processing areas revealed by sinewave speech and acoustically similar non-speech. *Neuroimage* 31:342–353.
- Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET (2000) Human temporal lobe activation by speech and nonspeech sounds. *Cereb Cortex* 10:512–528.
- Dahl CD, Logothetis NK, Kayser C (2009) Spatial organization of multisensory responses in temporal association cortex. *J Neurosci* 29:11924–11932.
- Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. *J Neurosci* 23:3423–3431.
- Davis MH, Johnsrude IS (2007) Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hear Res* 229:132–147.
- Davis MH, Coleman MR, Absalom AR, Rodd JM, Johnsrude IS, Matta BF, Owen AM, Menon DK (2007) Dissociating speech perception and comprehension at reduced levels of awareness. *Proc Natl Acad Sci U S A* 104:16032–16037.
- Desai R, Liebenthal E, Waldron E, Binder JR (2008) Left posterior temporal regions are sensitive to auditory categorization. *J Cogn Neurosci* 20:1174–1188.
- Dolan RJ, Fink GR, Rolls E, Booth M, Holmes A, Frackowiak RS, Friston KJ (1997) How the brain learns to see objects and faces in an impoverished context. *Nature* 389:596–599.
- Driver J, Noesselt T (2008) Multisensory interplay reveals crossmodal influences on “sensory-specific” brain regions, neural responses, and judgments. *Neuron* 57:11–23.
- Fadiga L, Craighero L, Olivier E (2005) Human motor cortex excitability during the perception of others' action. *Curr Opin Neurobiol* 15:213–218.
- Friston K (2009) The free-energy principle: a rough guide to the brain? *Trends Cogn Sci* 13:293–301.
- Friston K (2010) The free-energy principle: a unified brain theory? *Nat Rev Neurosci* 11:127–138.

- Friston KJ, Holmes AP, Worsley KJ, Poline JB, Frith CD, Frackowiak RSJ (1995) Statistical parametric maps in functional imaging: a general linear approach. *Hum Brain Mapp* 2:189–210.
- Friston KJ, Penny WD, Glaser DE (2005) Conjunction revisited. *Neuroimage* 25:661–667.
- Ghazanfar AA, Chandrasekaran C, Logothetis NK (2008) Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. *J Neurosci* 28:4457–4469.
- Ghazanfar AA, Chandrasekaran C, Morrill RJ (2010) Dynamic, rhythmic facial expressions and the superior temporal sulcus of macaque monkeys: implications for the evolution of audiovisual speech. *Eur J Neurosci* 31:1807–1817.
- Gonzalo D, Shallice T, Dolan R (2000) Time-dependent changes in learning audiovisual associations: a single-trial fMRI study. *Neuroimage* 11:243–255.
- Hasson U, Skipper JI, Nusbaum HC, Small SL (2007) Abstract coding of audiovisual speech: beyond sensory representation. *Neuron* 56:1116–1126.
- Henson R, Shallice T, Dolan R (2000) Neuroimaging evidence for dissociable forms of repetition priming. *Science* 287:1269–1272.
- Hertrich I, Dietrich S, Ackermann H (2011) Cross-modal interactions during perception of audiovisual speech and nonspeech signals: an fMRI study. *J Cogn Neurosci* 23:221–237.
- Hickok G, Poeppel D (2000) Towards a functional neuroanatomy of speech perception. *Trends Cogn Sci* 4:131–138.
- Hickok G, Poeppel D (2007) The cortical organization of speech processing. *Nat Rev Neurosci* 8:393–402.
- Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat* 6:65–70.
- Kass RE, Raftery AE (1995) Bayes factors. *J Am Stat Assoc* 90:773–795.
- Kiebel SJ, Kloppel S, Weiskopf N, Friston KJ (2007) Dynamic causal modeling: a generative model of slice timing in fMRI. *Neuroimage* 34:1487–1496.
- Kohler E, Keysers C, Umiltà MA, Fogassi L, Gallese V, Rizzolatti G (2002) Hearing sounds, understanding actions: action representation in mirror neurons. *Science* 297:846–848.
- Lieberman AM, Mattingly IG (1985) The motor theory of speech perception revised. *Cognition* 21:1–36.
- Maier JX, Di Luca M, Noppeney U (2011) Audiovisual asynchrony detection in human speech. *J Exp Psychol Hum Percept Perform* 37:245–256.
- Massaro DW, Cohen MM, Smele PM (1996) Perception of asynchronous and conflicting visual and auditory speech. *J Acoust Soc Am* 100:1777–1786.
- Meredith MA, Stein BE (1983) Interactions among converging sensory inputs in the superior colliculus. *Science* 221:389–391.
- Miller LM, D'Esposito M (2005) Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J Neurosci* 25:5884–5893.
- Möttönen R, Calvert GA, Jääskeläinen IP, Matthews PM, Thesen T, Tuomainen J, Sams M (2006) Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *Neuroimage* 30:563–569.
- Nelissen K, Luppino G, Vanduffel W, Rizzolatti G, Orban GA (2005) Observing others: multiple action representation in the frontal lobe. *Science* 310:332–336.
- Noesselt T, Tyll S, Boehler CN, Budinger E, Heinze HJ, Driver J (2010) Sound-induced enhancement of low-intensity vision: multisensory influences on human sensory-specific cortices and thalamic bodies relate to perceptual enhancement of visual detection sensitivity. *J Neurosci* 30:13609–13623.
- Noppeney U (2011) Characterization of multisensory integration with fMRI—experimental design, statistical analysis and interpretation. In: *Frontiers in the neural bases of multisensory processes* (Wallace MT, Murray MM, eds), pp 233–252, in press. Boca Raton, FL: CRC.
- Noppeney U, Ostwald D, Werner S (2010) Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. *J Neurosci* 30:7434–7446.
- Penny WD, Stephan KE, Mechelli A, Friston KJ (2004) Comparing dynamic causal models. *Neuroimage* 22:1157–1172.
- Penny WD, Stephan KE, Daunizeau J, Rosa MJ, Friston KJ, Schofield TM, Leff AP (2010) Comparing families of dynamic causal models. *PLoS Comput Biol* 6:e1000709.
- Petkov CI, Logothetis NK, Obleser J (2009) Where are the human speech and voice regions, and do other animals have anything like them? *Neuroscientist* 15:419–429.
- Petrides M, Cadoret G, Mackey S (2005) Orofacial somatomotor responses in the macaque monkey homologue of Broca's area. *Nature* 435:1235–1238.
- Poeppel D, Idsardi WJ, van Wassenhove V (2008) Speech perception at the interface of neurobiology and linguistics. *Philos Trans R Soc Lond B Biol Sci* 363:1071–1086.
- Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat Neurosci* 12:718–724.
- Remez RE, Rubin PE, Pisoni DB (1983) Coding of the speech spectrum in three time-varying sinusoids. *Ann N Y Acad Sci* 405:485–489.
- Rizzolatti G, Craighero L (2004) The mirror-neuron system. *Annu Rev Neurosci* 27:169–192.
- Schroeder CE, Foxe J (2005) Multisensory contributions to low-level, “unisensory” processing. *Curr Opin Neurobiol* 15:454–458.
- Scott SK, Johnsrude IS (2003) The neuroanatomical and functional organization of speech perception. *Trends Neurosci* 26:100–107.
- Scott SK, Blank CC, Rosen S, Wise RJ (2000) Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* 123:2400–2406.
- Skipper JI, Nusbaum HC, Small SL (2005) Listening to talking faces: motor cortical activation during speech perception. *Neuroimage* 25:76–89.
- Skipper JI, van Wassenhove V, Nusbaum HC, Small SL (2007) Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb Cortex* 17:2387–2399.
- Spitsyna G, Warren JE, Scott SK, Turkheimer FE, Wise RJ (2006) Converging language streams in the human temporal lobe. *J Neurosci* 26:7328–7336.
- Stein BE, Meredith AM, eds (1993) *The merging of the senses*. Cambridge, MA: MIT.
- Stein BE, Stanford TR, Ramachandran R, Perrault TJ Jr, Rowland BA (2009) Challenges in quantifying multisensory integration: alternative criteria, models, and inverse effectiveness. *Exp Brain Res* 198:113–126.
- Stekelenburg JJ, Vroomen J (2007) Neural correlates of multisensory integration of ecologically valid audiovisual events. *J Cogn Neurosci* 19:1964–1973.
- Stephan KE, Penny WD, Moran RJ, den Ouden HE, Daunizeau J, Friston KJ (2010) Ten simple rules for dynamic causal modeling. *Neuroimage* 49:3099–3109.
- Stevenson RA, James TW (2009) Audiovisual integration in human superior temporal sulcus: inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage* 44:1210–1223.
- Sugihara T, Diltz MD, Averbach BB, Romanski LM (2006) Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. *J Neurosci* 26:11138–11147.
- Szyck GR, Jansma H, Münte TF (2009) Audiovisual integration during speech comprehension: an fMRI study comparing ROI-based and whole brain analyses. *Hum Brain Mapp* 30:1990–1999.
- Tuomainen J, Andersen TS, Tiippana K, Sams M (2005) Audio-visual speech perception is special. *Cognition* 96:B13–B22.
- van Atteveldt N, Formisano E, Goebel R, Blomert L (2004) Integration of letters and speech sounds in the human brain. *Neuron* 43:271–282.
- van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. *Proc Natl Acad Sci U S A* 102:1181–1186.
- von Kriegstein K, Giraud AL (2006) Implicit multisensory associations influence voice recognition. *PLoS Biol* 4:e326.
- Wallace MT, Ramachandran R, Stein BE (2004) A revised view of sensory cortical parcellation. *Proc Natl Acad Sci U S A* 101:2167–2172.
- Werner S, Noppeney U (2010a) Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *J Neurosci* 30:2662–2675.
- Werner S, Noppeney U (2010b) Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cereb Cortex* 20:1829–1842.
- Werner S, Noppeney U (2011) The contributions of transient and sustained response codes to audiovisual integration. *Cereb Cortex* 21:920–931.
- Wilson SM, Saygin AP, Sereno MI, Iacoboni M (2004) Listening to speech activates motor areas involved in speech production. *Nat Neurosci* 7:701–702.
- Wright TM, Pelphrey KA, Allison T, McKeown MJ, McCarthy G (2003) Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cereb Cortex* 13:1034–1043.